Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων
Εργαστηριο Συστηματων Τεχνητης Νοημοσυνης και Μαθησης

Enhancing Vision-Language Models: The role of LLMs in Augmenting Performance and Reasoning

Diploma Thesis

by

Penelope Stamou

Επιβλέπων: Αθανάσιος Βουλόδημος
Επίκουρος Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2025

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης

Enhancing Vision-Language Models: The role of LLMs in Augmenting Performance and Reasoning

Diploma Thesis

by

Penelope Stamou

Επιβλέπων: Αθανάσιος Βουλόδημος
Επίκουρος Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 21^{η} Μαρτίου, 2025.

....	....	....
Αθανάσιος Βουλόδημος	Α.-Γ. Σταφυλοπάτης	Κ. Νικήτα
Επίκουρος Καθηγητής Ε.Μ.Π.	Ομότιμος Καθηγητής Ε.Μ.Π.	Καθηγήτρια Ε.Μ.Π.

Αθήνα, Μάρτιος 2025

....
Πηνελοπη Σταμου
Διπλωματούχος Ηλεκτρολόγος Μηχανικός
και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Τα Μοντέλα Όρασης-Γλώσσας (VLMs) παρουσιάζουν πολύ καλές επιδόσεις σε σύνθετες οπτικο-γλωσσικές εργασίες. Τα αποτελέσματα πολλών ερευνητικών εργασιών δείχνουν πως οι τεχνικές προτροπής και οι μέθοδοι λεπτομερούς προσαρμογής μπορούν να χρησιμοποιηθούν για να ενισχύσουν την απόδοση των VLMs. Από την άλλη πλευρά, τα σύγχρονα πολυτροπικά LLMs εξακολουθούν να αντιμετωπίζουν δυσκολίες σε εργασίες που απαιτούν σύνθετη λογική, εξωτερική γνώση και απαντήσεις ευθυγραμμισμένες με τον άνθρωπο. Στην παρούσα διπλωματική εργασία, εξετάζουμε τους περιορισμούς των μεγάλων πολυτροπικών μοντέλων στην αντιμετώπιση προβλημάτων που απαιτούν εξωτερική γνώση και κοινή λογική. Εστιάζοντας στα σύνολα δεδομένων Stanford Image Paragraph Captioning και OK-VQA, διαπιστώνουμε ότι αν και τα πολυτροπικά LLMs παρουσιάζουν γνωστικές, γλωσσικές και λογικές ικανότητες, η απόδοσή τους περιορίζεται όταν αντιμετωπίζουν πολλές σύνθετες εργασίες ταυτόχρονα, ή όταν προσπαθούν να δώσουν απαντήσεις σε συγκεκριμένη μορφή, ακολουθώντας προκαθορισμένους κανόνες. Τα αποτελέσματα και η ανάλυσή μας δείχνουν ότι τα πολυτροπικά LLMs τελευταίας τεχνολογίας ξεπερνούν πολλές φορές τα υπάρχοντα σύνολα δεδομένων στην παραγωγή παραγράφων, ειδικά στις λεπτομέρειες που δίνουν, αλλά δυσκολεύονται στην εξαγωγή των σημαντικών στοιχείων του οπτικού περιεχομένου. Ομοίως, δυσκολεύονται με σύνολα δεδομένων περιγραφών εικόνων που στηρίζονται σε ερωτήματα που απαιτούν γνώση, όπως το OK-VQA. Για να ενισχύσουμε την απόδοσή τους σε αυτήν την περίπτωση, χρησιμοποιούμε ένα συνεργατικό πλαίσιο που περιλαμβάνει τρία μοντέλα: τον **Ανιχνευτή**, ένα LVLM που δέχεται μια εικόνα ως είσοδο και την περιγράφει σε μια παράγραφο, τον **Αναλυτή**, ένα LLM που δημιουργεί μια αρχική απάντηση στην ερώτηση με βάση την περιγραφή της εικόνας και τον **Διαμορφωτή**, ένα LLM που εξάγει και μορφοποιεί την τελική απάντηση με βάση ένα σύνολο προκαθορισμένων κανόνων.

Λέξεις-κλειδιά — Μεγάλα Γλωσσικά Μοντέλα, Πολυτροπικά Γλωσσικά Μοντέλα, Μοντέλα Όρασης-Γλώσσας, Συστήματα Πολλαπλών Δραστών, Απάντηση Οπτικών Ερωτήσεων με Χρήση Εξωτερικής Γνώσης, Περιγραφή Εικόνας σε Παράγραφο

Abstract

Vision-Language Models (VLMs) have demonstrated remarkable capabilities in complex visio-linguistic tasks. An extensive body of work has explored how prompting techniques and fine-tuning methods can be used to enhance their performance. However, modern multimodal LLMs still struggle with tasks that require complex reasoning, external knowledge, and human-aligned responses. In this work, we investigate the limitations of large-scale, multimodal models in handling open-ended tasks that demand external knowledge and commonsense reasoning. Focusing on the Stanford Image Paragraph Captioning and OK-VQA datasets, we find that although these models demonstrate substantial cognitive, linguistic, and reasoning abilities, their performance deteriorates when managing complex tasks simultaneously while adhering to specific response formats. Our analysis reveals that state-of-the-art multimodal models surpass existing datasets in paragraph generation but continue to face challenges in generating high-quality paragraphs. Similarly, they continue to struggle with knowledge-based, open-ended benchmarks such as OK-VQA. To boost their performance in the latter, we employ a collaborative framework comprising three models: the **Scout**, an LVLM that takes an image as input and describes it in a paragraph; the **Analyser**, an LLM that generates an initial answer to the question based on the image description; and the **Resolver**, an LLM that extracts and formats the final answer based on a set of predefined rules. Our framework yields improved performance over the single-agent baseline, indicating the effectiveness of a collaborative approach.

Keywords — Large Language Models (LLMs), Multimodal Large Language Models (MLLMs), Vision-Language Models (VLMs), Multi-Agent Systems (MAS), Knowledge-Based Visual Question Answering (K-VQA), Image Paragraph Captioning (IPC)

Ευχαριστίες

Με την παρούσα διπλωματική εργασία ολοκληρώνεται η φοίτησή μου στη Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου. Θα ήθελα να ευχαριστήσω θερμά όλους τους ανθρώπους που με στήριξαν όλα αυτά τα χρόνια και πίστεψαν σε εμένα και στα όνειρά μου.

Ιδιαίτερες ευχαριστίες οφείλω στον επιβλέποντά μου, κ. Αθανάσιο Βουλόδημο, για την πολύτιμη καθοδήγηση και υποστήριξή του, τόσο εντός όσο και εκτός του πλαισίου της διπλωματικής εργασίας. Επίσης, θα ήθελα να ευχαριστήσω ιδιαίτερα τη Μαρία Λυμπεραίου και τον Γιώργο Φιλανδριανό για την καθοριστική συμβολή τους τόσο στην έμπνευση όσο και στην ολοκλήρωση της παρούσας εργασίας.

Θα ήθελα ακόμη να εκφράσω την ευγνωμοσύνη μου στους δικούς μου ανθρώπους και ιδιαίτερα στους γονείς μου και τον αδερφό μου, για την αμέριστη υποστήριξή τους σε κάθε επίπεδο και για το ότι με έκαναν εξαρχής να πιστεύω πως δεν υπάρχει τίποτα που να μην μπορώ να πετύχω. Ευχαριστώ ιδιαίτερα την ξαδέρφη, συμμαθήτρια και συμφοιτήτριά μου, Έλενα, χωρίς την οποία η φοίτηση στη σχολή θα ήταν μία τελείως διαφορετική εμπειρία. Τέλος, ευχαριστώ θερμά τους φίλους μου, εντός και εκτός σχολής, για όλες τις στιγμές ξεγνοιασιάς και μη που μοιραστήκαμε αυτά τα χρόνια.

Πηνελόπη Στάμου, Μάρτιος, 2025

Contents

List of Figures

List of Tables

Chapter 1

Εκτεταμένη Περίληψη στα Ελληνικά

1.1 Θεωρητικό Υπόβαθρο

1.1.1 Εισαγωγή

Κίνητρο

Η συνεχώς αυξανόμενη παρουσία των Μεγάλων Γλωσσικών Μοντέλων (LLMs) επηρεάζει σχεδόν κάθε πτυχή της σύγχρονης ζωής, φέρνοντας επανάσταση στην αναζήτηση γνώσης, στη συλλογιστική και στη διεξαγωγή χειρωνακτικών εργασιών. Ωστόσο, η κατανόηση, η πρόβλεψη και η πλαισίωση αυτών των αδιαφανών συστημάτων παραμένουν ιδιαίτερα περίπλοκες. Μάλιστα, η πολυπλοκότητα εντείνεται καθώς μεταβαίνουμε από μονολιθικά μοντέλα σε σύνθετα συστήματα τεχνητής νοημοσύνης με πολλαπλά εξειδικευμένα στοιχεία, καθιστώντας την ανίχνευση λαθών ακόμη δυσκολότερη. Στο πλαίσιο αυτό, τα Πολυπρακτορικά Συστήματα (MAS) προσφέρουν μία πιο διαφανή εναλλακτική, επιτρέποντας την κατανομή και τη διεκπεραίωση εργασιών σε πολλά βήματα μέσω πολλαπλών πρακτόρων. Η παρούσα έρευνα επικεντρώνεται στη μελέτη της δυνατότητας των LLMs να ενισχύσουν τα μοντέλα Οπτικο-Γλωσσικής Κατανόησης (VLMs) για τη διεξαγωγή σύνθετων εργασιών γνώσης και συλλογιστικής, όπως η Περιγραφή Εικόνας σε Παραγράφο (IPC) και η Απάντηση Οπτικών Ερωτήσεων με Χρήση Εξωτερικής Γνώσης (Knowledge-Based VQA). Παράλληλα, διερευνά τις επιδόσεις των σύγχρονων πολυτροπικών LLMs, συγκρίνοντας ατομικές και πολυπρακτορικές αρχιτεκτονικές, προκειμένου να αναδείξει την αποτελεσματικότητα της συνεργασίας για την επίλυση οπτικο-γλωσσικών προβλημάτων.

Ερευνητικά Ερωτήματα και Υποθέσεις

Τα σύγχρονα πολυτροπικά μοντέλα, όπως τα GPT-4, Claude 3 και Gemini 1.5, συνδυάζουν γλωσσική ευχέρεια και οπτική κατανόηση, επιτυγχάνοντας υψηλές επιδόσεις σε εργασίες που απαιτούν γνώση και συλλογιστική ([106], [143], [142]). Με αυτή τη βάση, η παρούσα μελέτη εξετάζει σε ποιο βαθμό τα μοντέλα αυτά μπορούν να περιγράψουν επαρκώς μια εικόνα μέσω μιας παραγράφου. Αξιολογώντας τις δυνατότητές τους και εντοπίζοντας τα βασικά σημεία αδυναμίας τους, η έρευνα συμβάλλει προτείνοντας κατευθύνσεις για την καλύτερη κατασκευή συνόλων δεδομένων και μετρικών αξιολόγησης στον χώρο της περιγραφής εικόνας σε παράγραφο, ούτως ώστε να αντανακλώνται οι πραγματικές ικανότητες των σύγχρονων πολυτροπικών LLM.

Τα γλωσσικά μοντέλα έχουν αποθηκευμένα διαφορετικά είδη γνώσης στις παραμέτρους τους ([117]; [3]), την οποία μπορούν να ανακτήσουν για την απάντηση ερωτήσεων που απαιτούν εξωτερική γνώση ([106], [143], [142]). Σε αυτό το πλαίσιο, εξερευνούμε τη δυνατότητά τους να απαντάνε σε σύνθετες ερωτήσεις για εικόνες του συνόλου δεδομένων OK-VQA, οι οποίες βασίζονται σε εξωτερική γνώση. Κατά αυτόν τον τρόπο, προσπαθούμε να κατανοήσουμε τους παράγοντες που περιορίζουν την επίδοση των πολυτροπικών γλωσσικών μοντέλων στο παρόν σύνολο δεδομένων, καθώς επίσης και να αναδείξουμε στρατηγικές για την αξιοποίηση του πλήρους εύρους των δυνατοτήτων τους.

Τα Πολυπρακτορικά Συστήματα έχουν χρησιμοποιηθεί για τη βελτίωση της επίδοσης των μοντέλων σε εργασίες που συνδυάζουν όραση και γλώσσα και απαιτούν γνώση και συλλογιστική ικανότητα (Πίνακας 3.1). Υποθέτουμε ότι μια ανάλογη συνεργατική προσέγγιση, η οποία βασίζεται στη συνεργασία πρακτόρων LLM, θα μπορούσε να βελτιώσει την επίδοση των πολυτροπικών LLM στο σύνολο OK-VQA. Η επιβεβαίωση αυτής της υπόθεσης θα πρόσφερε μια πιο αποτελεσματική προσέγγιση για την απάντηση οπτικών ερωτήσεων, αλλά και σαφέστερη κατανόηση των παραγόντων που περιορίζουν τις επιδόσεις των πολυτροπικών LLM στο σύνολο OK-VQA.

Πειραματικό Πλαίσιο

Για την πειραματική φάση της εργασίας, δοκιμάζουμε το Claude ως ένα μεγάλης κλίμακας, πολυτροπικό μοντέλο σε δύο εργασίες: την περιγραφή εικόνας σε παραγράφο και την απάντηση οπτικών ερωτήσεων με χρήση εξωτερικής γνώσης. Για την πρώτη εργασία, προτρέπουμε το μοντέλο να παράγει παραγράφους που περιγράφουν τις εικόνες ενός υποσυνόλου του Stanford Image Paragraph Captioning Dataset. Αξιολογούμε την επίδοση του μοντέλου με τη χρήση των γλωσσικών μετρήσεων που έχουν καθοριστεί από τους δημιουργούς του συνόλου δεδομένων Stanford. Παράλληλα, διεξάγουμε μια γλωσσική ανάλυση για να αξιολογήσουμε πόσο μακροσκελείς, λεπτομερείς, σημασιολογικά πλούσιες, ποικιλόμορφες και γραμματικά ορθές είναι οι παράγραφοι του μοντέλου, σε σχέση με αυτές του συνόλου δεδομένων.

Για τη δεύτερη εργασία, αξιοποιούμε το σύνολο δεδομένων OK-VQA, το οποίο περιλαμβάνει ερωτήσεις για εικόνες που απαιτούν εξωτερική γνώση για να απαντηθούν. Χρησιμοποιούμε τη μετρική που προτείνεται από τους δημιουργούς του συνόλου δεδομένων και αξιολογούμε την επίδοση ενός μεγάλης κλίμακας πολυτροπικού μοντέλου. Παράλληλα, προτείνουμε ένα σύστημα πολλαπλών πρακτόρων που συνίσταται από τρία μοντέλα: τον Ανιχνευτή, ένα πολυτροπικό γλωσσικό μοντέλο που λαμβάνει μια εικόνα ως είσοδο και την περιγράφει σε μια παράγραφο, τον Αναλυτή, ένα γλωσσικό μοντέλο που προτείνει μια αρχική απάντηση στην ερώτηση με βάση την περιγραφή της εικόνας και τον Διαμορφωτή, ένα γλωσσικό μοντέλο που εξάγει και μορφοποιεί την τελική απάντηση με βάση ένα σύνολο προκαθορισμένων κανόνων.

Συνεισφορά

Η παρούσα εργασία συνεισφέρει στον διάλογο για τα πολυτροπικά και τα πολυπρακτορικά συστήματα αναδεικνύοντας βασικούς περιορισμούς στην ταυτόχρονη διαχείριση εργασιών που απαιτούν οπτική συλλογιστική και εξωτερική γνώση, όπως η περιγραφή εικόνας σε παραγράφο και η απάντηση οπτικών ερωτήσεων με χρήση εξωτερικής γνώσης. Για την αντιμετώπιση των περιορισμών αυτών, προτείνουμε ένα συνεργατικό σύστημα που συνδυάζει ένα πολυτροπικό μοντέλο και δύο μεγάλα γλωσσικά μοντέλα. Η προσέγγιση αυτή οδηγεί σε βελτιωμένη απόδοση στο σύνολο δεδομένων OK-VQA. Παράλληλα, η γλωσσική ανάλυσή που διεξάγουμε για το σύνολο δεδομένων Stanford Image Paragraph Captioning μάς αποκαλύπτει ότι τα υπάρχοντα σύνολα δεν είναι αρκετά απαιτητικά για τα σύγχρονα γλωσσικά μοντέλα, υπογραμμίζοντας την ανάγκη για νέα σύνολα και μετρικές. Σε αυτό το πλαίσιο, προτείνονται κατευθυντήριες γραμμές για τη δημιουργία τους, με στόχο την αντιμετώπιση των αδυναμιών των σύγχρονων μοντέλων.

1.2 Θεωρητικό Υπόβαθρο

1.2.1 Μεγάλα Γλωσσικά Μοντέλα (ΜΓΜ)

Εισαγωγή

Τα Μεγάλα Γλωσσικά Μοντέλα (LLMs) είναι γλωσσικά μοντέλα με μεγάλο αριθμό παραμέτρων, τα οποία εκπαιδεύονται με τεχνικές αυτοεπιβλεπόμενης μάθησης σε τεράστια σώματα κειμένου για την εκμάθηση στατιστικών μοτίβων στη γλώσσα. Τα περισσότερα σύγχρονα γλωσσικά μοντέλα βασίζονται στην αρχιτεκτονική μετασχηματιστών [142], η οποία μοντελοποιεί τις εξαρτήσεις μεταξύ των ακολουθιών εισόδου και εξόδου μέσω ενός μηχανισμού αυτοπροσοχής. Η προεκπαίδευση είναι το πρώτο κρίσιμο βήμα στην ανάπτυξη ενός γλωσσικού μοντέλου και περιλαμβάνει την εκπαίδευσή του σε σώματα κειμένου βάσει κάποιων στόχων αυτοεπιτήρησης και προεκπαίδευσης. Ο κύριος στόχος εκπαίδευσης ενός γλωσσικού μοντέλου συνήθως αφορά την πρόβλεψη της πιθανότητας εμφάνισης ενός κειμένου x [84]. Ωστόσο, καθώς τα LLMs κλιμακώνονται σε μέγεθος, συχνά εμφανίζουν αναδυόμενες ικανότητες, όπως η υψηλή συλλογιστική ή η μάθηση εντός πλαισίου, οι οποίες δεν υπάρχουν σε μικρότερα μοντέλα και δεν αποτελούν άμεσους στόχους της εκπαίδευσής τους [161]. Αυτές οι ικανότητες μπορούν να αξιοποιηθούν μέσω τεχνικών προτροπής, επιτρέποντας την εκτέλεση σύνθετων εργασιών χωρίς πρόσθετη ρύθμιση των παραμέτρων.

Δυνατότητες

Τα LLMs έχουν αναδειχθεί σε ιδιαίτερα ικανά συστήματα, επιδεικνύοντας ένα εύρος προηγμένων δυνατοτήτων που προσομοιάζουν την ανθρώπινη νοημοσύνη. Η αρχιτεκτονική μετασχηματιστών τους προσδίδει εξαιρετική ευφράδεια και γλωσσική επάρκεια, επιτρέποντάς τους να κατανοούν και να αναπαράγουν τον ανθρώπινο διάλογο, να εκδηλώνουν ενσυναίσθηση και να εμφυσούν την εμπιστοσύνη στον άνθρωπο ([108]; [94]; [185]; [52]; [14]). Πέρα από τις γλωσσικές τους δυνατότητες, τα μεγάλα γλωσσικά μοντέλα διαπρέπουν σε εργασίες που απαιτούν υψηλή συλλογιστική ([32]; [159]), ενώ παράλληλα ενσωματώνουν μεγάλο όγκο γνώσης στις παραμέτρους τους, λειτουργώντας ως ευέλικτες βάσεις γνώσης ([117]; [5]). Επιπλέον, διαθέτουν ισχυρές δεξιότητες σχεδιασμού και συντονισμού: μπορούν να αναλύουν εργασίες σε επιμέρους βήματα, να αναθέτουν υποκαθήκοντα σε άλλα μοντέλα, να χρησιμοποιούν εξωτερικά εργαλεία και να σχηματίζουν πολυπρακτορικά συστήματα για τη συνεργατική επίλυση προβλημάτων. Έτσι, η γλώσσα λειτουργεί όχι μόνο ως μέσο έκφρασης, αλλά και ως εργαλείο δομημένης σκέψης και λήψης αποφάσεων.

Μηχανική Προτροπών

Τα τελευταία χρόνια, η εκμάθηση μέσω προτροπών (prompting) έχει αναδειχθεί ως μια ισχυρή εναλλακτική απέναντι στη συμβατική προσέγγιση της προεκπαίδευσης και λεπτής ρύθμισης (pretrain and fine-tune), επιτρέποντας στα μεγάλα γλωσσικά μοντέλα να εκτελούν σύνθετα καθήκοντα με ελάχιστη εποπτεία [85]. Αυτή η μετατόπιση στο πεδίο αναδεικνύει όχι μόνο την αποτελεσματικότητα της μηχανικής προτροπών (prompt engineering) ως προσέγγιση, αλλά και τις πραγματικές δυνατότητες των LLMs. Η συνεργασία μεταξύ πολλαπλών πρακτόρων, όπου διαφορετικά μοντέλα αλληλεπιδρούν, συλλογίζονται και εξειδικεύονται για να επιλύσουν σύνθετα καθήκοντα, θεωρείται το νέο υπόδειγμα στον χώρο των LLMs. Τα MAS εσωτερικεύουν εγγενώς πολλές από τις παραδοσιακές τεχνικές προτροπών που ενισχύουν τη συλλογιστική, τη λήψη αποφάσεων και την πρόσβαση σε γνώση, χωρίς να απαιτούν ρητή εκπαίδευση. Η εσωτερική δομή των MAS, μέσω της ανάθεσης ρόλων, της ανάλυσης εργασιών, της εξωτερικής ανατροφοδότησης και της δυναμικής συνεργασίας, προσομοιώνει πτυχές της ανθρώπινης νόησης ([71]; [49]; [16]). Ωστόσο, ζητήματα όπως η αστάθεια ρόλων και η προτροπή των πολυπρακτορικών συστημάτων αναδεικνύουν νέες προκλήσεις.

Προκλήσεις και Ηθικά Ζητήματα

Παρά την πρόοδο των μεγάλων γλωσσικών μοντέλων, εξακολουθούν να υφίστανται σημαντικές προκλήσεις και περιορισμοί. Σε τεχνικό επίπεδο, τα LLMs επιδεικνύουν ισχυρή τυπική γλωσσική επάρκεια, αλλά στερούνται αληθινής κατανόησης και λειτουργικών γλωσσικών ικανοτήτων, εγείροντας ερωτήματα για το κατά πόσο πραγματικά συλλογίζονται ή απλώς αναπαράγουν μοτίβα από τα δεδομένα προεκπαίδευσης ([95]; [50]; [184]; [17]). Παρουσιάζουν επίσης κενά γνώσης και περιορισμένη εξειδίκευση σε επιμέρους τομείς, όπως η υγεία, το δίκαιο και η οικονομία ([83]; [149]). Επίσης, είναι επιρρεπή σε παραισθήσεις (hallucinations), παράγοντας πειστικά αλλά εσφαλμένα αποτελέσματα, ιδίως όταν αμφιβάλλουν για κάποια πληροφορία ή δίνουν προτεραιότητα στην ολοκλήρωση μίας εργασίας ([120]). Οι παραισθήσεις αποτελούν ένα ακόμα φαινόμενο που έχει υπονομεύσει τόσο την απόδοση όσο και την αξιοπιστία των LLMs, ιδίως στην περίπτωση των πολυτροπικών μοντέλων, όπου οι παραισθήσεις είναι όχι μόνο εντονότερες αλλά και πιο εύκολα αντιληπτές από τον άνθρωπο ([93]). Η "αγνωσία", η αποτυχία δηλαδή ορθής ερμηνείας αισθητηριακών εισροών, ιδιαίτερα η αναγνώριση οπτικών στοιχείων, όπως αντικείμενα, χρώματα και χωρικές σχέσεις, είναι μία ακόμα έλλειψη που έχει παρατηρηθεί ([131]). Παράλληλα, η καθοδηγησιμότητα των μοντέλων (steerability) παραμένει περιορισμένη, υπονομεύοντας την παραγωγικότητα και τον έλεγχο. Σε ηθικό επίπεδο, τα LLMs ενισχύουν ζητήματα προκατάληψης και παραπληροφόρησης, μέσω της διαιώνισης στερεοτύπων και μετάδοσης εσφαλμένης ή επιβλαβούς πληροφορίας. Παρά την πρόοδο, τα μοντέλα παραμένουν αδιαφανή, συχνά παρέχοντας παραπλανητικές αιτιολογήσεις που μιμούνται την ανθρώπινη σκέψη χωρίς να αποκαλύπτουν την πραγματική διαδικασία λήψης αποφάσεων. Αυτές οι προκλήσεις αναδεικνύουν τα τεχνικά και θεωρητικά όρια των σύγχρονων LLMs και υπογραμμίζουν την ανάγκη για προσεκτικό σχεδιασμό, αξιολόγηση και διακυβέρνηση ώστε να διασφαλιστεί η ασφαλής και υπεύθυνη χρήση τους.

1.2.2 Μοντέλα Οπτικής-Γλωσσικής Κατανόησης

Τα Μοντέλα Οπτικής-Γλωσσικής Κατανόησης (Vision-Language Models, VLMs) είναι μοντέλα που εκπαιδεύονται να κατανοούν και να παράγουν πληροφορία μέσω οπτικών και κειμενικών μορφών, μαθαίνοντας ευθυγραμμισμένες αναπαραστάσεις μέσω συνόλων δεδομένων εικόνας-κειμένου [58]. Τα πρώιμα VLMs, όπως τα CLIP και ALIGN, βασίστηκαν στην αντιστοίχιση εικόνων και κείμενων σε έναν κοινό χώρο ενσωμάτωσης ([119]; [58]). Πιο πρόσφατα, εμφανίστηκαν τα Μεγάλα Μοντέλα Όρασης-Γλώσσας (Large Vision-Language Models, LVLMs), τα οποία ενσωματώνουν αρχιτεκτονικές βασισμένες σε μετασχηματιστές, επιτρέποντας την προεκπαίδευση σε μεγάλα πολυτροπικά σώματα δεδομένων [4]. Τεχνικές προτροπής (prompting) μπορούν να ξεκλειδώσουν σύνθετες πολυτροπικές ικανότητες συλλογισμού. Μοντέλα όπως το BLIP-2 [73] γεφυρώνουν τα VLMs και τα LLMs, συνδυάζοντας "παγωμένα" LLMs με κωδικοποιητές οπτικών δεδομένων. Αυτή η εξέλιξη οδήγησε στα μεγάλης κλίμακας πολυτροπικά LLMs όπως το GPT-4 [106], τα οποία είναι ικανά να επεξεργάζονται ταυτόχρονα οπτικό περιεχόμενο και κείμενο.

Κύριες προκλήσεις στην έρευνα των VLMs περιλαμβάνουν τις παραισθήσεις, την ασφάλεια, τη δικαιοσύνη, την ευθυγράμμιση των πολυτροπικών δεδομένων, τη συλλογιστική, την αποδοτικότητα της εκπαίδευσης και τη σπανιότητα δεδομένων ([80]; [43]). Αρκετές μελέτες έχουν εξερευνήσει τη χρήση βάσεων γνώσης και ειδικότερα μεγάλων γλωσσικών μοντέλων για την ενίσχυση των δυνατοτήτων των VLMs ([22]; [149]).

1.2.3 Πολυπρακτορικά Συστήματα

Σε αντίθεση με τα παλαιότερα μοντέλα μηχανικής μάθησης, τα οποία λειτουργούσαν ανεξάρτητα, τα σημερινά συστήματα που βασίζονται σε LLMs ενσωματώνουν πολλαπλά εξειδικευμένα υποσυστήματα, όπως ανακτητές γνώσης, διερμηνείς κώδικα και εξωτερικά εργαλεία, ώστε να εκτελούν πιο σύνθετες εργασίες [182]. Αυτή η μετάβαση οδήγησε στην ανάπτυξη πρακτόρων βασισμένων σε LLMs (LLM agents), οι οποίοι είναι ικανοί να λαμβάνουν αυτόνομες αποφάσεις και να επιλύουν προβλήματα συνεργατικά, είτε ανεξάρτητα σε Συστήματα Μονού Πράκτορα (Single-Agent Systems) είτε σε συντονισμό με άλλους πράκτορες σε Συστήματα Πολλαπλών Πρακτόρων (Multi-Agent Systems) [77]. Τα LLMs βρίσκονται στην αιχμή αυτών των καινοτομιών, ενισχύοντας τις δυνατότητες της σύνθετης τεχνητής νοημοσύνης (compound AI) και των Συστημάτων Πολλαπλών Πρακτόρων, δρώντας ως πράκτορες και συντονιστές. Η σύνθετη τεχνητή νοημοσύνη και τα Συστήματα Πολλαπλών Πρακτόρων έχουν σημειώσει αξιοσημείωτη επιτυχία σε διάφορους τομείς. Ωστόσο, παρά την πρόοδο, παραμένουν προκλήσεις στη βελτιστοποίηση αυτών των συστημάτων και στην κατανόηση της δυναμικής της συνεργασίας ανθρώπου-πράκτορα.

1.3 Προσεγγίσεις

1.3.1 Περιγραφή Εικόνας σε Παράγραφο και Πυκνή Περιγραφή Εικόνας

Χρησιμοποιούμε ένα τυχαία επιλεγμένο υποσύνολο 5.000 εικόνων από το σύνολο δεδομένων Stanford Image Paragraph Captioning. Για την αξιολόγηση της επίδοσης του μοντέλου, επιλέγουμε το Claude ως ένα μοντέλο αιχμής στη κατηγορία των πολυτροπικών LLMs και διεξάγουμε το μεγαλύτερο μέρος των πειραμάτων μας χρησιμοποιώντας την έκδοση 3.7 Sonnet. Το Claude πρόκειται για μια οικογένεια μεγάλων πολυτροπικών μοντέλων που έχει αναπτυχθεί από την Anthropic και περιλαμβάνει τρεις τύπους μοντέλων (Haiku, Sonnet και Opus). Και τα τρία μοντέλα υποστηρίζουν είσοδο τόσο κειμένου όσο και εικόνας και παρουσιάζουν υψηλές επιδόσεις σε δείκτες οπτικού συλλογισμού, γεγονός που μας οδηγεί στην υπόθεση ότι είναι ικανά να παράγουν λεπτομερείς περιγραφές εικόνων. Πραγματοποιούμε πειράματα με όλα τα μέλη της οικογένειας Claude 3.7, αλλά επιλέγουμε το Claude 3.7 Sonnet για τα κύρια πειράματά μας στο σύνολο δεδομένων Stanford Image Paragraph Description, καθώς προσφέρει την καλύτερη ισορροπία μεταξύ απόδοσης και κόστους.

Δίνουμε ως είσοδο στο Claude Sonnet 3.7 τις εικόνες του υποσυνόλου του Stanford και το προτρέπουμε να παράγει περιγραφές σε επίπεδο παραγράφου με προτροπή μηδενικής βολής (zero-shot). Αξιολογούμε τις παραγόμενες περιγραφές χρησιμοποιώντας τους δείκτες METEOR, BLEU-{1,2,3,4} και τη συνημιτονική ομοιότητα SBERT. Στη συνέχεια, αναλύουμε γλωσσικά τόσο τις αυθεντικές παραγράφους του συνόλου δεδομένων όσο και τις παραγράφους που παράγονται από το Claude.

Για την γλωσσική μας ανάλυση, αξιολογούμε τα εξής:

- **Μέσο μήκος παραγράφου**: Μακρύτερες παράγραφοι υποδηλώνουν πλουσιότερη λεπτομέρεια ή πιο σύνθετη αφηγηματική δομή, καθώς και την παρουσία άλλων γλωσσικών φαινομένων.
- **Τυπική απόκλιση μήκους**: Μετρά τη μεταβλητότητα στο μήκος των περιγραφών μεταξύ συνόλου δεδομένων. Η χαμηλή τυπική απόκλιση υποδηλώνει συνέπεια στις παραγόμενες περιγραφές, ενώ η υψηλή απόκλιση υποδεικνύει ευελιξία στην προσαρμογή στην ποσότητα οπτικών πληροφοριών που περιέχει η εικόνα.
- **Μέγεθος λεξιλογίου**: Αναφέρεται στον συνολικό αριθμό μοναδικών λέξεων που χρησιμοποιούνται για όλες τις παραγόμενες παραγράφους και αποτελεί βασικό δείκτη της εκφραστικής εμβέλειας του μοντέλου. Ένα μεγάλο λεξιλόγιο υποδηλώνει ότι το μοντέλο μπορεί να περιγράψει μια ευρεία ποικιλία εννοιών, ενεργειών, χαρακτηριστικών και σχέσεων, αποφεύγοντας την επαναληπτικότητα και τη γενίκευση. Αντίθετα, ένα μοντέλο με μικρό μέγεθος λεξιλογίου μπορεί να καταφεύγει σε υπερβολικά γενικές ή ακόμη και ανακριβείς περιγραφές.
- **Λεξιλογική ποικιλία (Αναλογία Τύπων προς Δείγματα, TTR)**: Η λεξιλογική ποικιλία είναι ένα μέτρο του πόσο ποικίλο είναι το λεξιλόγιο σε ένα κείμενο. Ένας υψηλότερος δείκτης TTR υποδηλώνει μεγαλύτερο λεξιλογικό πλούτο και πιο ποικιλόμορφη, ανθρώπινη γλώσσα. Αντίθετα, ένας χαμηλότερος δείκτης TTR μπορεί να υποδεικνύει πιο επαναληπτική και άκαμπτη γλώσσα. Για να αξιολογήσουμε τη

λεξιλογική ποικιλία ανά παράγραφο, επιλέγουμε ως μέτρο τον δείκτη αναλογίας τύπων προς δείγματα (TTR), τον οποίο ορίζουμε ως εξής:

– **Τύποι (Types)** = ο αριθμός των μοναδικών λέξεων σε μια παράγραφο

– **Δείγματα (Tokens)** = το σύνολο των λέξεων που παράγονται στην παράγραφο

– **TTR** = Τύποι / Δείγματα

- **Ρυθμός γραμματικών λαθών**: Αξιολογεί τη γραμματική ορθότητα υπολογίζοντας το μέσο ποσοστό γραμματικών λαθών ανά πρόταση. Χρησιμοποιούμε το LanguageTool API, έναν ελεγκτή γραμματικής που μπορεί να ανιχνεύσει ένα ευρύ φάσμα σφαλμάτων σε αγγλικά κείμενα, όπως γραμματικά λάθη, σφάλματα στίξης ή ορθογραφικά λάθη. Ο τελικός ρυθμός γραμματικών λαθών προκύπτει ως εξής:

  – **Ρυθμός Γραμματικών Λαθών** = Συνολικός Αριθμός Γραμματικών Λαθών / Αριθμός Προτάσεων

- **Συχνότητα υποτακτικών συνδέσμων**: Σύνθετες παράγραφοι χρησιμοποιούν υποτακτικούς συνδέσμους (π.χ., επειδή, αν και, ενώ, προκειμένου να) για την έκφραση αιτιακών, χρονικών, υποθετικών και αντιθετικών σχέσεων, μεταξύ άλλων. Η χρήση αυτών των συνδέσμων αντανακλά πιο προχωρημένη δομή, μεγαλύτερη γλωσσική επιδεξιότητα και επικοινωνία βαθύτερων ιδεών. Για να αξιολογήσουμε τη συχνότητα αυτών των δεικτών, δημιουργήσαμε ένα λεξικό που περιλαμβάνει διάφορους τύπους συνδέσμων, όπως συνδέσμους αιτίας και σκοπού (π.χ., επειδή, προκειμένου να), και προϋπόθεσης (π.χ., αν, εκτός αν) (Πίνακας 5.16).

- **Αναλογία ουσιαστικών, ρημάτων και αντωνυμιών**: Η αναλογία ουσιαστικών, ρημάτων, αντωνυμιών και άλλων μερών του λόγου παρέχει πληροφορίες για τη γλωσσική εστίαση της περιγραφής. Υψηλότερη συχνότητα ουσιαστικών υποδηλώνει περιγραφές επικεντρωμένες σε αντικείμενα, οι οποίες συνδέονται με χαμηλού και μεσαίου επιπέδου σημασιολογία. Αντίθετα, υψηλή αναλογία ρημάτων υποδεικνύει περιγραφές προσανατολισμένες στη δράση και ισχυρότερη αφηγηματική ροή (π.χ., τρέχει, κρατά, ρίχνει) και συνδέεται με σημασιολογία μεσαίου επιπέδου.

1.3.2 Συστήματα Πολλών Πρακτόρων για Απάντηση Οπτικών Ερωτήσεων με Χρήση Εξωτερικής Γνώσης

Για το δεύτερό μας πείραμα, εξερευνούμε το σύνολο δεδομένων OK-VQA, το οποίο περιλαμβάνει ερωτήσεις για εικόνες που απαιτούν εξωτερική γνώση για την απάντησή τους. Οι ερωτήσεις προέρχονται από 10 κατηγορίες, οι οποίες παρουσιάζονται στο Σχήμα 4.1.6. Διεξάγουμε όλα τα πειράματα στο σύνολο validation, το οποίο αποτελείται από 5.046 εικόνες, και χρησιμοποιούμε την ορισμένη από τους δημιουργούς του συνόλου μετρική ώστε να διασφαλίσουμε δίκαιη σύγκριση. Η προτεινόμενη προσέγγισή μας είναι ένα σύστημα πολλαπλών πρακτόρων (Multi-Agent System) που αποτελείται από τρία στάδια. Στο πρώτο στάδιο, χρησιμοποιούμε ένα MLLM (π.χ. Claude 3.7), το οποίο λαμβάνει εικόνες και καλείται να τις περιγράψει σε μια παράγραφο. Καλούμε αυτό το μοντέλο Ανιχνευτή (Scout), καθώς εξερευνά και συλλέγει πληροφορίες από την εικόνα χωρίς συγκεκριμένη καθοδήγηση. Στο επόμενο στάδιο, προτρέπουμε ένα LLM (π.χ. Llama 3.3) να απαντήσει στην ερώτηση σχετικά με την εικόνα, χρησιμοποιώντας μόνο την περιγραφή της εικόνας από τον Ανιχνευτή και την ίδια την ερώτηση. Ονομάζουμε αυτό το μοντέλο Αναλυτή (Analyser), καθώς λειτουργεί ως ο κύριος μηχανισμός συλλογισμού. Τέλος, χρησιμοποιούμε ένα ακόμη LLM για να διαμορφώσει την τελική απάντηση στην ερώτηση, βάσει ενός προκαθορισμένου συνόλου κανόνων. Ονομάζουμε αυτό το μοντέλο Διαμορφωτή (Resolver), καθώς είναι υπεύθυνο για τη λήψη της τελικής απόφασης ως προς την απάντηση στην ερώτηση. Το προτεινόμενο συνεργατικό σύστημα παρουσιάζεται στο Σχήμα 1.3.1.

Για κάθε μέθοδο, αναλύουμε την κατανομή του μήκους των λέξεων και τον αριθμό των περιπτώσεων μη-απάντησης (no-attempts), ώστε να εκτιμήσουμε κατά πόσο τα μοντέλα μπορούν να καθοδηγηθούν προς την παραγωγή απαντήσεων συγκεκριμένου μήκους και να προσπαθούν πάντα να απαντούν στις ερωτήσεις. Για την ανάλυση του μήκους λέξεων, υπολογίζουμε το ποσοστό των απαντήσεων που περιέχουν μία λέξη, δύο λέξεις, τρεις λέξεις και τέσσερις ή περισσότερες λέξεις. Για τον υπολογισμό των περιπτώσεων μη-απάντησης, χρησιμοποιούμε ένα σύνολο προκαθορισμένων φράσεων (παρουσιάζεται στον Πίνακα 4.5) που παράγουν συχνά τα μοντέλα όταν δεν απαντούν στην ερώτηση.

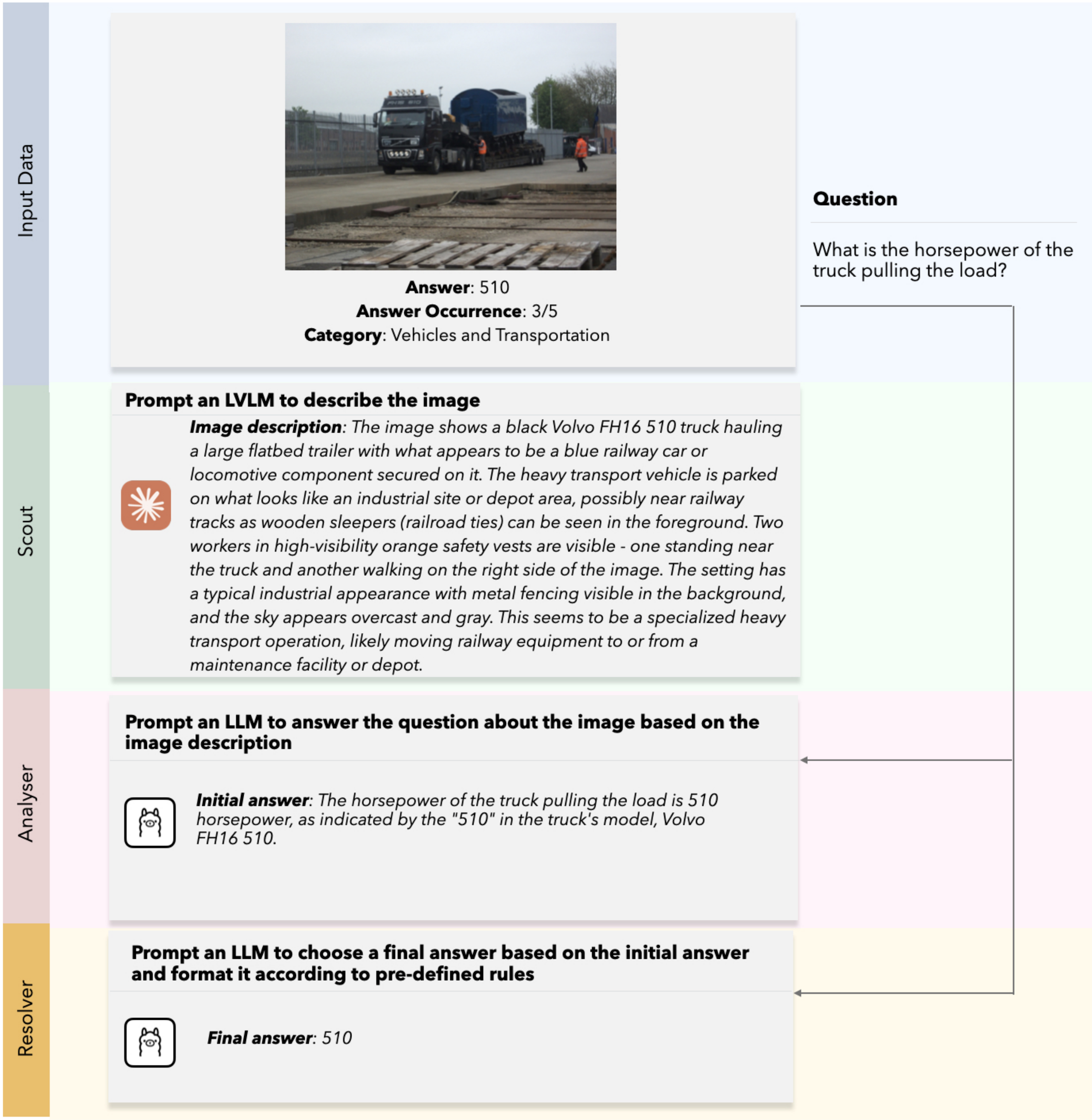

Figure 1.3.1: Σχηματική αναπαράσταση του πλαισίου συνεργασίας των Scout, Analyser και Resolver.

1.4 Πειράματα

1.4.1 Απάντηση Οπτικών Ερωτήσεων με Χρήση Εξωτερικής Γνώσης

Παρουσιάζουμε τα αποτελέσματα αξιολόγησης για το Claude 3.7 Sonnet στο υποσύνολο του Stanford Image Paragraph Captioning, για τις μετρικές METEOR και BLEU-{1,2,3,4}.

Είναι σαφές ότι το Claude 3.7 παράγει παραγράφους με σημασιολογική συνάφεια, όπως αποδεικνύεται από την υψηλή βαθμολογία METEOR, η οποία υπερβαίνει εκείνη των περιγραφών που έχουν γραφτεί από ανθρώπους. Ωστόσο, οι παραγόμενες παράγραφοι του μοντέλου διαφέρουν σημαντικά από την ακριβή διατύπωση των ανθρώπινων αναφορών, με αποτέλεσμα χαμηλές βαθμολογίες BLEU, ιδιαίτερα για μεγαλύτερα n-grams. Αυτό υποδηλώνει ότι, αν και το Claude αποτυπώνει το νόημα της εικόνας, αποκλίνει από τις παραγράφους αναφοράς

Model	METEOR	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Claude 3.7 - Sonnet	24.35	21.15	9.30	4.01	1.84
Human	19.22	42.88	25.68	15.55	9.66

Table 1.1: Σύγκριση απόδοσης μεταξύ του Claude 3.7 - Sonnet και των ανθρώπινων αναφορών με χρήση των μετρικών METEOR και BLEU.

ως προς τη διατύπωση. Αντίθετα, επιτυγχάνει υψηλή βαθμολογία SBERT (70,18%), γεγονός που υποδεικνύει ότι οι παραγόμενες παράγραφοι ταυτίζονται με τις παραγράφους αναφοράς όσον αφορά το πραγματικό περιεχόμενο.

Τα Σχήματα 1.4.1 και 1.4.2 απεικονίζουν τις διαφορές στους επικρατέστερους όρους μεταξύ των παραγόμενων και των παραγράφων αναφοράς, ενώ ο Πίνακας 1.2 παρουσιάζει τα αποτελέσματα της γλωσσικής ανάλυσης.

Τα αποτελέσματα υποδεικνύουν ότι το Claude πετυχαίνει ανώτερη γλωσσική απόδοση σε σύγκριση με τις παραγράφους αναφοράς, όπως φαίνεται από την συντριπτική πλειοψηφία των μετρικών. Πράγματι, οι παράγραφοι που παράγονται από το Claude είναι σχεδόν διπλάσιες σε μήκος σε σχέση με τις παραγράφους αναφοράς, παρουσιάζουν μεγαλύτερη λεξιλογική ποικιλία και σημαντικά χαμηλότερο ρυθμό γραμματικών λαθών. Αξιοσημείωτο είναι, επίσης, το μέγεθος του λεξιλογίου τους, το οποίο είναι σχεδόν διπλάσιο. Τέλος, οι παράγραφοι του μοντέλου, χρησιμοποιούν περισσότερους από τριπλάσιους υποτακτικούς συνδέσμους.

Figure 1.4.1: Σχηματική αναπαράσταση μέσω σύννεφου λέξεων των παραγράφων που σχολιάστηκαν από ανθρώπους στο υποσύνολο 5.000 εικόνων του Stanford Image Paragraph Captioning, αναδεικνύοντας τάση προς αντικειμενοκεντρική και κυριολεκτική περιγραφή.

Figure 1.4.2: Σχηματική αναπαράσταση μέσω σύννεφου λέξεων των παραγράφων που παρήχθησαν από το Claude 3.7 Sonnet για το υποσύνολο 5.000 εικόνων του Stanford Image Paragraph Captioning, υποδεικνύοντας τάση προς αφηρημένη και ερμηνευτική απόδοση.

	Avg. Length	**St. Dev. Length**	**Diversity (TTR)**	**Nouns (%)**	**Verbs (%)**	**Pronouns (%)**	**Vocab Size**	**Subord. Conj.**	**Grammar Error Rate**
Claude 3.7 Sonnet	118.96	11.38	0.6788	29.36	12.17	3.83	11,739	2,831	0.06
Stanford Dataset	61.36	23.54	0.6172	27.92	9.81	5.82	7,034	903	0.21

Table 1.2: Γλωσσική σύγκριση μεταξύ των παραγράφων του Claude 3.7 Sonnet και των παραγράφων αναφοράς του Stanford Image Paragraph Captioning συνόλου δεδομένων.

Image ID: 2356347 **Source**: Stanford Image Captioning Dataset

Human Annotators: A large building with bars on the windows in front of it. There is people walking in front of the building. There is a street in front of the building with many cars on it.

Claude 3.7 Sonnet: This image shows the modern architecture of what appears to be the Art Institute of Chicago's Modern Wing. The building features a striking design with clean lines, extensive glass panels, and a distinctive overhanging roof structure that creates a sheltering canopy. The exterior combines light-colored stone or concrete with steel and glass elements. The street view shows pedestrian crossings marked with yellow warning signs, ornate street lamps, and several vehicles parked along the curb. The scene is captured on what looks like a sunny day with blue skies, and there's some greenery visible around the building's perimeter. The overall design reflects contemporary architectural principles with its emphasis on transparency, geometric forms, and the integration of natural light.

Table 1.3: Σύγκριση παραγράφων για την εικόνα ID 2356346.

1.4.2 Συστήματα Πολλών Πρακτόρων για Απάντηση Οπτικών Ερωτήσεων με Χρήση Εξωτερικής Γνώσης

Χρησιμοποιούμε το Claude 3.7 Sonnet ως τη βασική γραμμή αναφοράς (baseline), παρέχοντάς του εικόνες και ερωτήσεις από το σύνολο validation του OK-VQA, χρησιμοποιώντας τις παραμέτρους επαγωγής (inference parameters) που αναφέρονται στον Πίνακα 5.3. Το μοντέλο καλείται να παράγει απαντήσεις αποτελούμενες από μία ή δύο λέξεις, καθώς επίσης και να προσπαθεί πάντα να απαντήσει στην ερώτηση, μέσω προτροπών μηδενικής βολής (zero-shot). Η ακρίβεια μετράται χρησιμοποιώντας τη μετρική soft accuracy που προτείνεται από τους δημιουργούς του dataset. Η μέση ακρίβεια ανέρχεται σε 43,59%, η οποία καθορίζεται ως η βασική γραμμή αναφοράς μας.

Τα αποτελέσματα του συνεργατικού συστήματος Ανιχνευτή-Αναλυτή-Διαμορφωτή, καθώς και τα αποτελέσματα άλλων μεθόδων με διαφορετικές τεχνικές προτροπής, παρουσιάζονται στους Πίνακες 1.8 και 1.9. Παρατηρούμε ότι όλες οι μέθοδοι αποδίδουν καλύτερα αποτελέσματα σε σύγκριση με τη χρήση του Claude, αναδεικνύοντας την αποτελεσματικότητα της προσέγγισης συνεργασίας πολλών πρακτόρων για αυτό το σύνολο. Μέσω της συνεργασίας, η απόδοση βελτιώνεται κατά 7% σε σχέση με το baseline, αναδεικνύοντας την αποτελεσματικότητα της χρήσης LLM για την εξαγωγή της τελικής απάντησης και την μορφοποίηση της απάντησης.

Η έξοδος του Ανιχνευτή για ένα παράδειγμα εικόνας φαίνεται στον Πίνακα 1.7, ενώ η γλωσσική ανάλυση για τις περιγραφές παρουσιάζεται στον Πίνακα 1.4. Επίσης, η αναπαράσταση των επικρατέστερων λέξεων των περιγραφών σε σύννεφο λέξεων παρουσιάζεται στο Σχήμα 1.4.3. Παραδείγματα για την έξοδο του συστήματος ανά βήμα παρουσιάζονται στους Πίνακες 1.5 και 1.6.

Avg. Length	**St. Dev. Length**	**Diversity (TTR)**	**Nouns (%)**	**Verbs (%)**	**Pronouns (%)**	**Vocab Size**	**Subord. Conj.**	**Grammar Error Rate**
115.65	18.15	0.6774	29.29	12.00	4.26	11,904	2,357	0.06

Table 1.4: Γλωσσική ανάλυση των παραγράφων του Ανιχνευτή για το σύνολο δεδομένων OK-VQA.

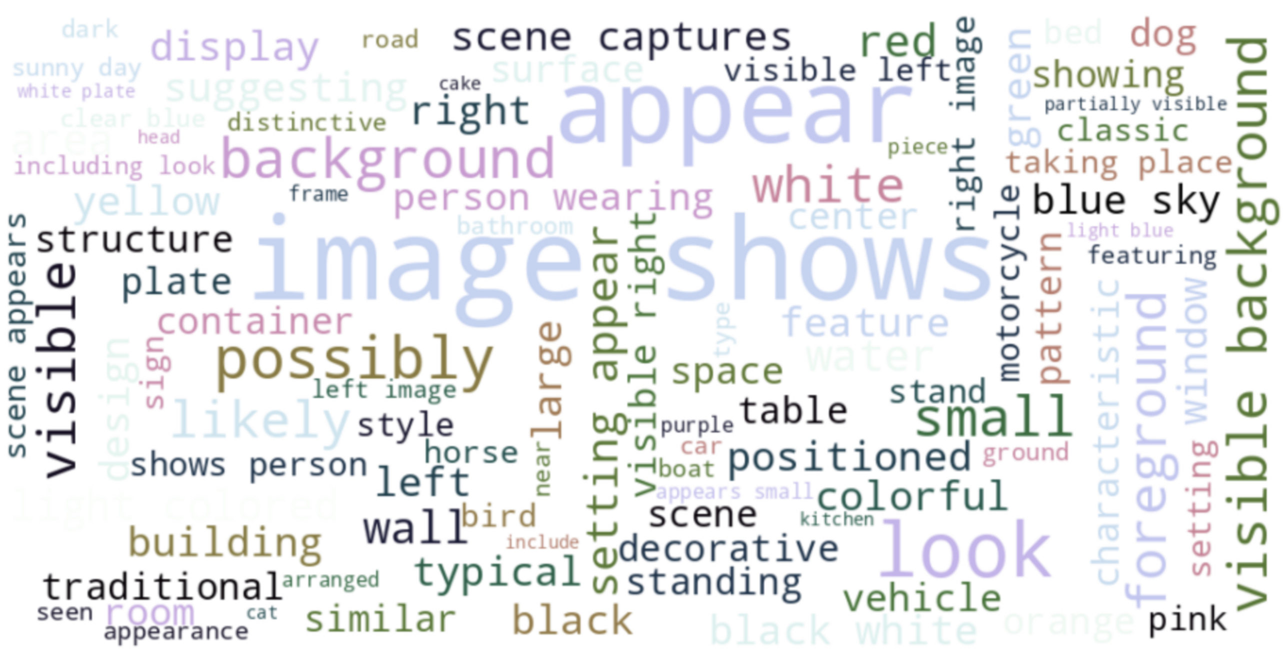

Figure 1.4.3: Επικρατέστεροι όροι σε σύννεφο λέξεων για τις περιγραφές παραγράφων του Ανιχνευτή στο OK-VQA.

Παράδειγμα 1: Σωστή απάντηση

Question: Which part of this animal would be in use if it was playing the game that is played with the items the man is holding?

Image Description (Scout): The image shows a person playing frisbee in a grassy yard while a dog is perched on their back. The person is wearing a white t-shirt, dark pants, and a black cap, and is bent forward holding two white frisbees - one in each hand. The dog appears to be a black and white spotted breed, possibly an Australian Cattle Dog or similar mix, and has impressively balanced itself on the person's back while they're in motion. In the background, there's a wooden fence. This looks like a fun moment of dog training or freestyle disc dog practice, where the dog has learned to perform tricks while the handler manipulates the frisbees.

Initial Answer (Analyser): If the dog was playing the game that is played with the items the man is holding (frisbees), the part of the animal that would be in use is its mouth, as dogs typically catch frisbees in their mouths.

Final Answer (Resolver): Mouth ✓

Ground Truth: mouth

Table 1.5: Παράδειγμα σωστής απάντησης του συστήματος Ανιχνευτή, Αναλυτή, Διαμορφωτή.
Notes: most salient sentence
second most salient sentence

Παράδειγμα 2: Εσφαλμένη απάντηση

Question: Which phase is that moon in?

Image Description (Scout): The image shows a clock tower with two visible clock faces on adjacent sides. The tower appears to be made of dark wooden shingles or similar material and has a small cupola or bell housing at the top with vented openings. The clock faces have Roman numerals, with one clock showing approximately 4:35. In the background, there's a crescent moon visible in the gray sky, and the silhouette of an evergreen tree can be seen next to the tower. The photo has a moody, atmospheric quality with its muted colors and twilight setting. ✗

Initial Answer (Analyser): The moon is in the crescent phase.

Final Answer (Resolver): crescent

Ground Truth: half

Error Type: Inaccurate Information

Table 1.6: Παράδειγμα εσφαλμένης απάντησης του συστήματος Ανιχνευτή, Αναλυτή, Διαμορφωτή
Notes: most salient sentence
second most salient sentence

Σχετικές προτάσεις στην παράγραφο του Αναλυτή

Question: What sport can you use this for?

Image Description (Scout): The image shows a black Honda motorcycle, likely a Varadero or similar adventure touring model, parked on a sandy/paved surface. The motorcycle features a prominent windshield, comfortable seat, and silver accents on the rear section. It has the characteristic dual-purpose design with wire-spoke wheels that are suitable for both on and off-road riding. In the background, there's a white metal container or storage unit, and what appears to be an orange flag or banner visible on the left side. The setting looks like an outdoor area with some dirt/sand terrain and a paved section where the motorcycle is standing on its kickstand. The scene suggests this might be at a motorcycle event, test riding area, or off-road riding location.

Table 1.7: Σχετικές προτάσεις στην παράγραφο του Αναλυτή.

Method	**Word distribution***		**Accuracy (%)**	**# of no attempts**
Claude Solo	1	80.04	43.59	60
	2	6.44		
	3	0.67		
	4+	12.84		
Analyser + Resolver Zero-Shot	1	96.27	49.65	61
	2	3.47		
	3	0.10		
	4+	0.16		
Analyser + Resolver Few-Shot	1	96.21	50.02	128
	2	3.09		
	3	0.10		
	4+	0.59		
Analyser CoT + Resolver Zero-Shot	1	81.81	46.50	130
	2	17.90		
	3	0.26		
	4+	0.04		
Analyser CoT + Resolver Few-Shot	1	87.44	46.86	48
	2	12.33		
	3	0.24		
	4+	0.00		
Scout + Analyser + Resolver Zero-Shot	1	92.63	47.05	1
	2	7.09		
	3	0.22		
	4+	0.06		
Scout + Analyser + Resolver Few-Shot	1	92.11	48.73	2
	2	7.59		
	3	0.20		
	4+	0.10		

Table 1.8: Συγκριτικά αποτελέσματα όλων των μεθόδων για το OK-VQA σύνολο δεδομένων.

Method	**Avg. (%)**	**VT**	**BCP**	**OMC**	**SR**	**CF**	**GHLC**	**PEL**	**PA**	**ST**	**WC**	**Other**
Claude Solo	43.59	41.43	44.88	45.70	40.63	47.53	44.54	41.12	41.42	44.05	46.67	46.01
Analyser-Resolver Zero-Shot	49.65	47.06	47.67	51.68	46.74	52.94	46.81	**48.36**	50.33	**50.71**	52.56	50.44
Analyser-Resolver Few-Shot	**50.02**	**47.38**	48.26	**51.92**	47.62	**53.14**	46.10	**48.36**	50.53	**50.71**	53.80	**51.24**
Analyser CoT-Resolver Zero-Shot	46.50	40.54	46.63	46.40	46.07	51.25	44.96	44.53	46.16	43.10	54.11	49.66
Analyser CoT-Resolver Few-Shot	46.86	41.60	46.98	47.20	48.29	49.84	46.67	44.63	45.47	44.29	**55.66**	50.37
Scout-Analyser-Resolver Zero-Shot	47.05	44.38	52.21	43.13	47.51	47.70	**53.76**	46.26	47.77	35.00	48.84	49.82
Scout-Analyser-Resolver Few-Shot	48.73	46.03	**54.42**	43.97	**48.78**	49.89	52.62	46.82	**51.00**	35.71	50.39	51.15

Table 1.9: Συγκριτικά αποτελέσματα των μεθόδων για το σύνολο δεδομένων OK-VQA ανά κατηγορία. *Notes:* VT = Vehicles and Transportation, BCP = Brands, Companies and Products, OMC = Objects, Material and Clothing, SR = Sports and Recreation, CF = Cooking and Food, GHLC = Geography, History, Language and Culture, PEL = People and Everyday Life, PA = Plants and Animals, ST = Science and Technology, WC = Weather and Climate.

1.5 Συμπεράσματα και Μελλοντικές Προεκτάσεις

Τα πολυτροπικά γλωσσικά μοντέλα αποτελούν τα πιο προηγμένα συστήματα τεχνητής νοημοσύνης που έχουν αναπτυχθεί μέχρι σήμερα. Ως αποτέλεσμα, αναπτύσσονται υψηλές προσδοκίες για επίλυση προβλημάτων από τα πολυτροπικά γλωσσικά μοντέλα, σχεδόν σε κάθε εφαρμογή. Η ισχυρή απόδοσή τους δικαιώνει τις περισσότερες φορές τις προσδοκίες αυτές και αφήνει ισχυρές ελπίδες για την ανάπτυξη ενός μοντέλου γενικής τεχνητής νοημοσύνης στο μέλλον. Επιπλέον, παρατηρούμε ότι πιθανά προβλήματά τους διορθώνονται από την προσαρμογή τους (μέσω μικρής κλίμακας εκπαίδευσης στο πεδίο γνώσης) ή αρχιτεκτονικές ανάπτυξης και συνεργασίας πολλαπλών πολυτροπικών γλωσσικών μοντέλων. Με την εργασία αυτή, προσπαθούμε να διερευνήσουμε τις δυνατότητες αυτές, με βάση ένα σύγχρονο πολυτροπικό γλωσσικό μοντέλο πολύ υψηλής επίδοσης, όπως το Claude, το οποίο χρησιμοποιούμε ως ένα πολυτροπικό μοντέλο μεσολάβησης σε δύο εργασίες: τη δημιουργία λεζάντας παραγράφων εικόνας και την απάντηση οπτικών ερωτημάτων (Visual Question Answering - VQA) που βασίζεται στη γνώση. Επιπλέον, αναπτύσσουμε ένα πλαίσιο πολλαπλών πρακτόρων για να αντιμετωπίσουμε αποτελεσματικότερα την απάντηση οπτικών ερωτημάτων, κάνοντας μάλιστα συγκρίσεις μεταξύ των προσεγγίσεων ενός μονολιθικού μοντέλου και μοντέλου πολλαπλών πρακτόρων.

Συγκεκριμένα, για το πρώτο πρόβλημα, ερευνήσαμε τη δημιουργία λεζάντας παραγράφων εικόνων, η οποία περιλαμβάνει τη δημιουργία περιγραφών μήκους παραγράφου για εικόνες. Ζητήσαμε από ένα μεγάλης κλίμακας, πολυτροπικό γλωσσικό μοντέλο να δημιουργήσει περιγραφές παραγράφων για τις εικόνες σε ένα υποσύνολο ενός σημαντικού συνόλου δεδομένων υποτίτλων παραγράφων εικόνας του που έχει παρουσιαστεί από το Πανεπιστήμιο Stanford σε μια ρύθμιση μηδενικής λήψης (zero-shot). Παρατηρούμε ότι οι γλωσσικές περιγραφές που πρότειναν οι δημιουργοί του συνόλου δεδομένων δεν ευθυγραμμίζονται πάντα με τις λεπτομερείς, πλούσιες και ρέουσες παραγράφους που εμπειρικά παρατηρήσαμε ότι δημιουργούνται από το μοντέλο. Ως εκ τούτου, πραγματοποιήσαμε μια γλωσσική ανάλυση για να συγκρίνουμε τις παραγράφους που γράφτηκαν από άνθρωπο, ως περιγραφή των εικόνων των δεδομένων, με τα αποτελέσματα περιγραφής των αντίστοιχων εικόνων που εξάγει το μοντέλο. Η ανάλυσή μας δείχνει ότι οι παράγραφοι που δημιουργούνται από το μοντέλο έχουν μεγαλύτερο μήκος, λεξιλογικό και σημασιολογικό πλούτο, ποικιλομορφία και γραμματική ορθότητα σε σύγκριση με τις παραγράφους που σχολιάζονται από τον άνθρωπο. Αυτά τα ευρήματα υποδηλώνουν ότι τα σύγχρονα πολυτροπικά γλωσσικά μοντέλα έχουν ξεπεράσει τα σύνολα δεδομένων και τις μετρήσεις που προϋπήρχαν, υπογραμμίζοντας την ανάγκη για νέα σημεία αναφοράς στην περιοχή.

Στο πλαίσιο του δεύτερου προβλήματος, εστιάσαμε στην απάντηση σε ερωτήματα που γίνονται σε εικόνες και απαιτούν για την απάντησή τους γνώση πεδίου. Συγκεκριμένα, ασχοληθήκαμε με το σύνολο δεδομένων OK-VQA, το οποίο αποτελεί σημείο αναφοράς στην περιοχή. Αξιολογήσαμε ένα πολυτροπικό γλωσσικό μοντέλο σε ερωτήματα μηδενικής λήψης και διαπιστώσαμε ότι, παρότι φαίνεται ότι έχει την απαραίτητη γνώση, δεν μπορεί να φτάσει στα επίπεδα απόδοσης της τεχνολογίας αιχμής (state-of-the-art), που επιτυγχάνεται κυρίως από συστήματα που έχουν προσαρμοστεί στα δεδομένα αυτά. Αυτό οφείλεται σε μεγάλο βαθμό στη μορφή των αναμενόμενων απαντήσεων, την οποία το μοντέλο δυσκολεύεται να τηρήσει μόνο μέσω της απλής προτροπής που χρησιμοποιήσαμε. Προκειμένου να αξιοποιηθεί πλήρως το δυναμικό του πολυτροπικού γλωσσικού μοντέλου, προτείνουμε ένα σύστημα πολλαπλών πρακτόρων που καθιερώνει τη συνεργασία μεταξύ του πολυτροπικού γλωσσικού μοντέλου και ενός απλού μεγάλου γλωσικού μοντέλου. Συγκεκριμένα, αναπτύξαμε και εκτελέσαμε πειράματα που χρησιμοποιούν τρία μοντέλα ως πράκτορες: τον **Ανιχνευτή**, που είναι ένα πολυτροπικό γλωσσικό μοντέλο που παίρνει μια εικόνα ως είσοδο και την περιγράφει σε μια λεζάντα παραγράφου, τον **Αναλυτή**, που είναι ένα απλό γλωσσικό μοντέλο που παράγει μια αρχική απάντηση στο ερώτημα με βάση την περιγραφή της εικόνας, και τον **Διαμορφωτή**, που είναι ένα απλό γλωσσικό μοντέλο που επιλέγει και μορφοποιεί την τελική απάντηση σύμφωνα με ένα σύνολο προκαθορισμένων κανόνων. Το σύστημα βελτιώνει την επίδοση σε σχέση με τη μοντέλο αναφοράς (baseline), υποδεικνύοντας ότι το πλαίσιο συνεργασίας μπορεί να γίνει αποτελεσματικό για τόσο σύνθετες εργασίες. Τα αποτελέσματα δείχνουν ότι τα πολυτροπικά γλωσσικά μοντέλα μπορούν να παράγουν περιγραφές εικόνων που είναι επαρκείς για να απαντήσουν στην ερώτηση στην πλειονότητα των περιπτώσεων.

Επιπλέον, τα γλωσσικά μοντέλα μπορούν να δώσουν απαντήσεις σε ερωτήματα σχετικά με εικόνες, αξιοποιώντας την περιγραφή της παραγράφου της εικόνας ως βάση, υπό την προϋπόθεση ότι οι οπτικές πληροφορίες που απαιτούνται για την απάντηση στην ερώτηση υπάρχουν στη λεζάντα. Τέλος, επικυρώνουμε ότι τα γλωσσικά μοντέλα μπορούν να χρησιμοποιηθούν αποτελεσματικά ως μορφοποιητές για να βοηθήσουν στην εξαγωγή της τελικής απάντησης και στη μεταγενέστερη επεξεργασία της για ευθυγράμμιση με την απαιτούμενη μορφή.

Για το μέλλον, σχεδιάζουμε να διεξάγουμε μελέτες επί μέρους ανάλυσης, για να κατανοήσουμε σε βάθος ποιο στοιχείο του συστήματός σχετίζεται με τις ορθές απαντήσεις και να αξιολογήσουμε ποιοι τύποι σφαλμάτων σχετίζονται με την πλειοψηφία των αποτυχιών των μοντέλων. Στοχεύουμε επίσης να διεξάγουμε πειράματα με επιπλέον μοντέλα για να διερευνήσουμε πώς η επεκτασιμότητα επηρεάζει την απόδοση των μοντέλων. Υποθέτουμε ότι τα μικρότερα γλωσσικά μοντέλα θα μπορούσαν να είναι αποτελεσματικά για αυτήν την εργασία, καθώς μπορεί να παρουσιάζουν λιγότερη αφαιρετική μεροληψία σε σύγκριση με μεγαλύτερα μοντέλα. Για την ελαχιστοποίηση των σφαλμάτων που σχετίζονται με ελλιπή περιγραφή, δηλαδή με περιγραφή που δεν περιέχει τα απαραίτητα στοιχεία για την απάντηση του ερωτήματος, σκοπεύουμε να εξετάσουμε τη χρήση *ελεγχόμενης* δημιουργίας λεζάντας εικόνας, όπου το κείμενο θα προσαρμόζεται με βάση το εκάστοτε ερώτημα ή προκαθορισμένους περιορισμούς που σχετίζονται με το εκάστοτε ερώτημα.

Τέλος, σκοπεύουμε να διαξάγουμε πειράματα που στηρίζονται στην αξιολόγησης από ανθρώπους (human evaluation), για να εκτιμήσουμε την απόδοση του συστήματός μας στην εργασία OK-VQA, βασιζόμενοι στην παραδοχή ότι οι υπάρχουσες μετρικές δεν αποτυπώνουν επαρκώς τις πραγματικές δυνατότητες των μοντέλων. Παράλληλα, σκοπεύουμε να διερευνήσουμε εναλλακτικές μεθόδους αυτόματης αξιολόγησης, όπως τη χρήση γλωσσικών μοντέλων ως αξιολογητών, οι οποίες θα μπορούσαν να υποστηρίξουν αποτελεσματικότερα εργασίες ανοιχτού τύπου. Επιπλέον, μία σημαντική μελλοντική κατεύθυνση αφορά την ανάπτυξη νέων συνόλων δεδομένων, ικανών να ανταποκριθούν στις αυξανόμενες δυνατότητες των πολυτροπικών γλωσσικών μοντέλων και συστημάτων πρακτόρων που στηρίζονται σε πολυτροπικά γλωσσικά μοντέλα. Ιδιαίτερη έμφαση σκοπεύουμε να δώσουμε στην ανάγκη δημιουργίας συνόλων δεδομένων στον τομέα της δημιουργίας λεζάντων εικόνων, τα οποία να θέτουν ουσιαστικές προκλήσεις για τα σύγχρονα πολυτροπικά γλωσσικά μοντέλα, στοχεύοντας ειδικότερα στα βασικά ζητήματα λειτουργίας τους που έχουμε εντοπίσει.

Chapter 2

Introduction

It is becoming increasingly difficult to find an aspect of modern life that remains untouched by the growing presence of Large Language Models (LLMs). With their close proximity to humans, combined with rapid articulation, knowledge retrieval, and reasoning capabilities, LLMs' promise resonates across domains. However, understanding, predicting, and framing these opaque models is not an easy task, and research aimed at establishing their true capabilities is more crucial than ever. The shift from monolithic models to compound AI systems is further complicating this endeavour, as architectures grow more and more complex while remaining largely undisclosed. Multimodality adds new layers of complexity by making it difficult to pinpoint the sources of model errors. Multi-Agent Systems (MAS) offer some transparency in that regard, as they allow for the decomposition of tasks and reasoning processes across agents, making it easier to inspect system behaviour.

LLMs are at the centre of MAS, using language as a vehicle for thought and communication [136], similar to how societies and civilisations evolved alongside the development of human language. LLM-based MAS have proven highly effective across a diverse range of tasks, outperforming single agents in many downstream applications (Table 3.1). This research sets out to explore the potential of LLMs in enhancing Vision-Language Models (VLMs) for reasoning-intensive and knowledge-based multimodal tasks, particularly those involving images and language. Focusing on two tasks, image paragraph captioning and knowledge-based VQA, this research seeks to benchmark the performance of current Multimodal Large Language Models (MLLMs), specifically Anthropic's Claude, on these tasks and compare solo and multi-agent architectures to address the latter.

2.1 Motivation

Language is at the core of advancements in LLMs, LVLMs, and MAS alike. Thanks to their transformer architecture, these models achieve levels of fluency and linguistic competence that surpass those of traditional statistical machine learning models ([108]; [94]; [185]). However, human language is only one of several ways in which LLMs mirror human behaviour. LLMs at scale are capable of storing multiple types of knowledge similar to traditional knowledge graphs ([117]; [5]). At the same time, they exhibit strong performance on a wide range of arithmetic, commonsense, logical, symbolic, and multimodal reasoning tasks [32]. They are also powerful problem-solvers, capable of formalising problems by translating natural language into mathematical models ([32]; [159]). That means they can plan actions, delegate tasks, and coordinate other models, all through the use of language ([156]; [168]). Thanks to these capabilities, LLMs can perform reasoning-intensive tasks, carry out manual operations, and coordinate other models ([162]; [31]; [182]; [22]).

However, it remains unclear whether LLMs' fluency necessarily reflects conscious reasoning and true knowledge or the reproduction of surface-level patterns learned during pretraining. In fact, they have been found to lack central aspects of linguistic competence, namely functional linguistic competence, which is associated with higher-level capacities, such as world knowledge [95]. Another crucial issue is the presence of cognitive gaps and the lack of domain specificity in the knowledge encoded within LLMs' parameters ([39]; [83]; [149]).

The combination of these two issues introduces a new challenge regarding the ability of LLMs to reason about factual knowledge. When it comes to multimodal models, hallucinations are also a major problem, involving the generation of content that is not visually grounded [120]. Finally, steerability is a crucial issue from an application-driven perspective, as we ideally want to be able to guide models toward specific actions and outputs through prompting ([99]; [17]).

Benchmarks that target these areas prove more challenging for LLMs and expose their shortcomings when it comes to handling reasoning-intensive and knowledge-based tasks. In Image Paragraph Captioning (IPC), the goal is to generate a coherent, multi-sentence paragraph that describes an image [65]. Paragraphs are linguistically complex structures, requiring the model to possess not only visual understanding of low- to mid-level picture elements but also a wide range of linguistic capabilities [109]. Hallucinations remain common in model-generated paragraphs, as models tend to follow a narrative flow and prioritise it over factual accuracy. In another VL task, Visual Question Answering (VQA), the model is provided with an image, a question about that image, and is tasked with predicting the answer. In Knowledge-Based Visual Question Answering (K-VQA), to generate the correct answer, the model must also draw on knowledge that is not directly present in the image. Both of these tasks present challenges for MLLMs because they require understanding, reasoning, and knowledge across both vision and text modalities.

MAS can improve accuracy in these benchmarks by delegating subtasks across models and modalities and performing step-by-step reasoning (Table 3.1). MAS have demonstrated superior performance over single-agent systems across many tasks ([47]; [77]) and offer a more transparent alternative than compound AI systems. However, establishing an effective collaboration framework can be challenging, especially when different modalities need to be aligned. Guo et al. [47] identified a notable gap in MAS research concerning multimodal settings. Addressing this gap could significantly advance the capabilities of multimodal AI systems and offer more effective and explainable solution across VL tasks.

2.2 Research Hypotheses and Questions

Current large-scale multimodal models, such as GPT-4, Claude 3, and Gemini 1.5, possess formal linguistic competence and visual understanding and already achieve high performance on VQA tasks and other VL benchmarks ([106], [143], [142]). We hypothesise that they could also generate detailed, semantically rich, and coherent paragraphs to describe images, a task for which not all models have been explicitly evaluated. Therefore, this study seeks to explore how well they perform on the task of image paragraph captioning and to analyse the linguistic characteristics of the paragraphs they generate. By evaluating these skills and identifying key challenges, this research contributes by proposing directions for datasets and metrics that better align with the capabilities of modern multimodal LLMs.

LLMs at scale also store multiple types of knowledge within their parameters ([117]; [3]) and perform well on many tasks that require external knowledge not directly present in the image (([106], [143], [142]). Therefore, we assume that multimodal LLMs possess much of the knowledge required for the OK-VQA task and explore their performance and prompt steerability toward specific formatting requirements, such as word length distributions, in a zero-shot setting. By doing so, we seek to understand the factors limiting their performance on the benchmark and explore strategies to unlock their full potential in open-ended tasks.

LLM-based MAS have been employed to improve performance in reasoning-intensive and knowledge-based VL tasks (Table 3.1). We hypothesise that a collaborative approach, incorporating text-only LLM agents at key stages, could improve results beyond the zero-shot performance of multimodal LLMs on the OK-VQA task. To establish an effective collaboration framework, we further hypothesise that: (a) MLLMs can generate linguistically strong and salient paragraph descriptions for images that contain the necessary information to answer the question, (b) that text-only LLMs can extract the correct answer by relying solely on the image description, and that (c) text-only LLMs can effectively format the final answer to meet specific formatting requirements. Validating these hypotheses would not only provide a more effective approach to the task but also offer a more transparent framework for understanding why multimodal LLMs fail to achieve state-of-the-art results on this task in a zero-shot setting. At the same time, establishing LLMs' effectiveness as formatters would also provide an easy, task-agnostic post-processing step to support open-ended tasks.

2.3 Experimental Setup

In the experimental phase of this work, we benchmark Claude as a large-scale, multimodal proxy model on two tasks: image paragraph captioning and knowledge-based VQA. For our first task, we explore image paragraph captioning, which involves generating paragraph-length descriptions for images. We prompt the model to generate paragraph descriptions for the images in a subset of the Stanford Image Paragraph Captioning Dataset in a zero-shot setting. We use the linguistic metrics established by the creators of the dataset to compare the human-written paragraphs from the dataset annotations with the model's outputs and conduct a linguistic analysis to evaluate how lengthy, detailed, semantically rich, diverse, and grammatically correct model-generated paragraphs are compared to human-annotated paragraphs.

For our second task, we focus on knowledge-based VQA and specifically, the OK-VQA dataset, an open-ended benchmark which features questions about images that draw on outside knowledge. We use the soft accuracy metric proposed by the creators of the dataset to evaluate a large-scale, multimodal model in a zero-shot setting and propose a multi-agent system that establishes collaboration between an LVLM agent and text-only LLM agents. Our framework comprises three models: the Scout, an LVLM that takes an image as input and describes it in a paragraph; the Analyser, an LLM that generates an initial answer to the question based on the image description; and the Resolver, an LLM that extracts and formats the final answer based on a set of predefined rules. We experiment with different prompting techniques, including Chain-of-Thought (CoT) prompting and few-shot prompting, and evaluate various methods to propose design elements.

2.4 Contributions

In this work, we explore the limitations of modern multimodal LLMs in handling open-ended tasks that require outside knowledge and visual commonsense reasoning. We examine the Stanford Image Paragraph Captioning and OK-VQA datasets and demonstrate that although these models possess substantial cognitive, reasoning, and linguistic capabilities, their performance deteriorates when they are managing multiple complex tasks and trying to conform to specific response formats. To address this, we propose a collaborative framework that combines a multimodal LLM with a text-only LLM in two key stages, image description and answer formatting, resulting in improved performance over the baseline. Our analysis highlights the need for new benchmarks in the areas of image paragraph captioning and knowledge-based VQA that explicitly test a model's ability to reason across abstraction levels and generate human-aligned responses.

First, we highlight the scarcity of datasets in the image paragraph captioning domain that truly challenge modern multimodal LLMs and emphasise the significance of this task as a benchmark. We conducted experiments on the most popular dataset of this area, the Stanford Image Paragraph Captioning dataset, and performed a linguistic analysis suggesting that state-of-the-art multimodal LLMs have long surpassed this benchmark. Our results demonstrate that models like Claude can generate paragraphs that are significantly more semantically rich, detailed, and diverse than the ground-truth paragraphs in the dataset, underscoring the need for a new and more challenging benchmark. We identify several areas that remain challenging for LLMs and present opportunities for further improvement. To this end, we propose directions for future datasets that could better evaluate their reasoning, knowledge, and multimodal understanding. We argue that high-quality image paragraph generation is still challenging for multimodal LLMs and relies on many of the same skills needed for VQA, hence our interest in both tasks. Therefore, with appropriately designed datasets, the image paragraph captioning task offers a promising framework for evaluating capabilities relevant to VQA.

In addition, we investigate collaborative frameworks for knowledge-based VQA and demonstrate that incorporating LLMs can improve the performance of LVLMs alone. We focus on the OK-VQA dataset because, as an open-ended benchmark, it poses a greater challenge than multiple-choice formats. Models must not only know the correct answer but also express it in a way that aligns with how humans would naturally respond. We observe that multimodal LLMs possess the knowledge and reasoning abilities required to perform well on this task. In fact, the vast majority of the time, the models understand the question and provide valid answers. The underlying problem is that they struggle to perform reasoning-intensive and cross-modal tasks simultaneously or to adhere to the required answer format or abstraction level. As a result, their answers can be overly specific, sound unnatural to humans, or be presented in a format that does not match expectations. This issue is particularly evident in open-ended tasks and is a key factor that limits the performance of models that have not been fine-tuned on the dataset. This is why existing benchmarks and evaluation metrics, such as OK-VQA and the proposed soft accuracy metric, fall short in capturing the full extent of the reasoning and cognitive abilities of modern multimodal LLMs that are prompted in a zero-shot manner.

However, understanding the appropriate level of abstraction in a question and formatting the answer as expected are capabilities we ideally want models to handle well without the need for fine-tuning. We find that models like Claude struggle to tackle complex multimodal questions, abstraction, and formatting all at once. Conversely, they are effective at describing images, while LLMs excel at finding answers hidden in text and are powerful formatters, capable of post-processing other models' outputs and converting them into the expected format. This motivates us to explore a collaborative approach to address the OK-VQA task, where an LVLM is first prompted to generate a paragraph description of the image without access to the question. Then, an LLM provides an initial answer based on the question and the generated description, followed by a second LLM that selects and formats the final answer to meet the requirements. Our collaborative framework improves performance over the baseline and demonstrates the potential of collaboration between models in tackling open-ended VQA.

We highlight that LLMs tend to default to more specific answers than humans, which significantly compromises their results in open-ended tasks. They also struggle to transition between abstraction levels and often fail to look for hypernyms when the information required for hyponyms is missing. Notably, if they can't find a specific hyponym match, they don't generalise upward like humans would. For example, when asked 'Who designed the statues?', humans might initially expect to identify a particular individual. However, upon examining the image and realising that this level of detail can't be inferred, they naturally generalise upward, providing a broader yet appropriate answer like 'an artist'. In contrast, we find that LLMs struggle with this kind of flexibility when it comes to abstraction levels. Instead, they tend to fixate on finding a precise answer. We notice that if they fail to locate the specific information they are looking for in the image, they don't generalise upward but rather choose to not answer the question at all.

Given that multimodal LLMs clearly possess much of the knowledge required for challenging multimodal tasks, their relatively low accuracy on the OK-VQA dataset cannot be attributed solely to a lack of knowledge or visual reasoning. We argue that the ability to interpret and transition between abstraction levels is a critical component of reasoning, and is a primary weakness of modern multimodal LLMs when evaluated on open-ended datasets. Through a collaborative approach, we improve accuracy over the baseline and highlight that while LLMs can effectively handle formatting, selecting the appropriate level of abstraction and reasoning at that level remains a significant challenge. Therefore, we also emphasise the need for datasets that explicitly assess models on their ability to understand and shift between different levels of abstraction.

Chapter 3

Theoretical Background

3.1 Large Language Models (LLMs)

3.1.1 Foundations

Large Language Models (LLMs) are language models with a large number of parameters, trained using self-supervised learning techniques on vast corpora of text to learn statistical patterns in language. Most modern LLMs are based on the transformer architecture [147], which models dependencies between input and output sequences through a self-attention mechanism. Pretraining is the first critical step in LM development and involves training the model on text corpora using self-supervision and pretraining objectives. The main training objective of a pre-trained LM typically consists of an objective predicting the probability of text x [84]. Common pretraining objectives include Standard Language Modeling (SLM), which trains models to predict text autoregressively, and denoising objectives, such as Corrupted Text Reconstruction (CTR) and Full Text Reconstruction (FTR), which train models to recover original text from corrupted inputs. However, as LLMs scale in size, they often exhibit emergent abilities, such as complex reasoning or in-context learning, that are not present in smaller models and were not directly targeted by their training objectives [161]. These abilities can be elicited through prompting, enabling LLMs to perform complex tasks without additional fine-tuning.

3.1.2 Capabilities

Fluency, Human-Likeness, and Empathy

With their fluency and human-like characteristics ([107]; [94]; [185]), LLMs have astounded the world, emerging as powerful tools with close proximity to humans. Thanks to their transformer architecture, LLMs can achieve high levels of fluency and linguistic competence beyond the reach of traditional statistical ML models. Transformers combine three key features that make them ideal for learning language: (i) a deep stacking of layers, which allows them to learn both low- and high-level linguistic patterns, (ii) an attention mechanism in each layer, enabling each token to selectively focus on relevant parts of the input sequence, and (iii) the parallel processing of input tokens, which allows for efficient training across vast corpora of data. Together these characteristics contribute to LLMs' ability to learn core aspects of human language and convincingly replicate it. Yet, human language is just one of several ways in which LLMs mirror human behaviour. *Human-likeness* is defined by [52] as the ability to understand dialogue context, leverage knowledge appropriately, detect user emotions and personality, and generate friendly and reasonable responses that are coherent and consistent with the dialogue context.

Similarly, *empathy*, defined by [103] as the ability to project another's feelings and ideas onto one's understanding, is expressed in empathetic systems through *emotion awareness, personality awareness, and knowledge accessibility* [94]. These qualities enable meaningful user engagement and support decision-making in applications, particularly through conversational agents, such as ChatGPT ([151]; [159]). Human-likeness and empathy also play a role in fostering trust in algorithmic recommendations—trust increases when machines emulate human characteristics and declines when they lack human-like traits like empathy [14]). Therefore, LLMs function as intuitive intermediaries that enhance trust and strengthen collaboration between humans and algorithms [47].

Reasoning and Cognitive Abilities

LLMs are also powerful reasoners and flexible, extendible, and expressive Knowledge Bases (KBs) ([117]; [5]). The ability to reason is central to human intelligence, yet machines still face challenges in drawing conclusions from given information and prior knowledge ([117]; [32]). Unlike many of their predecessors, LLMs demonstrate a strong performance on arithmetic, commonsense, logical, symbolic, and multimodal reasoning tasks [32] and can effectively formalise problems by translating natural language to mathematical models [159].

Moreover, they have significant amounts of knowledge encoded in their parameters, enabling direct retrieval through prompting and fine-tuning ([138]; [112]; [5]). These abilities can be leveraged to solve complex tasks presuming specialised knowledge, such as biomedical knowledge [132]. While traditional KBs are constructed manually and follow a rigid, predefined structure, LLMs offer a more flexible alternative that is easier to build, update, and maintain in real time ([138]; [5]). However, knowledge retrieval is not without challenges, and prompt engineering is of utmost importance to ensure the accuracy and completeness of retrieved facts. Nevertheless, with advances in factuality assessment, increased domain specialisation, and scaling in size and human feedback, LLMs are becoming increasingly reliable KBs ([48]; [167]).

Planning and Coordination

LLMs have also demonstrated remarkable abilities in reasoning, decision-making, and planning, which allow them to accomplish many tasks independently or delegate them to other components ([156]; [168]). Language lies at the heart of these capabilities; just as it was pivotal in the evolution of biological intelligence, it is now playing a similarly transformative role in the development of artificial intelligence. In this sense, language is more than a mere textual modality and functions as a vehicle for both thought and communication [136]. With this foundation, LLMs are capable of:

(i) utilising external tools such as search engines and code interpreters ([182]; [98],

(ii) decomposing complex tasks into sub-tasks ([113]; [180]; [185]),

(iii) delegating responsibilities to other models ([127]; [185]),

(iv) reasoning on complex tasks ([162]; [31]), and

(v) forming multi-agent systems spontaneously ([111]; [85]).

Thanks to these capabilities, LLMs are able to serve as coordinators in AI systems, improving performance and effectiveness ([182]; [23]).

3.1.3 Prompt Engineering

In recent years, prompt-based learning has emerged as a powerful alternative to supervised learning by eliminating the need for large amounts of labelled, task-specific data. As a result, attention has gradually shifted from the traditional 'pretrain and fine-tune' paradigm to a newer approach known as 'pretrain, prompt, predict' [85]. In this framework, a language model is first pre-trained to acquire general-purpose linguistic knowledge and then guided to perform downstream tasks through carefully crafted prompts, often without any additional fine-tuning. While compelling, this method introduces the challenge of prompt engineering, which involves identifying the most effective prompt to elicit the desired behaviour from the model and achieve the highest accuracy on the task at hand.

Formally, prompt engineering refers to the design of a prompting function $f_{prompt}(x)$ that achieves optimal performance on a downstream task [86]. The prompting function can be *static*, applying the same prompt template to every input, or *dynamic*, generating a custom template for each input. Based on their shape, prompts are grouped into two main types: *cloze prompts* [112], which involve filling in the blanks within a textual string (e.g., '______ is the capital of France.'), and *prefix prompts* [76], which require the model to continue a given text prefix (e.g., 'Describe the steps involved in solving a quadratic equation.'). Prompts can be constructed either manually or automatically. While manual prompts are intuitively written by humans, automated prompts are generated using models or rule-based strategies. These can be further categorised as *discrete* (or *hard*), in which the prompt is usually an actual text string within the discrete space of natural language, and *continuous* (or *soft*), in which the prompt is encoded directly in the continuous embedding space of the LM. Prompt engineering has grown into a dynamic field of both theoretical and practical interest, encompassing an expanding set of prompting techniques. However, the field is currently undergoing another shift. With the focus transitioning from monolithic models to multi-agent systems, collaboration among multiple models is becoming the new standard for leveraging LLMs in a zero-shot manner, achieving results that match or even exceed those of previous paradigms. These emergent collaborative behaviours of LMs suggest a need to move beyond traditional single-agent prompting techniques toward multi-agent and model-generated prompting.

Single-Agent Prompting

Numerous prompting techniques have proven effective in guiding language models to perform specific tasks, often matching or exceeding the performance achieved through fine-tuning. Additional prompting methods can be found in the comprehensive review by [121].

Zero-shot Prompting

Thanks to their large-scale pretraining, LLMs can be prompted directly to perform tasks in a zero-shot manner, without the need to include additional context, inputs, or examples. Instead, they can be prompted using only a task description, depending on the complexity of the task, the knowledge it requires, and the types of tasks the model was exposed to during training. Although zero-shot prompting might now appear to be the standard approach to prompting, it was a groundbreaking concept at the time of its introduction, as it demonstrated that LLMs could perform tasks without any explicit supervision [118].

Few-shot Prompting

LLMs are often capable of performing new language tasks after seeing only a few examples, much like humans [13]. Few-shot prompting provides the model with a small number of input-output examples to help it understand the task at hand and enable in-context learning, in contrast to zero-shot prompting, which offers no context, inputs, or examples. While effective across a variety of tasks, standard few-shot prompting has limitations, including the need for additional tokens to incorporate examples and the potential introduction of biases that may influence few-shot results.

Chain-of-Thought (CoT) Prompting

Despite their capabilities, tasks involving complex reasoning often expose the limitations of LLMs. Chain-of-Thought (CoT) prompting addresses this limitation by guiding the reasoning process of LLMs through a logical reasoning chain [160]. LLMs prompted with CoT have been shown to adopt a more structured approach to reasoning by breaking down problems and demonstrating a deeper understanding of the task, compared to those prompted using traditional methods. However, the authors argue that this is an emergent ability, observed only in sufficiently large models.

Self-Consistency

The observation that problems requiring thoughtful analysis often involve greater reasoning diversity led Wang et al. [155] to sample multiple, diverse reasoning paths from the language model's decoder using few-shot CoT. This prompting technique, known as self-consistency, boosts the performance of CoT across a range of challenging benchmarks involving arithmetic and commonsense reasoning.

Tree-of-Thoughts

Complex reasoning tasks often require exploration and look-ahead reasoning. To address this challenge, [178] and [87] expanded on CoT and proposed the Tree-of-Thoughts (ToT) framework, which introduces a tree structure of coherent language sequences serving as intermediate reasoning steps, known as thoughts. The LM model can then self-evaluate its progress and systematically explore thoughts by employing search algorithms, such as breadth-first search and depth-first search.

Graph-of-Thoughts

Despite its success, the ToT prompting technique has a key limitation: it follows a strictly linear reasoning process, unlike human thinking, which is often non-linear. Therefore, to better model the human thought process, [178] proposed Graph-of-Thought (GoT) reasoning, which adopts a graph representation where thought units are treated as nodes and the relationships between them as edges. By simulating non-sequential human thinking, the GoT reasoning model demonstrates substantial gains over the CoT baseline and sets a new state of the art in multimodal reasoning.

Retrieval Augmented Generation (RAG)

A common critique of LLMs concerns their ability to access and use knowledge effectively when tackling knowledge- and reasoning-intensive tasks. To improve capabilities and factual consistency, [70] proposed Retrieval Augmented Generation (RAG), a framework in which prompts are enriched with relevant external information, enabling LMs to generate more context-aware and factually accurate responses. RAG operates by combining the parametric memory of a pre-trained language model with non-parametric memory retrieved from a dense vector index of external knowledge sources, such as Wikipedia. Given a user query, the model retrieves relevant resources using a neural retriever and grounds its output on the retrieved context to produce more informed and accurate responses. Retrieval-augmented models generate more specific, diverse, and factual language, helping to mitigate hallucinations in LLMs and achieving strong, often state-of-the-art performance on many knowledge-intensive benchmarks.

Prompt Chaining

Prompt chaining, introduced in [114], involves identifying subtasks and prompting the LLM iteratively, by using each subtask's response as input for the next prompt. This approach allows the model to focus on one specific problem at a time, without being overwhelmed by multiple simultaneous tasks. In addition to improved performance, prompt chaining enhances the transparency, controllability, and reliability of LLM applications.

ReAct

Proposed by [177], ReAct is based on the observation that, in humans, reasoning and action are intertwined processes that drive learning and reasoning. According to the ReAct framework, LLMs are used to generate both reasoning traces, which help the model induce, track, and update action plans while handling exceptions and actions, which allow the model to retrieve information from external sources such as knowledge bases or environments. Experimental results demonstrate that ReAct outperforms several state-of-the-art baselines across a diverse set of language and decision-making tasks, while also enhancing the interpretability and trustworthiness of LLM outputs. Combining ReAct with CoT yields the best results, as it allows the model to leverage both internal knowledge and information retrieved from external sources during reasoning.

Reflexion

Reflexion is a prompting technique introduced by [128] that encourages LMs to learn from prior mistakes through verbal reinforcement. It operates through an iterative process involving three components: an actor model that performs tasks, an evaluator model that scores outcomes, and a self-reflection model that generates verbal reinforcement cues to guide the actor's improvement. This cycle helps the model improve over time, making Reflexion especially effective across various tasks, including sequential decision-making, coding, and language reasoning.

Multi-Agent Prompting

Multi-agent prompt engineering is a relatively new area of interest, referring to the design of multiple prompting functions $f^i_{prompt}(x)$ corresponding to multiple language agents that collaborate on a downstream task. Multi-agent prompting inherently builds upon single-agent prompting techniques, while also introducing strategies that are unique to multi-agent settings. This section explores how each single-agent prompting strategy discussed in the previous section relates to, or has been adapted for, multi-agent systems. In addition, it presents new prompting strategies introduced specifically for multi-agent settings. The prompting strategies used in various MAS applications for the collaboration of different language agents on reasoning tasks are presented in Table 3.1.

MAS Zero-shot Prompting

Zero-shot prompting has been used in MAS applications primarily to explore the extent to which agent collaboration can match or outperform both the 'pretrain and fine-tune' and 'pretrain, prompt, predict' paradigms. In fact, LLM-based MAS are capable of performing complex tasks in a zero-shot manner, without the need for additional context, inputs, or examples, and often achieve better results than a single fine-tuned model or a single prompted model.

MAS Few-shot Prompting

Few-shot prompting has been employed for few-shot learning within MAS pipelines, and has been shown to contribute to performance improvements. However, to the best of our knowledge, few-shot prompting has not yet been used explicitly to teach agents how to interact with one another. Given its ability to convey patterns through demonstration, few-shot prompting could potentially be a valuable tool for guiding inter-agent communication through examples of meaningful collaboration.

MAS Chain-of-Thought (CoT), Tree-of-Thought (ToT), and Graph-of-Thought (GoT) Prompting

In single-agent settings, CoT prompting better simulates human reasoning than zero-shot prompting, as it encourages step-by-step problem solving through a logical reasoning chain. ToT takes this a step further by encompassing branches of possible reasoning paths, thereby more closely simulating the exploratory and lookahead reasoning that humans often engage in during complex problem solving. To account for the non-linear and interconnected nature of human reasoning, GoT adopts a graph-based representation in which thoughts are modelled as nodes and the relationships between them as edges. The former three are all part of a growing family of prompting techniques aimed at improving reasoning by framing it as a structured, multi-step process. Multi-agent systems are inherently linked to CoT, ToT, and GoT, as they explore different reasoning paths by distributing cognitive processes across multiple agents. We hypothesise that a system composed of multiple models is also more aligned with human reasoning than a single component, as different functions in the brain are managed by specialised regions and pathways [11]. Therefore, multi-agent prompting techniques are inherently connected to, and could potentially expand upon, single-agent prompting strategies aimed at improving reasoning.

MAS Self-Consistency

In single-agent settings, self-consistency builds upon CoT by generating multiple reasoning paths and selecting the most frequently occurring answer among them. In MAS, a similar form of self-consistency can be achieved by distributing reasoning paths across multiple agents and then aggregating them, either through consensus, voting, or by delegating the final decision to an orchestrator agent. We hypothesise that self-consistency is well-suited to MAS, where different modalities, models, and roles come together to integrate diverse reasoning paths. In fact, a recurring theme in the literature is that the effectiveness of MAS is closely tied to agent diversity, which is aligned with self-consistency improving results for single-agent setting.

MAS Prompt Chaining

Prompt chaining works by decomposing a task into subtasks and prompting the language model iteratively, using the output of each subtask as input for the next. Prompt chaining naturally relates to MAS, as they both involve task decomposition and sequencing. However, in MAS, different agents can be assigned different subtasks, enabling both parallelisation and specialisation. Therefore, we find that MAS inherently implement prompt chaining through role assignment and task decomposition, and in some cases, may do so more effectively than a single agent.

MAS Reflection

The idea behind reflexion is self-critique and iterative improvement through verbal reinforcement. Whereas in single-agent settings reflexion is typically an internal mechanism, in multi-agent systems it can originate both from within an agent (self-reflection) and externally through feedback from other agents (external-reflection). External reflection may be more useful than internal reflection in certain contexts, as self-prompting tends to be generic and susceptible to bias, whereas external reflection can be more context-aware, targeted, and capable of challenging the model's initial errors.

MAS are able to implement prompting techniques used for single-agent settings. In fact, the reason prompting techniques such as CoT and Reflexion yield better results than zero-shot prompting in single-agent settings may also explain why MAS outperform single models in zero-shot contexts—because MAS implicitly internalise specific prompting strategies through their structure. That is to say, MAS labeled as zero-shot are not truly so, as they inherently implement prompting techniques already known to enhance performance. This complicates evaluation, making it difficult to draw clear comparisons between 'zero-shot' MAS and truly zero-shot single-agent baselines.

Other prompting techniques also exist that are specifically designed for MAS settings. For example, a common scenario in MAS collaboration is role-playing, according to which agents assume specific roles and are assigned respective tasks. Inception prompting implements role-playing by assigning roles and tasks to the language models at the beginning of the session [71]. It comprises task specifier prompts, which define the task, decompose it into subtasks, and delegate them to individual models, as well as system prompts, which instruct models to adopt specific roles (e.g., 'junior') and adhere to predefined rules (e.g., 'As a junior, your role is to assist, and you should not override decisions made by the senior.'). Role-playing requires minimal human intervention and is an effective way to establish collaboration between agents. However, models cannot always be steered through prompting to adhere to their assigned roles. Phenomena such as role flipping may occur, where models break character and deviate from their designated behaviour. Studies also highlight the importance of role diversity, noting that the use of identical role descriptions can lead to significant performance degradation ([49], [16], [174]). An alternative to inception prompting proposed by [25] is employing an agent as a recruiter, similar to a human resource manager, that gathers the most suitable group of agents and delegates tasks.

3.1.4 Challenges and Ethical Concerns

Technical and Theoretical Challenges

Lack of True Understanding and Reasoning

While LLM foundation models appear to perform well on many reasoning tasks, it remains unclear whether this reflects true reasoning or simply the reproduction of surface-level patterns learned during pretraining. The debate between reasoning and memorisation is deeply existential for humans and rooted in cognitive biases that influence how we perceive and interact with machines. Before the era of language models, and particularly LLMs, the use of language was generally assumed to reflect underlying thought. However, the tendency to associate language with thought has proven to be a logical fallacy that humans project onto machines [95]. In fact, the post-LLM era has taught us that strong language abilities do not necessarily indicate strong reasoning capabilities, and conversely, poor reasoning abilities do not necessarily imply poor language use. To dissociate language from thought, [95] describe two kinds of linguistic competence: *formal linguistic competence* refers to the knowledge of linguistic rules and statistical regularities, while *functional linguistic competence* refers to the ability to use language effectively in real-world situations, often drawing on non-linguistic capacities. The two types of competence correspond to distinct processes and paths in the human brain.

Viewed through this lens, it becomes evident that LLMs exhibit formal linguistic competence but lack functional linguistic competence ([95]; [69]). The fluency and near-human grammatical abilities of modern LLMs have been empirically and systematically evaluated through benchmarks. These models have achieved remarkable results not just on general NLP tasks but also on challenging benchmarks of linguistic competence, including the BLiMP and the SyntaxGym ([158]; [41]). LLMs' strong performance on these datasets is largely attributed to their transformer architecture, enabling them to extract complex linguistic patterns from vast amounts of data.

However, foundational LLMs lack equivalent functional linguistic competence, which, in real-world settings, typically requires four key capacities to sustain even the simplest of conversations: (i) **formal reasoning**, which includes abilities such as logical, mathematical, and relational reasoning, problem solving, as well as computational thinking (ii) **world knowledge**, referring to commonsense understanding of the physical, social, and conceptual world, (iii) **situation modelling**, the ability to track protagonists, locations, and events as the conversation unfolds, and (iv) **social reasoning**, which involves interpreting language within its social context [95]. LLMs have been widely criticised along these dimensions ([50]; [184]; [170]; [88]; [154]; [166]; [123]; [124]; [140]; [125]; [183]), and although newer versions show a markedly improved capacity to handle complex language tasks, a bottleneck remains that foundational models have yet to overcome. However, this is becoming increasingly difficult to assess as multi-component AI systems become the industry norm.

The central issue, then, is understanding why LLMs sufficiently achieve formal linguistic competence yet continue to fall short in functional linguistic competence. At a surface level, one might argue that this is simply because the training objective of transformer-based architectures is the prediction of the next or masked word in a sequence, not performing calculations, detecting user emotions, or having a notion of death. LLMs are trained to produce human-like language, not to talk, think, and act like humans [95]. Nonetheless, they have previously demonstrated the ability to learn patterns not directly incentivised by their training objectives. This raises the deeper question of which intrinsic properties make certain linguistic features more learnable by LLMs while others remain elusive. Or perhaps it is a matter of how certain properties and patterns of the world are represented in the large internet corpora used to train LLMs, rendering some more easily learnable than others. These questions are essential for determining what expectations are reasonable to place on foundational language models and for understanding when we ought to seek out alternatives or complementary components to achieve specific goals.

Knowledge Gaps and Lack of Domain Specificity

Critiques of foundational LLMs often point to their knowledge gaps and lack of domain specificity. LLMs have been tested on a range of knowledge-intensive benchmarks, including MMLU[1], TriviaQA[2], TruthfulQA[3], CommonsenseQA[4], VCR[5], OK-VQA[6], and A-OKVQA[7], and have often been found to perform inadequately,

[1] https://paperswithcode.com/dataset/mmlu
[2] https://paperswithcode.com/dataset/triviaqa
[3] https://paperswithcode.com/dataset/truthfulqa
[4] https://paperswithcode.com/dataset/commonsenseqa
[5] https://visualcommonsense.com/
[6] https://okvqa.allenai.org/
[7] https://paperswithcode.com/dataset/a-okvqa

particularly when multi-hop or multimodal reasoning and explicit knowledge is required. From a technical standpoint, this issue stems from limitations in training data coverage, accuracy and currency of information, as well as the lack of domain-specific fine-tuning ([83]; [149]). The vast majority of LLMs are trained on large-scale text corpora scraped from publicly available internet sources. Although large, these corpora are imperfect and often underrepresent certain domains such as medicine, law, and specialised scientific fields. In essence, LLMs cannot learn what they are not exposed to during training. Similarly, LLMs tend to learn what they are repeatedly exposed to. Thus, if the training data includes frequent instances of misinformation (e.g., climate change denial), common misconceptions (e.g., the belief that humans have only five senses or use only 10% of their brains), or harmful stereotypes (e.g., that women belong in the kitchen), such content may be internalised by the model and reproduced in its outputs. The absence of accurate, complete, up-to-date, and specialised knowledge is problematic not only because it compromises performance, but also because uncertainty makes LLMs more prone to hallucinate [120].

The challenge of data currency has been substantially mitigated in recent LLMs such as GPT-4[8], which incorporate real-time information access through integrated web browser plugins ([182]; [171]). Retrieval-Augmented Generation (RAG) has also emerged as a promising solution to address knowledge gaps, the need for up-to-date information, and the lack of domain specialisation by incorporating content from external knowledge databases [39]. Fine-tuning of pretrained models with human feedback has also been widely used to align the model with human values and preferences. Task-specific tuning and RAG have become the gold standard for LLM-based applications, ensuring robustness, improved factual accuracy, and better alignment with domain-specific requirements ([39]; [56]; [7]; [134]). However, from a more theoretical perspective, the problem can be traced back to surface-level learning and the broader debate between learning and memorisation. LLMs often give the impression of thinking and learning, yet in reality, they identify and reproduce statistical patterns in language rather than acquiring structured or grounded understanding of the real world. Therefore, the issue might be theoretical in nature, but it can still be effectively addressed through technical means.

Hallucinations

Rawte et al. in [120] define hallucination in foundational models as the generation of content that is not based on factual or accurate information. This manifests specifically through outputs containing fictional, misleading, or entirely fabricated claims, rather than reliable and truthful information. An *intrinsic* hallucination is a model output that contradicts its source inputs, representing a failure of faithfulness. Conversely, an *extrinsic* hallucination is a model output that contradicts world knowledge, representing a failure of factualness [57]. Hallucinations in large foundational models constitute a cross-modal challenge, affecting text, image, audio, and video generation alike. As noted by [120], these fabricated outputs pose a significant challenge due to their plausibility and the confidence with which they are stated, making them difficult to detect despite being factually incorrect or entirely fabricated.

Blind reliance in LLMs can compromise decision-making and trigger a cascade of misaligned outputs, as hallucinations tend to propagate across the entire interaction trajectory [187]. Hallucinations can also severely harm trust in algorithmic predictions. People hold algorithms to a *perfect automation schema*, expecting them to perform perfectly and quickly losing trust after seeing them perform and err ([61]; [29]). As a result, when encountering non-factual LLM outputs that combine assertive delivery with plausibility, suspicion is created about undetected errors in prior and future interactions, thereby reinforcing *algorithm aversion*—humans' inherent reluctance to trust algorithms ([29]; [61]).

Hallucinations in LLMs are attributed to multiple interconnected factors and are tied to persistent limitations in the field, including outdated knowledge, bias, limited reasoning capabilities, and domain-specific knowledge gaps ([120]; [149]). Surprisingly, LLMs can demonstrate latent knowledge of correct information yet still produce confident hallucinations—generating and committing to false answers they later reject when queried in a different session [187]. In this scenario, the knowledge exists within the system's parameters, but either the model struggles to retrieve it and use it appropriately or prompt engineers fail to probe it. For instance, prompts instructing the model to first answer a question and then provide an explanation may pressure it to

[8]https://cdn.openai.com/papers/gpt-4.pdf

commit to a wrong answer and construct justifications around it, rather than the other way around ([187]; [9]; [2]).

Therefore, understanding hallucinations and applying prompt engineering and other approaches to handle it effectively is a key consideration when it comes to LLM research. Current work on hallucinations primarily spans three directions: detection ([79]; [100]; [46]), factuality ([48]; [149]; [21]), and prevention ([191]; [100]; [46]). Standardised hallucination evaluation benchmarks include TruthfulQA (Question Answering), FactualityPrompt (Text Completion), FActScore (Task Instructions), KoLA-KC (Task Instructions), HaluEval (Question Answering & Task Instructions), FACTOR (Text Completion) ([187]; [149]; [79]). Object hallucination, which involves generating captions or responses that mention objects inconsistent or entirely absent from the visual content, is also a particularly important area of focus in computer vision and LVLMs due to its prevalence and persistence, its impact on user trust, and its potential for adverse consequences in high-stakes applications such as medical image analysis ([79]; [191]; [46];[74]). Hallucination research has progressed considerably, with advancements such as factuality checking and object hallucination detection showing promising results in mitigating their impact on tasks, yet it remains unclear whether hallucinations are a technical or a more fundamental theoretical issue.

Agnosia

Agnosia in Multimodal Large Language Models (MLLMs) is a concept closely related to hallucinations and was first introduced by Lu et al. in [89] to mirror the neuropsychological phenomenon in which individuals are unable to correctly process sensory inputs or recognise elements such as objects, colours, or spatial relations [131]. Agnosia in MLLMs refers to instances where the model misinterprets visual inputs or fails to comply with textual instructions even in straightforward cases, resulting in irrelevant outputs, errors, or ungrounded assertions. The authors of the paper reveal the deficiencies of MLLMs in interpreting multimodal inputs and observe six distinct types of agnosia—**Entity**, **Number**, **Colour**, **Material**, **Action**, and **Spatial**—yet additional types may be identified as the field develops [89]. Recognising agnosia as a critical challenge in MLLM research, Lu et al. in [89] formalistically define the phenomenon and propose a comprehensive framework that evaluates its presence across MLLMs and mitigates its impact by leveraging a multimodal instruction tuning method. Similar to hallucinations, the presence of agnosia might be a technical as well as theoretical issue and more research is needed to better understand its underlying causes and develop more effective mitigation strategies.

Steerability

Steerability is defined as the extent to which a model can be guided along a specific dimension [99]. Similarly, models are *steerable* when they can be easily made to adopt various behaviours, such as specific personas, tones, or content styles [133]. Chang et al. in [17] provide a formal definition of steerability as the product of two components: sensitivity and directionality. *Sensitivity* measures how much the model moves in goal-space in response to a user's request, while *directionality* captures how well the model's output aligns with the user's intended direction in goal-space. For a model to be considered steerable, it must exhibit both high sensitivity and high directionality. *Prompt steerability* refers specifically to the extent to which a model's behaviour can be guided through prompting alone [99].

Measuring steerability can be challenging, as it requires understanding the user's goals and differentiating between a model's inability to generate a given output and its inability to be steered towards it [146]. Although steerability benchmarking is a relatively new area of research, Vafa et al. [146] proposed a comprehensive framework and benchmark for evaluating general model steerability, while Miehling et al. [99] introduced a benchmark specifically designed to assess prompt steerability. Findings from these benchmarks indicate that, despite their ability to produce high-quality outputs and perform well on other tasks, many current models exhibit limited steerability ([99]; [146]). In fact, Chang et al. [17] argue that modern LLMs are, in general, not steerable.

However, despite steerability generally being regarded as a desirable property within the LLM community, it is not intrinsically positive or negative, and its effects on human–AI interaction can be either beneficial or harmful, depending on the application context and the user's intentions or goal. This, according to [17], is due to the fact that what is considered safe in one context may not be safe in another. Steerability is beneficial when it enables personalised responses—such as tailoring fertility advice to a user's specific situation—but potentially harmful when it allows users to bypass safety mechanisms or redefine critical concepts like medical terminology ([17]; [44]). It is also possible that we might want the same model, even within the same use case, to be steerable in generating personalised outputs—such as adapting advice based on a user's specific fertility concerns—while remaining unsteerable in ways that could compromise accuracy, such as incorporating unverified treatments or misinformation into its responses. Therefore, an ideal LLM should exhibit context-aware steerability by adjusting how much it can be influenced based on the specific use case and calibrating its responsiveness in accordance with what is considered ethical and appropriate in each context.

Ethical Concerns

LLMs amplify existing ethical concerns about AI due to their unprecedented power and widespread accessibility, while also introducing novel challenges. Key ethical concerns in AI include bias and fairness, transparency, privacy and data security, and accountability and governance. In addition to these concerns, the era of generative AI introduces further ethical challenges, including misinformation and disinformation, censorship, intellectual property and plagiarism, various forms of abuse such as hate speech and cyberbullying, as well as significant environmental impact, among others ([60]; [129]; [35]; [38]; [188]). This chapter focuses on the topics most relevant to the experimental section of the thesis, though all issues remain important.

Bias and Fairness

Bias in the context of NLP can generally be defined as the presence of systematic misrepresentations, attribution errors, or factual distortions that result in the favouring of certain groups or ideas, the reinforcement of stereotypes, or the reproduction of inaccuracies based on patterns learned during training [35]. Common types of bias in LLMs include demographic, cultural (or social), linguistic, temporal, ideological or political, and confirmation bias, as outlined by Ferrara in [35]. These forms of bias originate from multiple sources, including training data, algorithms, labelling and annotation practices, product design, and policy decisions [35].

In the context of generative AI, discussions of bias primarily refer to social bias. In traditional NLP, social bias is a subjective term referring to unequal treatment or outcomes among social groups that result from longstanding and systemic power imbalances. Gallegos et al. [38] expanded the definition of social bias and fairness in the context of LLMs by identifying two key types of social harms that may arise: *representational harms*, which involve denigrating or subordinating attitudes towards social groups, and *allocational harms*, which refer to the unequal distribution of resources or opportunities across social groups. Representational harms encompass misrepresentation, stereotyping, disparate system performance, derogatory language, and exclusionary norms, while allocational harms include both direct and indirect forms of discrimination. To encapsulate fairness in the context of LLMs, the authors formalise a set of fairness desiderata consisting of Fairness Through Unawareness, Invariance, Equal Social Group Associations, Equal Neutral Associations, and Replicated Distributions.

Completely eliminating bias from generative language models may be impossible, as human language itself is a reflection of society and inherently contains various biases, stereotypes, and assumptions. Drawing on [35], it can be argued that removing these elements could risk losing the rich, context-dependent, and culturally embedded nature of language as we know it. Moreover, the concept of bias is highly subjective, as values and norms vary significantly across individuals, communities, and historical periods. However, efforts should be made to mitigate harmful forms of bias, particularly social and ideological or political bias. A wide range of bias mitigation techniques for LLMs are identified in the literature, and [38] categorise them according to the stage at which they intervene in the model pipeline: pre-processing, in-training, intra-processing, and post-processing. Mitigation strategies for bias in LLMs span several stages of the model pipeline. Pre-processing techniques focus on modifying training data or prompts, such as through data augmentation or filtering to reduce biased patterns. In-training methods adjust the model itself, for instance by modifying the loss function or updating select parameters. At the intra-processing stage, techniques like decoding strategy modification alter model behaviour at inference without further training. Finally, post-processing approaches such as rewriting detect and revise biased outputs after generation.

Transparency

Transparency in an AI model refers to the ability to understand its training process and how it makes decisions. By nature, LLMs are notoriously opaque, due to their deep stacking of layers and transformer-based architecture, which enables training on vast corpora of data. This complexity, albeit central to the success of generative AI models, poses major challenges for explainability—the ability to explain or present model behaviour in a manner that is understandable to humans [188]. Although what constitutes a human-understandable explanation remains inconclusive, approaches to explainable AI (XAI) are generally classified into two categories: interpretable ML models and post-hoc interpretation techniques [45]. Statistical ML techniques such as linear regression, logistic regression, and decision trees fall under the category of interpretable ML, as they are inherently designed to learn from data in a transparent way. In contrast, deep learning models, particularly LLMs, are not interpretable by design, as their internal workings are largely opaque. Therefore, post-hoc interpretation techniques are required to approximate explainability in LLMs.

Modern benchmarks such as VCR[9] and OK-VQA[10] incorporate explainability by requiring models not only to answer a question but also to provide a rationale justifying their choice, either in multiple-choice or free-text form. This approach falls under the category of output-based explanations. While this approach represents a step forward in transparency, these types of output explanations generated by the model can be misleading. A model may arrive at the correct answer for the wrong reason or pair an incorrect answer with a seemingly appropriate rationale. Interestingly, even when a model chooses the correct answer and offers an aligned rationale, this does not necessarily indicate true reasoning. For instance, a model may associate certain words—like the co-occurrence of ‘teacher’ and ‘school’—and default to the correct answer through learned statistical patterns rather than reasoning. A similar issue could occur in high-stakes medical applications. Consider a multimodal LLM reviewing a chest X-ray of the marginal case of a 19-year-old Hispanic woman with public health insurance. The model reports no findings, citing clear lungs and low clinical risk due to her age. The explanation appears reasonable, but in reality, the model overlooks a subtle lung lesion because it has learned to associate this demographic with a lower likelihood of disease. Here, the rationale hides a statistical bias rooted in underrepresentation in the training data, a phenomenon previously reported in medical datasets such as MIMIC-CXR[11] and the CheXpert[12].

Therefore, there is a danger that LLMs might not disclose their actual reasoning and may instead offer explanations that align with their own motives or what they believe users want to hear. In fact, they have been shown to lie deliberately to appear more favourable, particularly when they recognise they are being evaluated ([10]; [122]). The expectation that LLMs should provide accurate rationales may simply reflect our assumption that humans would, revealing yet another bias we project onto machines. General trends in benchmarks with rationales may be useful to observe, but it is important to keep in mind that in high-stakes applications like medicine, the use of words like ‘because’ or ‘as such’ does not necessarily reflect the model’s actual reasoning. These phrases can create the appearance of explanation without revealing the true basis for a decision.

3.2 Vision-Language Models (VLMs)

3.2.1 Foundations

Vision-Language Models (VLMs) are models trained to understand and generate information across visual and textual modalities, learning aligned representations through image-text datasets [58]. Early VLMs, such

[9] https://visualcommonsense.com/
[10] https://okvqa.allenai.org/
[11] https://physionet.org/content/mimic-cxr/2.1.0/
[12] https://stanfordmlgroup.github.io/competitions/chexpert/

as CLIP and ALIGN, used contrastive objectives to map images and text into a shared embedding space ([119]; [58]). More recently, Large Vision-Language Models (LVLMs) have emerged, integrating transformer-based architectures to allow pretraining on large multimodal corpora [4]. Prompting techniques, such as few-shot prompting with task-specific examples, can unlock complex multimodal reasoning abilities. Models like BLIP-2 [73] bridge VLMs and LLMs by combining frozen LLMs with vision encoders. This transition culminates in multimodal LLMs like GPT-4 [106], capable of processing visual and textual content at the same time.

3.2.2 Tasks and Challenges

Traditional vision tasks such as object detection, image classification, and semantic segmentation remain long-standing challenges in the computer vision community. Beyond these tasks, the era of LVLMs and multimodal LLMs has shifted the emphasis toward more challenging tasks that require joint understanding, reasoning, and knowledge across both visual and textual modalities. Some of these tasks, which are relevant to this thesis, are summarised below.

- **Visual Question Answering (VQA)**: In Visual Question Answering (VQA), a model is provided with an image I and a textual question Q about that image, and is tasked with predicting a textual answer A, either as free-form text or by selecting from a list of predefined options [92]. While the former is framed as a classification task, the latter falls under the domain of language generation. VQA is a challenging multimodal task, requiring both a detailed understanding of the image and advanced reasoning to produce correct answers [1].

- **Knowledge-based Visual Question Answering (K-VQA)**: In addition to visual understanding and reasoning, certain questions rely on knowledge outside the image. In K-VQA, a model is provided with an image I, a textual question Q about that image, and is tasked with predicting a textual answer A. To produce the correct answer, the model must incorporate knowledge K, which is not directly present in the image itself. There are two main approaches to incorporating knowledge in the K-VQA task: employing knowledge embedded in the model's parameters, and retrieving an external knowledge source such as a database or knowledge base. Famous datasets in this task include K-VQA[13], OK-VQA[14], A-OKVQA[15], WebQA[16], and FVQA[17].

- **Visual Commonsense Reasoning (VCR)**: Visual Commonsense Reasoning (VCR) involves predicting an answer A and a corresponding rationale R, given an image I and a question Q about the image, which typically requires commonsense knowledge to answer correctly. Commonsense knowledge refers to the abstract entities, facts, and real-world events that are accepted by the majority of people about everyday life [92]. The task is structured in two stages: first predicting the correct answer ($Q \rightarrow A$), and then selecting the appropriate rationale that justifies it ($QA \rightarrow R$).

- **Image Captioning (IC)**: Image Captioning (IC) is the task of generating a caption c for an image I. While IC still presents challenges for traditional computer vision systems, certain VLMs, multimodal LLMs, and collaborative approaches, such as the CLIP-GPT-2 combination, demonstrate strong performance on IC tasks. VQA and IC are closely related, particularly in the context of LLMs. Because unimodal LLMs are restricted to textual input, when evaluated on VQA datasets such as VCR[18] or OK-VQA[19], they rely solely on captions, provided by the dataset or generated by other models. However, most training and testing IC datasets, such as COCO[20] and Flickr[21], consist of image-caption pairs that do not directly require external knowledge. As a result, captions often omit important visual information, significantly restricting LLMs' ability to answer correctly.

- **Image Paragraph Captioning (IPC)**: A single-sentence description often fails to capture the full

[13] https://paperswithcode.com/dataset/kvqa
[14] https://okvqa.allenai.org
[15] https://github.com/allenai/aokvqa
[16] https://webqna.github.io/
[17] https://github.com/wangpengnorman/FVQA
[18] https://paperswithcode.com/dataset/vcr
[19] https://okvqa.allenai.org
[20] https://cocodataset.org/#home
[21] https://paperswithcode.com/dataset/flickr30k

picture. As an alternative, Image Paragraph Captioning (IPC) has been proposed, where the goal is to generate a coherent, multi-sentence paragraph P that describes the image I. Paragraphs convey significantly more information about images compared to single-sentence captions. Krause et al. [65] observe that image paragraphs tend to use more verbs and pronouns, slightly fewer nouns, and a similar frequency of adjectives compared to captions. This reflects their richer structure, as paragraphs describe not just objects, but also their properties and relationships. They also note that complex linguistic features like coreference, which are characteristic of human-like language, cannot not be captured by single-sentence captions. Given these complexities, paragraph generation is more linguistically challenging than sentence-level image captioning.

Key challenges in VLM research include hallucination, safety, fairness, multimodal alignment, commonsense and physical reasoning, training efficiency, and data scarcity [80];[43]. Object hallucination, the mention of nonexistent objects in images, remains a pervasive issue in VLMs [93], often stemming from misaligned contextual representations across modalities. As such, hallucination detection and mitigation become critical considerations when it comes to VLM research. Another key consideration is safeguarding against unethical use and ensuring fairness toward marginalised groups. LVLMs continue to struggle with commonsense and physical reasoning and present technical challenges such as training and fine-tuning efficiency, as well as the scarcity of high-quality training data. VLM research is evolving rapidly, with increasingly sophisticated applications, such as the collaboration between VLMs and LLMs, poised to address these challenges.

3.2.3 Enhancing VLMs with LLMs

A number of studies have explored the use of knowledge bases, and specifically LLMs to enhance the capabilities of VLMs ([22];[149];[59];[51];[110];[175];[186];[92];[91]; [67];[8];[66];[135]). The complementary nature of LLMs and VLMs allows for multimodal interaction, where language informs perception and perception grounds language. In the human mind, language and perception are two central cognitive systems, fundamental to understanding the full spectrum of human behaviour [148]. The interplay between linguistic and perceptual information during processing is complex and bidirectional, with language influencing the interpretation of visual input and perception informing the understanding of language. Similarly, visual reasoning in machines relies on multimodal perception.

Although VLMs incorporate both language and vision components, their internal architectures and training objectives can vary. This suggests that not all VLMs are created equal, and that there is a certain level of complementarity even among different VLMs [22]. Furthermore, LLMs tend to exhibit superior performance in NLP tasks than VLMs, owing to their scale, architecture, and pretraining. To achieve complementarity across models and modalities, we are transitioning away from a single, large-scale generalist model intended to serve all purposes, and instead focusing on integrating smaller, specialised components. Ensemble methods have shown strong potential, but are ill-suited for heterogeneous models with divergent inputs and outputs [22]. In response, research is increasingly turning to multi-agent collaboration, made possible by the coordination and reasoning capabilities of LLMs.

In visio-linguistic tasks regarding QA, LLMs play a central role in enabling complementarity between components. In [22], LLMs are used to integrate and reason over VLM outputs, boosting performance in complex tasks like VQA, knowledge-based VQA, visual entailment, and spatial reasoning. Wang et al. [149] also presented a multi-agent framework for VQA, where an LLM acts as a Seeker agent—posing sub-questions to the VLM agent, identifying relevant issues, and constructing a knowledge base to guide reasoning. Jiang et al. in [59] introduced an adaptive pipeline in which a multimodal LLM, such as GPT-4, interacts step by step with a standard LLM like GPT-3.5, with the objective of analysing reasoning shortcomings, filling in missing information, and avoiding overconfidence. Overconfidence is a key factor that leads models to produce incorrect answers, even when they have the reasoning and cognitive capabilities to answer correctly. Hu et al. [51] approach the knowledge-based VQA task through a multi-agent framework composed of three LLM-based agents that simulate hierarchical team roles: Junior, Senior, and Manager. Each agent independently plans actions using a planner, invokes external tools, and contributes to the final decision through voting. The cited works surpass VLM architectures alone and achieve state-of-the-art performance on standardised VQA benchmarks, including OK-VQA.

These LLM-VLM collaboration frameworks have also demonstrated substantial capabilities in tasks involving multi-step reasoning and planning. Yang et al. [175] explore a novel agent-enhanced collaborative approach to image captioning, using an LLM as an orchestrator for a VQA model. To this end, the task is divided into a series of interconnected question-answer subtasks, demonstrating that a collaborative approach can effectively break down complex tasks into smaller, more manageable ones. Zhang et al. [185] further show that LLMs can work together to assist in planning and executing long-horizon tasks, as evidenced by the ThreeDWorld Multi-Agent Transport (TDW-MAT) benchmark. Additionally, they highlight how LLMs can enhance cooperative abilities and communication between agents, as seen in the Communicative Watch-And-Help (C-WAH) environment. These studies demonstrate that LLM-driven collaborative frameworks can facilitate task planning and execution. Furthermore, task decomposition inherently constitutes a form of few-shot CoT, which improves reasoning performance through its structured problem-solving approach [160]. Therefore, through planning and decomposing tasks, LLMs can turn complexity into performance gains.

3.3 From Monolithic Models to Multi-Agent Systems

LLMs have demonstrated outstanding performance across a diverse range of tasks. Unlike earlier ML models, which operated independently, today's LLM-based systems integrate multiple specialised components, including retrievers, code interpreters, and external tools, to perform more complex tasks [182]. This transition has led to the development of LLM agents, capable of autonomous decision-making and collaborative problem-solving, either independently in Single-Agent Systems or in coordination with other agents in Multi-Agent Systems [77]. LLMs are at the forefront of these breakthroughs, driving the capabilities of compound AI and Multi-Agent Systems, acting as agents, coordinators, and optimisers. Compound AI and Multi-Agent Systems have achieved remarkable success across various domains. However, despite these advances, several challenges remain in optimising these systems and understanding the dynamics of human-agent collaboration. This chapter explores the emerging trends in compound AI and LLM Multi-Agent Systems, focusing on the opportunities they present and the challenges that must be addressed to unlock their full potential.

3.3.1 Background

Recently, there has been a significant shift in both research and industry from monolithic AI models to compound AI systems. Compound AI systems are defined by [182] as systems that tackle AI tasks using multiple interacting components, including multiple calls to models, retrievers, or external tools. In contrast, AI models are simply defined as statistical models (e.g., transformers that predict the next token in text). State-of-the-art performance is increasingly driven by compound AI systems, with most well-established implementations today comprising multiple components. LLMs in particular, through natural language prompting, have unlocked unprecedented flexibility in these systems' capabilities. For instance, ChatGPT[22] integrates a number of specialised modules, including a web browser plugin, a code interpreter, and the DALL·E[23] image generator, all orchestrated by an LLM at the core of the system [106].

Compound AI systems in the LLM era are largely synonymous with AI agents, and [62] do not explicitly differentiate the two terms. Ferber et al. [34] offer a foundational definition of agents as *a physical or virtual entity capable of acting, perceiving its environment (albeit partially), communicating with others, exhibiting autonomy, and possessing the skills necessary to achieve its goals and tendencies.* A Multi-Agent System (MAS), according to [34], *comprises an environment, objects, and agents (with agents being the sole actors), along with the relations among these entities, a set of operations they can perform, and the changes in the universe over time due to these actions.* AI agents are still being defined, but they are generally understood by the community as AI systems capable of autonomously performing tasks on behalf of a user or another system, as opposed to traditional hard-coded logic.

Kapoor et al. [62] adhere to the general notion that agency is a continuum, rather than a binary, identifying several factors that influence the system's level of agency. AI systems are more agentic when the environment is complex, such as those with a broad range of tasks or multiple stakeholders ([126]; [37]). Furthermore, systems that independently pursue complex goals without explicit instructions are also regarded as more agentic ([126]; [15]; [37]). User interface and supervision play a crucial role as well—AI systems that can be

[22] https://chatgpt.com
[23] https://openai.com/index/dall-e-2/

directed using natural language and act autonomously on the user's behalf display more agency ([126]; [15]; [37]). Lastly, systems incorporating design patterns like tool use or planning, as well as those with a dynamic control flow driven by an LLM, exhibit more agentic behaviour ([164], [18], [102]).

LLMs have demonstrated remarkable abilities in reasoning, decision-making, and planning, which allow them to accomplish many tasks independently or delegate them to other components ([151]; [168]). In light of these capabilities, LLM-based agent systems have become a dynamic field, with LLMs used in both Single-Agent Systems (LLM-SA), where one LLM handles planning or decision-making, and Multi-Agent Systems (LLM-MA), where multiple LLMs or other agents collaborate [47]. LLM-SA systems have demonstrated significant cognitive abilities [137], while LLM-MA systems build on this by combining specialised LLMs and enabling interactions among them, which allows for more effective simulation of complex real-world environments [47]. Notable applications of LLM-based MAS include improving factuality [31], enhancing medical reasoning [141], and simulating macroeconomic activities [75].

As members of this multi-agent society, LLMs achieve their effectiveness by mimicking human-like collaborative intelligence [186]. LLMs do not just resemble humans' fluency: they are empathetic, human-like dialogue systems capable of emulating societies, individual traits, and thought processes, closely resembling human behaviour ([186]; [52]; [94]). LLM agents can be easy-going or overconfident, self-reflective or argumentative, conforming or dissenting. Notably, within close-knit groups, they often prioritise harmony and consensus over objective evaluation of differing opinions, sliding into phenomena like 'groupthink' [55], much as humans do. These traits form the foundation of the human-like collaborative capabilities of LLMs and highlight the potential of LLM-based systems in solving problems by emulating human intelligence.

3.3.2 The Potential of Compound AI and LLM-based MAS

Compound AI systems are shaping the industry as popular consumer products, outperforming both humans and specialised models across many domain-specific tasks. AlphaGeometry competes with silver and gold medalists in the International Math Olympiad [144], and Gemini Ultra[24] outperforms human experts across 57 subjects on the MMLU (Massive Multitask Language Understanding) dataset. Meanwhile, MedPrompt excels in medical benchmarks [104], surpassing monolithic LLM models like Med-PaLM-2 [130]. These achievements are largely attributed to their advanced system design, ability to integrate dynamic data, and integration of multiple, specialised components [182]. Monolithic models have limited performance improvement beyond their initial level, with substantial gains typically only achieved after exhaustive training. On the other hand, compound AI's sophisticated system design can lead to greater improvements with less effort. For example, DeepMind's AlphaCode 2 [6], a competitive programming tool, generates up to 1 million potential solutions for a problem, which are then filtered and ranked, thereby outperforming traditional LLMs. Thus, compound AI systems surpass the limitations of monolithic models and offer the best path toward enhancing the quality and reliability of AI applications.

Similarly, due to their distributed decision-making and problem-solving capabilities, MAS have demonstrated superior performance over single-agent systems ([47]; [77]). LLM-MAS, in particular, hold great promise for addressing complex problems and simulating various aspects of the world in fields like finance, healthcare, and policy-making. In finance, LLM agents serve as implicit computational models of human behaviour and have been used to simulate economic and trading environments [47]. These simulations encompass a wide range of scenarios, including macroeconomic activities, information marketplaces, financial trading, and the dynamics of virtual towns ([75]; [163]; [78]; [189]). LLM agents are also employed in scientific discussions and debates, where they analyse complex medical data, make decisions, and provide insights based on available medical knowledge [141]. Another recent study by [165] leveraged these agents to model the propagation of disease. Furthermore, LLM-MA systems can simulate virtual governments and assess the impact of various policies on communities, providing valuable insights to policymakers ([33]; [42]; [169]; [165]). What this progress suggests is that these AI systems are likely to remain a dominant trend in the foreseeable future.

3.3.3 Collaboration Patterns and Prompting Techniques in LLM-based MAS

Table 3.1 summarises notable LLM-based MAS applications on downstream reasoning-intensive tasks, outlining their collaboration patterns and prompting techniques.

[24] https://deepmind.google/technologies/gemini/ultra/

Paper	Task	Modalities	Method	Design	Prompting	Representation	Results	Insights
[22]	VQA, outside knowledge VQA, visual entailment, visual spatial reasoning	text, vision	• **Ensemble:** Averages outputs from multiple VLMs • **Cola-Zero:** Frozen LLM coordinates multiple VLMs • **Cola-FT:** Instruction-tuned LLM coordinates multiple VLMs	• Cooperation • Rule-based • **Ensemble:** Decentralised • **Cola:** Centralised	Zero-shot, Few-shot	Agent dialogue illustration	Outperforms the Ensemble method	• Reasoning performance improves through LLM finetuning or in-context learning • Explicitly prompting the LLM to generate rationales does not necessarily enhance performance
[157]	VQA	text, vision	• **SIRI framework:** Collaboration of 3 agents simulating top-down human reasoning • Responder (VLM): generates answer candidates and replies to queries • Seeker (LLM): queries the Responder • Integrator (MVKB): builds final answer	• Cooperation • Role-based • Decentralised	Zero-shot, CoT	System, Reasoning process illustration	Outperforms baseline	• VLMs struggle with event correlation in images • Top-down reasoning improves VLM performance
[31]	Arithmetic reasoning, mathematical reasoning, chess move prediction	text	Multi-Model debate over multiple rounds (short and long form)	• Cooperation • Rule-based • Decentralised • Debate	Zero-shot, Few-shot, CoT, reflection	Agent debate illustration	Outperforms baselines	• Debate improves mathematical and strategic reasoning • Zero-shot CoT prompting boosts performance • Debate enhances accuracy and reduces hallucinations • Performance scales with agents and rounds
[59]	VQA	text, vision	Multi-agent system that adaptively calls agents to analyse shortcomings, fill in missing information, and discover the final answer step-by-step	• Cooperation • Rule-based • Decentralised	Zero-shot, CoT	Reasoning process illustration	The proposed multi-agent pipeline improves accuracy by ∼10% over frozen LVLMs, but remains ∼10% below fine-tuned LVLMs	• Most VQA models depend heavily on dataset-specific fine-tuning with low zero-shot generalization • Focuses on zero-shot performance without fine-tuning • Optimizes inference by skipping modules when unnecessary

Paper	Task	Modalities	Method	Design	Prompting	Representation	Results	Insights
[51]	Outside knowledge VQA	text, vision	Three LLM-based agents simulate team roles (Junior, Senior, Manager), each using distinct tools and planners	• Cooperation • Role-based • Decentralised	Few-shot, generated knowledge prompting, inception prompting, tool selection	System diagram, Prompt construction diagram	Increased performance over the baseline	• Tools automatically create prompts and inject knowledge • Requires a single V100 GPU to run
[175]	Image Captioning	text, vision	MoColl: LLM-based agent queries a VQA model iteratively, then generates caption based on accumulated Q&A pairs	• Cooperation • Role-based • Decentralised	Few-shot, RAG	System diagram	Outperforms all baseline methods across all metrics	• Adapting to domain-specific data improves learning • More few-shot examples enhance captioning • Number of sub-questions boosts vocabulary • LLM decomposes captioning into subtasks • LLM optimizes the VQA module
[151]	Recom-mendation	text	MACRec: multi-agent framework of specialized agents (Manager, Analyst, Reflector, etc.)	• Cooperation • Role-based • Hierarchical	Zero-shot, reflection, inception role prompting	Agent role illustration, web interface	Qualitative improvements in task-specific coordination	• Applicable to rating, explanation, conversation tasks • Agent customization adapts to recommendation needs
[190]	Scientific research	text	ChatGPT Research Group: seven LLM agents for literature search, planning, lab design, analysis, etc.	• Cooperation • Role-based • Centralised • Human-AI collaboration	Generated prompting, prompt chaining, inception prompting	Agent role illustration	Optimized synthesis of MOF-321, MOF-322, COF-323	• Supports intensive research workflows • Addresses human bias and multitasking • Built in under a month
[172]	Commonsense reasoning	text	FORD: formal debate with three stages — fair, mismatched, roundtable	• Cooperation • Rule-based • Centralised • Debate	Zero-shot, Few-shot, inter-consistency	Agent debate illustration	Outperforms baselines across datasets	• Debates improve LLM performance and consistency • Stronger LLMs can be misled in mismatched settings
[16]	Open-ended QA, dialogue response generation	text	ChatEval: referee agents evaluate model responses	• Cooperation • Rule-based • Decentralised • Debate • Communication: one-by-one, simultaneous, summarizer	Inception prompting	Graph illustration	Improves text evaluation accuracy	• Repeating roles can degrade performance • Varying strategies boosts effectiveness • Performance increases with more roles

Paper	Task	Modalities	Method	Design	Prompting	Representation	Results	Insights
[141]	Medical reasoning (anatomy, clinical knowledge, college medicine, medical genetics, professional medicine, and college biology)	text	MedAgents: LLM-based agents follow a five-stage process: gathering domain experts, proposing individual analyses, summarising findings, conducting collaborative consultation with iterative revisions, and reaching a final unanimous decision	• Cooperation • Role-based • Decentralised	Zero-shot	System pipeline illustration	Outperforms Zero-Shot baselines and matches Few-Shot baselines across nine datasets	• Medical knowledge can be integrated through role-playing agents • Systematic errors attributed to misretrieved knowledge, consistency issues, and CoT flaws
[24]	Problem-solving	text	CoMM: LLMs take on distinct roles as two problem solvers from different disciplines and one summariser, forming a collaborative problem-solving team	• Cooperation • Role-based • Decentralised	Zero-shot, Few-shot, inception prompting, CoT	Agent role illustration	Outperforms state-of-the-art baselines in both Zero-shot and Few-shot settings	• Multiple domain experts across disciplines improve performance • Discussion rounds' impact varies by dataset • Assigning distinct reasoning paths to roles is an effective strategy
[53]	Code generation	text	MapCoder: Uses four LLM agents assigned to retrieval, planning, coding, and debugging	• Cooperation • Rule-based • Decentralised	Few-shot, CoT, analogical prompting	System diagram	Achieves new state-of-the-art across benchmarks, languages, and problem difficulty	• Modular decomposition significantly boosts performance over single-agent prompting
[72]	Problem-solving	text	Three LLM-based agents take turns interacting with the text game interface, updating beliefs, and generating actions and messages	• Cooperation • Rule-based • Decentralised	Zero-shot, CoT, scratchpad prompting (dialogue belief state tracking)	Game environment diagram	Achieves performance comparable with SOTA reinforcement learning algorithm	• Emergent collaboration and Theory of Mind capabilities • Belief state tracking mitigates hallucinations and long-horizon failures
[28]	Complex reasoning (arithmetic, commonsense, symbolic)	text	Master–worker framework: master agent decomposes tasks and delegates to worker agents using KQML-based communication	• Cooperation • Rule-based • Hierarchical • Synchronous and asynchronous communication	Few-shot, CoT, GoT, Tree-of-Thoughts, ReAct	Flowchart, System diagram	Improves coordination in complex reasoning tasks	• Prompt-based frameworks enhance LLM reasoning • Structured communication prevents context overflow • Complex frameworks hard to evaluate

Paper	Task	Modalities	Method	Design	Prompting	Representation	Results	Insights
[49]	QA, Causal Reasoning	text	LEGO: Five LLM agents in two modules — knowledge integration (Cause, Effect, Master) and refinement (Explainer, Critic)	• Cooperation • Role-based • Decentralised	Few-shot, inception prompting	Agent role illustration	Outperforms baselines in human evaluation	• Role-based agents handle complex reasoning better than single LLM • One round of feedback boosts precision; more rounds may introduce noise
[81]	Commonsense MT, counter-intuitive arithmetic reasoning	text	MAD: Multi-Agent Debate framework with tit-for-tat LLMs overseen by an LLM judge	• Cooperation • Rule-based • Centralised • Debate	Zero-shot, CoT, reflection, meta-prompting	Agent debate illustration	Improves performance on challenging reasoning tasks	• Mitigates Degeneration-of-Thought in self-reflection • Weak judge + strong debaters > strong judge + weak debaters • Complex questions require more debate rounds
[20]	Spatial reasoning, strategic planning, numerical reasoning, risk assessment, communication, opponent modelling, and team collaboration	text	LLMArena: Introduces the LLMArena benchmark to evaluate LLM agents in seven multi-agent, dynamic game environments and presents baseline experiments comparing different model sizes and types	Competition	Zero-shot, without hints, self-refinement, self-consistency	Game environment illustration	GPT-4 leads across all environments with a normalised average score of 100; capabilities improve with model scale	• LLMs exhibit weaknesses in opponent modelling and team collaboration • Bid and Hanabi expose larger model performance gaps • Numerical reasoning + team tasks are hard for small LLMs
[115]	Multi-skill QA	text	MetaQA: A collaborative QA system that combines expert agents by aggregating their answers and confidence scores to select the best final answer	• Cooperation • Rule-based • Centralised	Zero-shot	System architecture diagram	Outperforms prior multi-agent and multi-dataset QA systems	• Collaborative answer aggregation improves QA • Confidence scoring is key • Meta-level reasoning boosts agent selection
[174]	Mathematical, commonsense, symbolic reasoning	text	A multi-agent collaboration framework emulating academic peer review where agents generate, review with scores, and revise solutions	• Cooperation • Rule-based • Decentralised	CoT, inception prompting	Agent dialogue illustration	Superior accuracy across ten datasets vs. existing methods	• Feedback > solution sharing • Too much discussion can reduce accuracy • Smaller model gap = better collaboration • Model diversity boosts collaboration • Letting one model think divergently helps

Paper	Task	Modalities	Method	Design	Prompting	Representation	Results	Insights
[25]	Text understanding, reasoning, coding, tool use, embodied AI	text	AgentVerse: A multi-agent framework coordinating experts via expert recruitment, collaborative decision making, action execution, and evaluation	• Cooperation • Role-based • Decentralised / Hierarchical • Vertical / Horizontal structure	Zero-shot	System, tool use, collaboration, graph illustrations	Outperforms CoT and solo settings	• Recruiter agents as alternative to inception prompting • Emergent behaviours: volunteering, conformity, sabotage • Reflects complex social dynamics
[139]	Software development, court simulation	text	A general multi-agent framework where LLM agents are dynamically assigned roles, connect via plugins, and interact in a black-box environment	Role-based	Inception prompting	–	Could enhance existing AGI-style systems	• Effective collaboration needs modular design and protocols • Key concerns: agent overpopulation, scalability, evaluation metrics
[116]	Multi-hop QA, outside knowledge VQA	text, vision	AutoAct: An agent learning framework that synthesises planning from limited data using a tool library, then creates specialised sub-agents via division-of-labour	• Cooperation • Rule-based • Centralised	Zero-shot, Few-shot, reflection, self-instruct, tool selection prompting, self-differentiation, trajectory synthesis	System diagram	Better or comparable to all baselines	• Bootstrapping agent skills needs no large-scale data • Self-instruction and planning suffice for high QA accuracy
[36]	Task decomposition, tool integration, autonomous execution	text	Proposes a framework that decomposes user queries into task graphs, integrates tool selection, and introduces new evaluation metrics and a dataset	• Cooperation • Role-based • Hierarchical	Tool selection	System architecture diagram	Outperforms baselines in decomposition and tool accuracy; new metrics improve evaluation	• Coarse + fine-grained graphs reduce redundancy • Structural metrics (SSI) matter for sequential tasks • Tool metrics (Tool F1) matter for parallel tasks

Table 3.1: Overview of collaboration patterns and prompting techniques in LLM-based Multi-Agent System applications involving reasoning-intensive downstream tasks

3.3.4 Challenges and Open Questions

The field of compound AI and LLM-MA is relatively new and the design, optimisation, and operation of these systems remain challenging. In both compound AI and LLM-MA systems, many shortcomings arise from the intrinsic limitations of LLMs, namely their black-box nature, hallucinations, reasoning abilities, knowledge gaps, and role-playing capabilities ([151]; [47]; [117]; [187]). However, the intricacies of their design introduce additional challenges.

In compound AI systems, developers must choose the best system design from a wide range of options and determine how to efficiently allocate resources across multiple components. Moreover, since compound AI systems often include non-differentiable elements, optimising components to function efficiently together is no trivial task and often requires specialised, ad hoc tools. Machine Learning Operations (MLOs), such as assessing performance or debugging, are also more complex for compound AI, as it is often challenging to attribute mistakes or successes to a specific component [182].

In LLM-MA systems, developers have to design coordination mechanisms that avoid groupthink and cognitive biases [75]. Establishing a comprehensive framework for coordination can be challenging, particularly in human-agentic workflows [3]. A terminological framework for human-agentic collaboration has only recently been proposed by [3], and there still remains a lack of empirical insight into which human-agentic workflows are effective and the underlying reasons for their success. Guo et al. [47] also reported a notable gap in multi-modal settings, with most LLM-MA research focusing on text-based environments. Multimodal environments pose additional challenges, requiring agents to process and integrate various types of data, such as text, images, audio, and video, each with different formats and structures. Another cited gap is the lack of standardised evaluation frameworks for AI agents ([47]; [62]). AI agent benchmarking is still an emerging field, as agents differ significantly from models, and a comprehensive evaluation framework, particularly one that accounts for humans-in-the-loop, has yet to be established ([62]; [47]). Kapoor et al. [62] have taken the first step toward this by proposing measures to enhance the rigour of AI agent benchmarking and shifting the focus away from narrow accuracy.

The design and prompting of compound AI and LLM-MA systems is still evolving, though a number of tools and techniques are available. As the focus has shifted from models to AI systems, so too has the programming approach, transitioning from hard-coded programming to *language model programming*, where developers call models and tools from component libraries. DSPy [63] was the first programming paradigm for end-to-end optimisation of LLM systems with multiple calls and tool definitions. To this end, DSPy leverages the linguistic abilities of LLMs and user inputs in natural language to generate prompts that maximise performance. Recent advancements also enable end-to-end optimisation of system parameters by using an LLM to generate and refine instructions and tool definitions, enhancing the performance of compound AI systems [82]. MLOs, such as debugging and performance tracking, are now considerably easier thanks to software advancements. There are also significant efforts to theorise human-agentic workflows. Recent advancements have extended the Business Process Model and Notation (BPMN) to effectively formalise and represent human-agentic collaborative workflows [3]. However, there remains significant potential for experimentation and refinement, especially in real-world settings.

Chapter 4

Approach

4.1 Tasks, Datasets, and Related Work

4.1.1 Image Paragraph Captioning and Dense Captioning

Tasks

Image Paragraph Captioning (IPC) refers to the task of generating a coherent, multi-sentence paragraph P that describes the image I. Paragraphs are linguistically complex structures that convey both high-level semantics and low-level visual details from the image. Depending on the complexity of the image, the model needs to possess the following abilities in order to produce a high-quality paragraph description:

- **Visual understanding**: At the most basic level, the model needs to understand low-level semantics such as edges, regions, and textures, as well as mid-level semantics such as objects, trajectories, and locations within the image. Beyond basic recognition, it must also be able to grasp spatial relationships, temporal dynamics, and overall scene context in order to reason effectively about the identified elements and their properties.

- **Linguistic capabilities**: Paragraphs are linguistically complex structures. This complexity is reflected in the higher use of verbs, subordinating conjunctions, and other linguistic features such as coreference, which appear more frequently than in traditional single-sentence captions. Structure, coherence, and flow across multiple sentences are also essential, often achieved through the use of discourse markers that connect ideas and highlight relationships. As a result, producing a high-quality paragraph demands not only fluency but also a deep understanding of complex linguistic phenomena.

- **Knowledge**: High-quality image paragraph captioning goes beyond low- and mid- level semantics and requires at least five clusters of knowledge that are associated with the top level of the visual semantic pyramid, as shown in Figure 4.1.1 and Figure 4.1.2. We adopt the four knowledge clusters identified by [109] and introduce a fifth—outside knowledge—which we argue is essential for addressing knowledge-based image paragraph captioning tasks. The first cluster identified by [109] is **commonsense knowledge**, such as understanding actions, events, object relationships, interactions, and purposes that are either explicit or implied in the scene. **Emotional knowledge** is also required for high-level paragraph captioning, allowing the model to perceive emotions, moods, emotional significance, and affective tones. At the same time, **aesthetic knowledge** plays a role in describing the overall atmosphere and visual appeal of the image as a unified whole. Positioned at the uppermost level of high-level semantics, **inductive interpretative knowledge** is needed to capture complex, often subjective or culturally encoded meanings such as symbolism, 'aboutness', or abstract concepts. Finally, **outside knowledge** is associated with mid-level semantics and is essential for accurately describing images that depend on real-world context, such as historical landmarks, public figures, or artworks. This type of knowledge typically requires either prior domain knowledge or retrieval-based access.

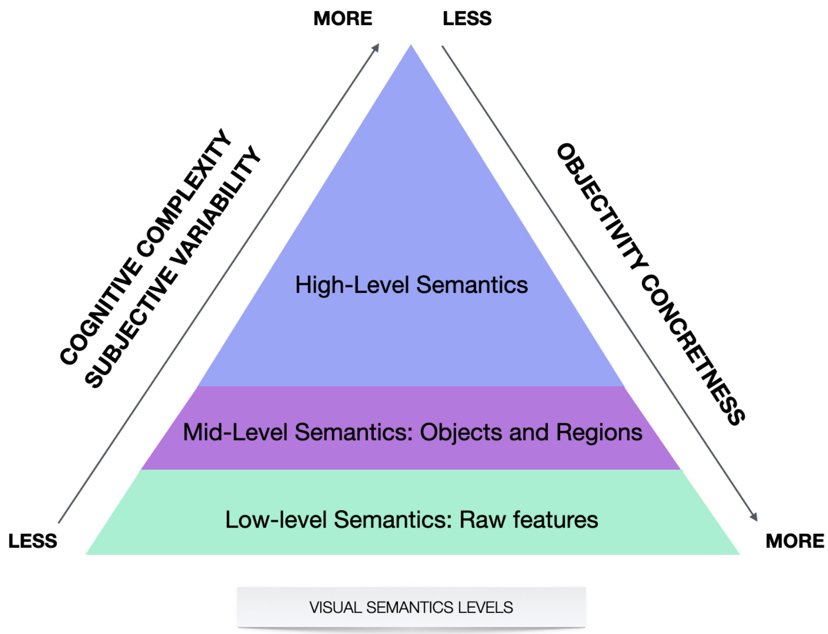

Figure 4.1.1: Visual semantics levels, (adjusted from [109]).
Notes: The low-level includes raw or elemental features, the mid-level involves individual objects and regions, and the high-level encompasses purpose, emotion, aesthetics, and interpretation.

- **Abstraction-Level Awareness**: A high-quality paragraph needs to strike the right level of abstraction for each element while effectively conveying the meaning of the image. This means that the model must be able to understand which aspects of the image are important and the level of detail or abstraction at which they should be presented. Describing every element in detail would result in a text that exceeds the scope of a paragraph and would likely compromise coherence and flow. Therefore, the model must be able to decide which elements to describe and to what extent. We define two key types of decisions regarding abstraction levels that models have to make during image paragraph captioning: **concept selection**, determining which aspects of the image to include (e.g., 'Should I talk about the plane in the sky?'), and **name specificity**, deciding how precisely to name them (e.g., 'Should I call this a plane, commercial plane, or Boeing 737?').

Datasets

Historically, image captioning has been one of the most widely studied tasks in computer vision involving language. However, in the era of LLMs, image captioning no longer poses a significant challenge for multimodal LLMs. On the other hand, image paragraph captioning requires diverse abilities and should be seen as a more demanding benchmark for evaluating models' capacity for complex visual understanding and commonsense reasoning.

Stanford's Image Paragraph Captioning dataset [65] is the most prominent in the area of image paragraph captioning and comprises 19,561 images from the MS COCO[1] and Visual Genome dataset[2], each annotated with a paragraph description. Annotations were collected via Amazon Mechanical Turk using experienced U.S. workers with a high acceptance rate, and were subject to both automatic and manual quality checks. Suggested language metrics for this task include METEOR, CIDEr, and Bleu-{1,2,3,4}. More recent datasets for dense image captioning include human-annotated sets like DCI [145], DOCCI [105], and IIW [40], as well as model-generated datasets such as PixLore [12].

[1] https://paperswithcode.com/dataset/coco
[2] https://paperswithcode.com/dataset/visual-genome

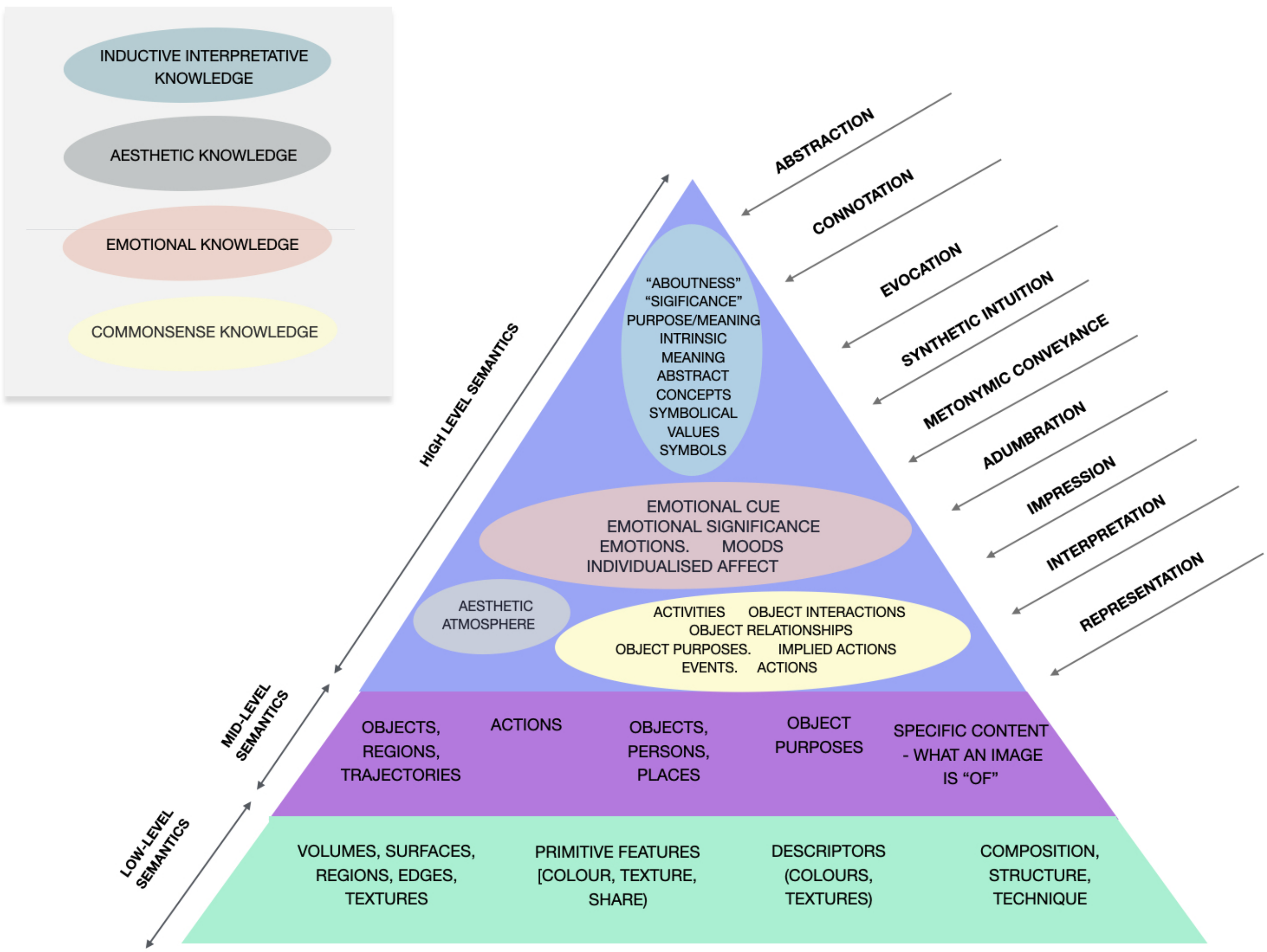

Figure 4.1.2: The three levels of visual semantics (adjusted from [109]).

We focus on the Stanford Image-Paragraph Captioning dataset, which is the most well-studied in the area and one that played a defining role in shaping the image paragraph captioning task. We evaluate this dataset and its successors through the lens of knowledge-based VQA. We argue that the ability to generate detailed, context-aware, and knowledge-rich paragraph descriptions of images is a crucial step toward improving model performance in tasks that require deeper reasoning, hence our interest in the intersection of these two tasks. Therefore, we seek datasets that are human-annotated, VQA-focused, and capture key challenges such as optical character recognition (OCR), the use of outside knowledge, and varying levels of abstraction. In Table 4.1, we provide a brief overview of notable datasets and how well they align with these goals.

	Stanford IPC	**DCI**	**DOCCI**	**IIW**	**PixLore**
Human annotated	Yes	Yes	Yes	Yes	No
VQA focus	Low	Low	Low	Low	Low
Knowledge-based	Low	Low	Yes	Low	Low
OCR	Low	Low	Yes	Yes	Low
Abstraction Level	No	No	No	No	No

Table 4.1: Overview of notable image-to-text datasets evaluated for their relevance to knowledge-based VQA.

Related Work

What is particularly striking about the Stanford's Image Paragraph Captioning dataset is that most of the frameworks explicitly tested on it, including those reporting the highest results, are based on traditional architectures such as RNNs, CNNs, LSTMs, and GANs ([176]; [97]; [173]; [150]; [19]. An overview of the results achieved by more traditional architectures is presented in Table 4.2. State-of-the-art LVLMs and MLLMs, despite achieving state-of-the-art performance among VLMs in many other tasks, have not been extensively evaluated on the Stanford Image Paragraph Captioning dataset, although a few exceptions do exist [64]. We argue that, despite their value, Stanford's dataset and similar ones that predate LLMs no longer pose a sufficient challenge for modern multimodal LLMs or reflect their actual capabilities. In this work, we use Claude-3.7 as a proxy to demonstrate that LLMs are capable of generating image paragraph captions that surpass the scope and complexity of existing datasets and evaluation metrics. Based on these findings, we identify key areas that are still challenging for LLMs and propose future directions for benchmarking in image paragraph captioning.

Rank	**Model**	**BLEU-4**	**METEOR**	**CIDEr**	**Method**
1	HSGED (SLL) [176]	11.26	18.33	36.02	RNN
2	SCST training, w/rep. penalty [97]	10.58	17.86	30.63	CNN, LSTM
3	IMAP [173]	10.29	17.36	24.07	RNN
4	CAE-LSTM [150]	9.67	18.82	25.15	LSTM, CNN
5	Diverse and Coherent Paragraph Generation from Images [19]	9.43	18.62	20.93	VAE, GAN, RNN, CNN

Table 4.2: Notable results across linguistic metrics for the Stanford Image Paragraph Captioning dataset.

4.1.2 Knowledge-based VQA

Task

Knowledge-based VQA is a challenging subdomain of VQA that draws upon commonsense and outside knowledge to answer questions. Unlike traditional VQA tasks, where answers can be derived directly from visual cues (e.g., object recognition, counting, attribute detection), knowledge-based VQA tests not only perception but also inference and retrieval, either from the model's parameters or from external knowledge bases. In doing so, it exposes critical limitations in multimodal models, such as multi-hop reasoning, knowledge retrieval, and hallucinations.

Q: Which transportation way in this image is cheaper than taxi?
A: Bus

Q: What was the name of the first cloned type of this animal?
A: Dolly

Q: Do all the people in the image have a common occupation?
A: No

Q: List common properties of these two images.
A:
Background: snow
Scene: ski slope, ski resort, mountain snowy
Object concepts: racing, winter sports, outdoor recreation.

Figure 4.1.3: Image-question pairs across datasets that draw on commonsense and outside knowledge. The images are taken from FVQA, A-OKVQA, KVQA, and KB-VQA respectively.

Besides the **visual** extraction of the bus, the first image of Figure 4.1.3 requires **general world** knowledge that *buses are a form of public transportation and are typically more affordable than taxis*. Similarly, alongside visual recognition of the sheep, answering the second question demands **factual** knowledge that *Dolly* was the first cloned sheep. The third image goes a step further, requiring not only the **visual** extraction of the two people present, but also **named entity** knowledge to identify them as *Hillary Clinton* and *Aamir Khan*, followed by **factual and comparative** knowledge to determine that their occupations differ—Hillary Clinton being a *politician*, and Aamir Khan being an *actor*. Finally, answering what common properties the pair of images share requires identifying transitive categories, as well as applying **comparative and commonsense** knowledge to infer the objects, attributes, and scenes present in both. Answering all of these questions relies on external knowledge, which must be extracted and applied to infer the correct answer.

Datasets

To benchmark this ability, datasets such as OK-VQA and A-OKVQA have been developed to test reasoning with open-ended knowledge, while others like KB-VQA and FVQA evaluate reasoning grounded in a given knowledge base. An overview of notable datasets for knowledge-based VQA is provided in Table 4.3.

We chose to focus on OK-VQA because, despite being a long-standing benchmark for knowledge-based VQA, no existing framework, to the best of our knowledge, has achieved human-level performance. Despite progress, the highest reported accuracy on OK-VQA, achieved by the PaLI-X-VPD model, is still just 66.88%, which is significantly lower than state-of-the-art results on related tasks such as A-OKVQA. Besides testing high-level reasoning and retrieval, what makes this dataset particularly challenging is its open-ended format, which requires models to generate answers in a specific format that also align with human-annotated answers. This dataset also offers an opportunity to evaluate emerging multimodal models, such as Claude, which have not been extensively studied in this context.

OK-VQA comprises 14,055 questions about images, with 9,009 in the training split and 5,046 in the validation split. The dataset includes a wide range of questions spanning ten distinct knowledge categories, with category distributions shown in Figure 4.1.6 and representative examples provided in Figure 4.1.4. Each question is

Dataset	Knowledge Type	Format	Rationale	Goal
OK-VQA [96]	factoid	DA	No	visual reasoning with open-world knowledge
A-OKVQA [125]	common/world	MC/DA	Yes	visual reasoning with open-world knowledge
KB-VQA [152]	fixed KB	DA	No	visual reasoning with given knowledge base
FVQA [153]	fixed KB	DA	Yes	visual reasoning with given knowledge base
VCR [183]	people actions	MC	Yes	visual commonsense reasoning
ScienceQA [90]	scientific	MC	Yes	multimodal reasoning over images, diagrams, text
MMMU [181]	mixed (text, image, graph)	multi-format	Yes	multidisciplinary multimodal reasoning
S3VQA [54]	factoid	DA	Yes	situated visual question answering with OCR + text
KVQA [68]	world knowledge	DA	No	world knowledge about named entities

Table 4.3: Various datasets in knowledge-based VQA and their characteristics.

associated with a specific knowledge category and paired with an image. For every question, there are ten answers provided by human-annotators along with their confidence scores. The dataset includes both the raw answers, the extracted answers, and an associated confidence score. The distribution of answer lengths is presented in Figure 4.1.5.

The creators of the dataset suggest this accuracy metric to account for inter-human variability in phrasing the answers:

$$\mathrm{Acc}(\mathit{ans}) = \min\left\{\frac{\#\text{humans that said } \mathit{ans}}{3}, 1\right\} \tag{4.1.1}$$

Before evaluating machine-generated answers, the following pre-processing steps are applied:

- All characters are converted to lowercase.
- All periods are removed, except when they appear as decimal points.
- Number words are converted to digits.
- All articles (a, an, the) are removed.
- An apostrophe is added if a contraction is missing one (e.g., dont → don't).
- All punctuation, except apostrophes and colons, is replaced with a space character. In the case of commas, no space is inserted if the comma appears between digits (e.g., 100,978 → 100978).

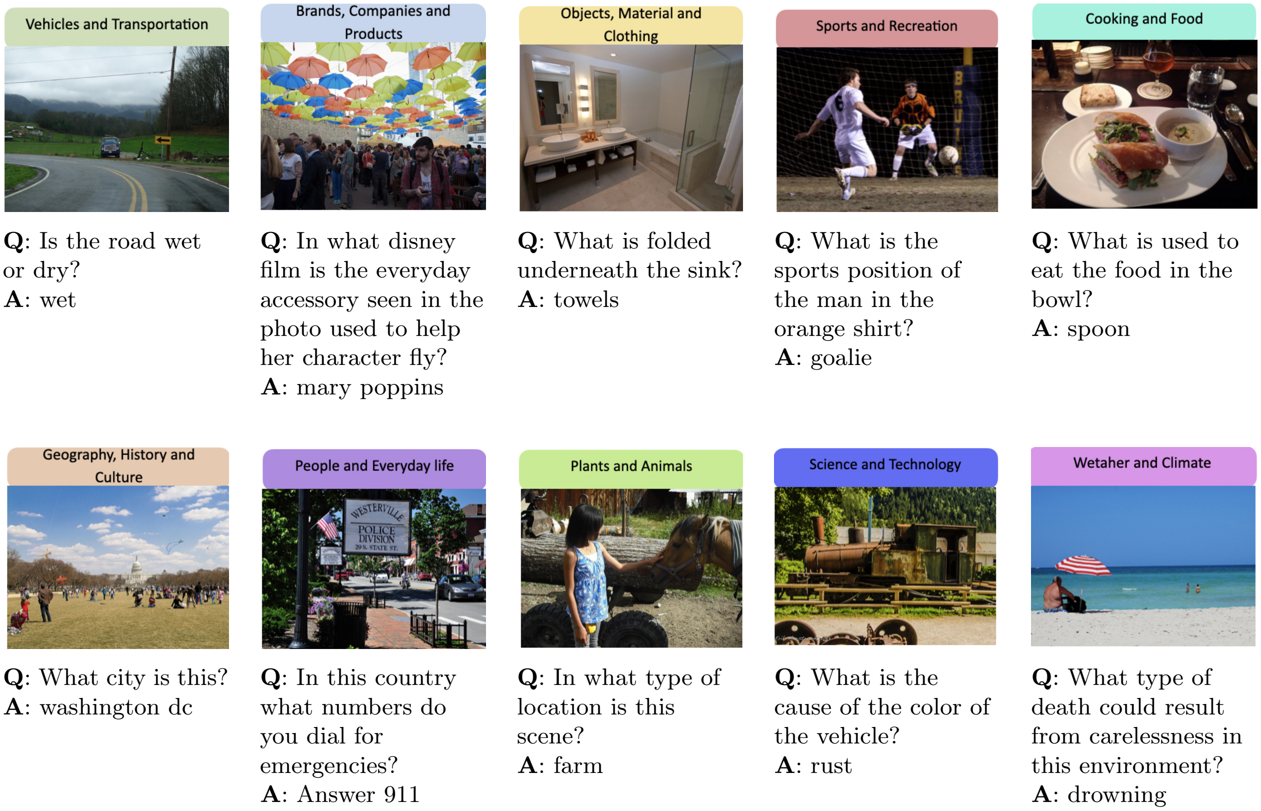

Figure 4.1.4: Examples drawn from the OK-VQA dataset across categories.

Related Work

An overview of the highest results achieved by various papers on the OK-VQA dataset is presented in Table 4.4. Recent work on OK-VQA shows that the most successful approaches combine strong vision-language models (like PaLI-X or PaLM-E) with either fine-tuning on OK-VQA or external knowledge retrieval, often using dense passage retrievers or curated databases. Retrieval-augmented generation and fine-grained visual reasoning (e.g., using object-centric features) have also consistently boosted performance. A common trend is moving from specialised architectures (e.g., KRISP, MAVEx) toward scaling large pre-trained multimodal transformers with minimal task-specific design. Critiques of OK-VQA point to its lack of annotated knowledge, free-form answer evaluation challenges, and biases that allow guessing. Newer work like A-OKVQA addresses some of these by introducing multiple-choice answers and rationales. For evaluating generalist multimodal LLMs like GPT-4V, Gemini, and Claude without fine-tuning, researchers often test them in zero-shot or few-shot setups on OK-VQA and A-OKVQA, finding that while they perform reasonably well, explicit retrieval or prompting with external knowledge remains critical to close the gap with fine-tuned models.

4.2 Method

4.2.1 Image Paragraph Captioning with Multimodal LLMs

We work with a randomly selected subset of 5,000 images from the Stanford Image Paragraph Captioning dataset. To evaluate model performance, we select Claude as a representative state-of-the-art multimodal LLM and conduct the majority of our experiments using the 3.7 Sonnet version. Claude is a family of large multimodal models developed by Anthropic that features three types of models: Haiku (prioritising speed and affordability), Sonnet (balancing capability and cost), and Opus (geared toward complex reasoning tasks). All three models support both text and image inputs and exhibit strong performance on visual reasoning benchmarks, leading us to hypothesise that they are capable of generating detailed, context-aware,

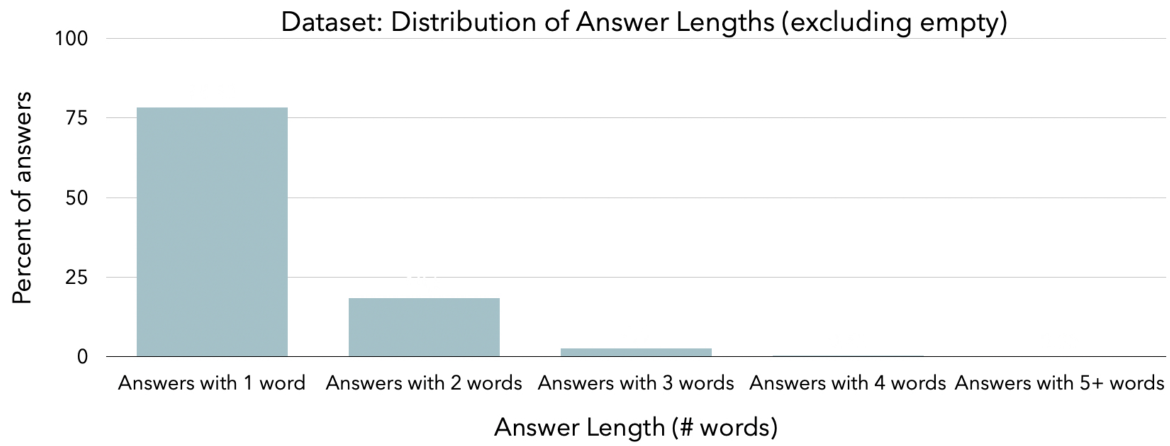

Figure 4.1.5: Distribution of answers with 1, 2, 3, 4 or 5+ words in the OK-VQA dataset

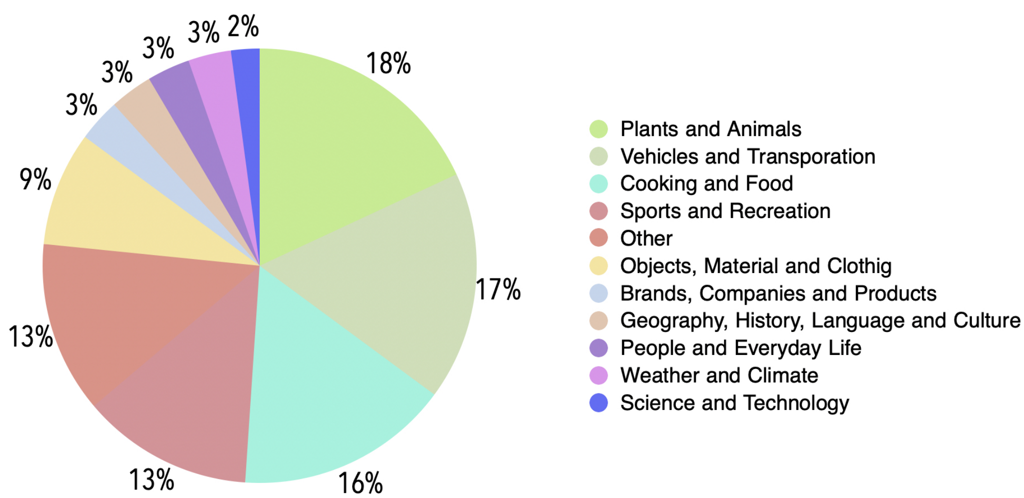

Figure 4.1.6: Distribution of questions across categories in the OK-VQA dataset

and factually grounded image paragraphs. To our knowledge, no formal results on this dataset have been reported by commercial model evaluations or prior academic work, and Claude 3.7 remains a relatively recent release. We experiment with all members of the Claude 3.7 family but choose Claude 3.7 Sonnet for our main experiments on the Stanford Image Paragraph Description dataset, as it offers the best balance between performance and cost-effectiveness.

We input images from the Stanford subset to Claude Sonnet 3.7 and prompt it to generate paragraph-length descriptions in a zero-shot fashion. We evaluate the generated descriptions using METEOR, BLEU-{1,2,3,4}, and SBERT-based cosine similarity scores. We then conduct a linguistic analysis of both the ground-truth paragraphs from the dataset subset and the paragraphs generated by Claude.

For our analysis, we examine the following linguistic metrics to assess the quality and human-likeness of the generated descriptions and to draw comparisons with the ground-truth annotations in our Stanford subset:

- **Average paragraph length**: Measures the overall verbosity or conciseness of the descriptions. Longer paragraphs may indicate richer detail or more complex narrative structure and other linguistic phenomena. However, they may also increase the risk for hallucinations, as models tend to hallucinate more when talking in excessive details about images. The average length of human-written paragraphs

Rank	Model	Accuracy (%)
1	PaLI-X-VPD [30]	66.80
2	PaLM-E-562B [97]	66.10
3	PaLI-X (Single-task FT) [26]	66.10
4	PaLI 17B [27]	64.50
5	Prophet [179]	62.50

Table 4.4: Overview of results for the OK-VQA dataset

typically ranges from 60 to 120 words, depending on the context and the amount of information conveyed in the image. In the Stanford dataset, human annotators produce paragraphs averaging around 60–70 words.

- **Standard deviation of length**: Captures variation in description length across paragraph samples. A low standard deviation suggests consistency in the generated descriptions, while a high deviation may indicate flexibility in adapting to the amount of visual information present in the image. Consistency is a key quality we strive for in LLMs. That said, when a model generates responses of roughly the same length it might be a symptom of template-based descriptions or a lack of adjustment to image complexity. Moreover, when LLMs are forced to talk in excessive detail about images with limited meaningful content, they are also more susceptible to hallucinations.

- **Vocabulary size**: Refers to the total number of unique words used across all generated paragraphs and is a key indicator of the model's expressive range. A large vocabulary suggests the model can describe a wide variety of concepts, actions, attributes, and relationships, avoiding redundancy and generalisation. On the other hand, a model with a small vocabulary size may resort to dull, overly generic, unnatural-sounding, or even inaccurate descriptions.

- **Lexical diversity (Type-Token Ratio, TTR)**: Lexical diversity is a measure of how varied the vocabulary is in a piece of text. A higher TTR indicates greater lexical richness and more diverse, human-like language. Conversely, a lower TTR may indicate more repetitive and rigid language. To assess lexical diversity per paragraph, we choose the type-token ratio (TTR) metric, which we define as follows and then average:

$$\begin{aligned}&\textbf{Types} = \text{the number of unique words in a paragraph}\\&\textbf{Tokens} = \text{the total number of words generated in the paragraph}\\&\textbf{TTR} = \text{Types} \,/\, \text{Tokens}\end{aligned}$$

- **Grammar error rate**: Assesses grammatical correctness. A lower error rate signals higher fluency and linguistic competence, important for readability and user trust. To calculate the average grammar error rate per sentence we use the LanguageTool API, a rule-based grammar checker that can detect a wide range of issues in English text, such as grammatical, punctuation, or spelling mistakes. We then use SpaCy to parse the paragraph and count the number of sentences. The final measure of grammaticality is computed as the average number of grammar errors per sentence across the subset:

$$\textbf{Grammar Error Rate} = \text{Total Grammar Errors} \,/\, \text{Total Sentences}$$

- **Frequency of subordinating conjunctions**: Complex paragraphs use subordinating conjunctions (e.g., because, although, while, in order to) to ensure narrative flow and express causal, temporal, conditional, and contrastive relationships, among others. The use of these conjunctions reflects more advanced sentence structure, greater linguistic sophistication, and the communication of deeper ideas. In order to assess the frequency of these markers, we created a lexicon that includes different types of conjunctions, such as concession (e.g., although, even though), reason and purpose (e.g., because, in order to), and condition (e.g., if, unless) (Table 5.16)

- **Proportion of nouns, verbs, and pronouns**: The proportion of nouns, verbs, pronouns, and other parts of speech provides insight into the linguistic focus of the description. A higher frequency of nouns indicates object-centric descriptions, which are associated with low- and mid-level semantics. Conversely, a high verb ratio points to action-oriented descriptions and a stronger narrative flow (e.g., is running, holds, throws), and is associated with mid-level semantics. A moderate use of pronouns is common in natural, cohesive writing to avoid repetition. In paragraphs, complex linguistic phenomena occur, such as coreference, where different expressions refer to the same entity. However, too many pronouns might lead to ambiguity or compromise readability.

4.2.2 Multimodal Collaboration for Knowledge-based VQA

We conduct all experiments and ablations on the validation split, which consists of 5,046 images, and use the established accuracy metric to ensure a fair comparison. Our proposed approach is a multi-agent system consisting of three stages. In the first stage, we employ an MLLM (e.g., Claude 3.7), providing it with images and prompting it to generate a paragraph-length description for each in a zero-shot manner. We refer to these as *greedy descriptions*, as they are not guided by the questions. Instead, the model transfers all the information from the image into the description, regardless of the question. We call this model the **Scout**, as it explores and gathers information from the image without specific direction. In the next stage, we prompt a text-only LLM (e.g., Llama 3.3) to answer the question about the image, using only the image description provided by the Scout and the question itself. We refer to this model as the **Analyser**, as it serves as our main reasoner. Finally, we employ another text-only LLM to resolve and format the final answer to the question, based on a set of predefined rules. We call this model the **Resolver**, as it is responsible for making the final decision on the answer to the question. An illustration of our proposed multi-agent framework is shown in Figure 4.2.1.

Since the dataset expects answers of a specific word length, and we also want models to consistently attempt to answer, we also evaluate how well they meet these criteria. For each method, we analyse the word length distribution and the number of no-attempts to assess how effectively the models can be steered toward producing responses of a specific length and always attempting to answer the question. For the word length distribution, we simply calculate the percentage of answers with one word, two words, three words, and four or more words. To calculate the number of no-attempts, we use a set of predefined phrases, shown in Table 4.5, that models often produce when not answering the question, and use exact matching to count their occurrences. We use the following prompts for our models and experiment with zero-shot, few-shot, and CoT prompting (Table 4.7 to Table 4.10).

No-attempt phrases
unknown, not visible, can't be inferred, cannot be inferred,
not evident, unclear, indeterminate, not sure, no information,
n/a, insufficient information, not enough info, undetermined, uncertain, unsure

Table 4.5: No-attempt phrases.

Scout Prompt
Describe the image in a paragraph. Image: <image>

Table 4.6: Scout prompt.

Analyser CoT Prompt
You are given an image and a question about the image. Image: <image> Question: <question> Think step-by-step about the contents of the image and then answer the question as clearly as possible. You need to answer the question with only one or two words. You should attempt to answer the question with one or two words even if the actual answer is more nuanced.

Table 4.7: Analyser CoT prompt.

Resolver Zero-Shot Prompt
You are given the following: 1. A question about an image 2. A response from another model Question: <question> Model response: <response> Your task: Extract a concise answer to the question based only on the response. When formatting your answer, follow these rules: 1. Use a single word in most cases. Only use two words when they form a natural and necessary phrase. 2. Use the singular form unless plural is clearly appropriate. 3. Use the base form of verbs, unless the -ing form is a noun. 4. Do not answer with vague responses. Always choose the most intuitive and the most likely answer and make your best guess based on the available information.

Table 4.8: Resolver zero-shot prompt.

Resolver Few-Shot Prompt

You are given the following:

A question about an image

A response from another model

Question: <question>

Model response: <response>

Your task: Extract a concise answer to the question based only on the response.

When formatting your answer, follow these rules:

1. Use a single word in most cases. Only use two words when they form a natural and necessary phrase.

2. Use the singular form unless plural is clearly appropriate.

- Prefer: horse

- Acceptable: scissors (when singular form doesn't apply)

3. Use the base form of verbs, unless the -ing form is a noun.

- Correct: run, jump, dance

- Acceptable: gardening (when functioning as a noun)

- Avoid: jumped, running fast

4. Do not answer with vague responses like unknown, unclear, or not visible

Always choose the most intuitive and the most likely answer and make your best guess based on the available information.

Table 4.9: Resolver few-shot prompt.

Claude Solo Prompt

You are given an image and a question about the image.

Image: <image>

Question: <question>

Answer the question as clearly as possible.

You need to answer the question with only one or two words.

You should attempt to answer the question with one or two words, even if the actual answer is more nuanced.

Table 4.10: Claude Solo Prompt.

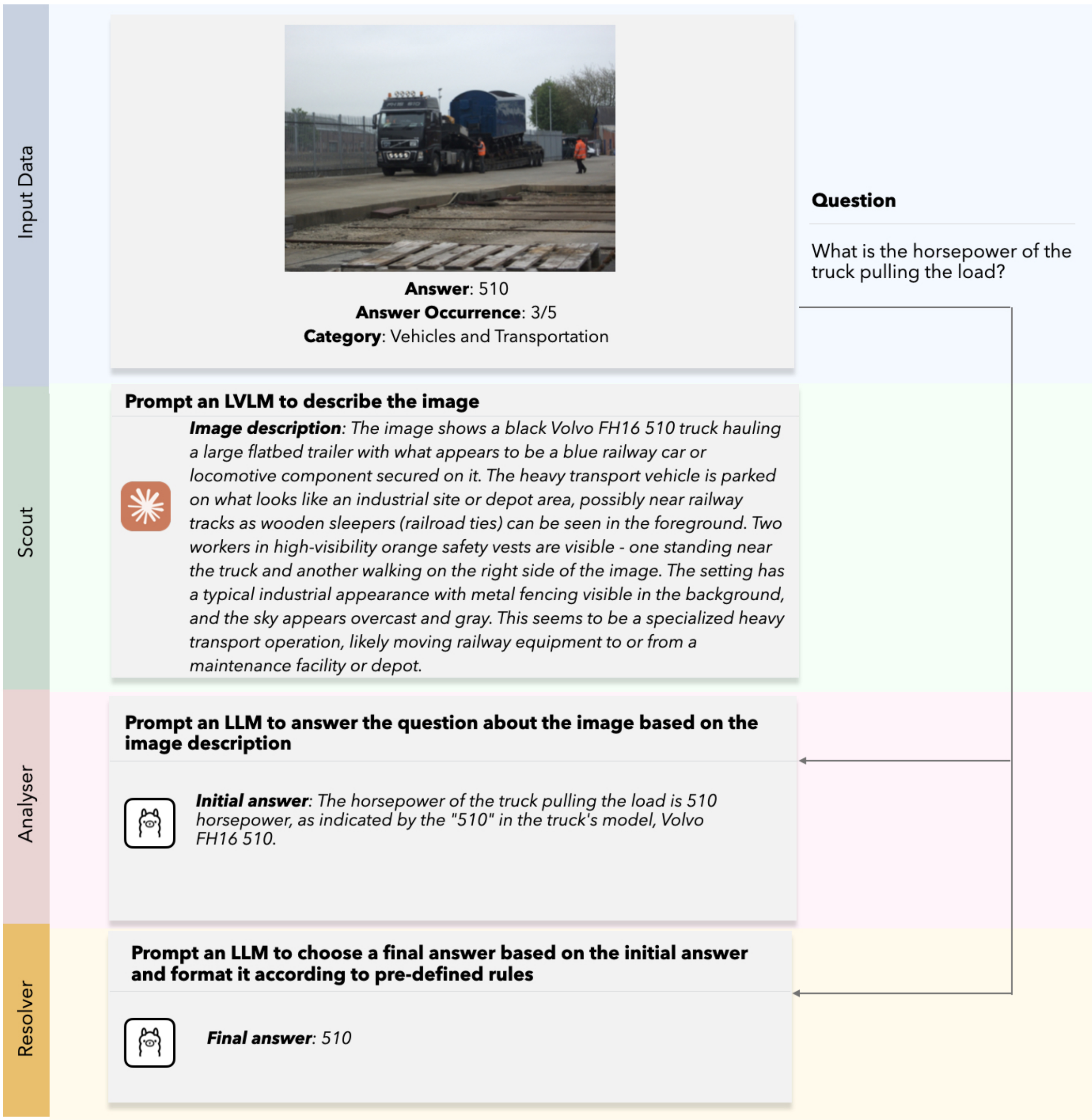

Figure 4.2.1: Illustration of our Scout, Analyser, Resolver collaboration framework.

Chapter 5

Experiments

5.1 Preliminaries

In this section, we detail the inference parameters and models used in our experiments (Table 5.1 to Table 5.6). We employed Claude 3.7 Sonnet as our vision model, while Llama 3.3 served as our text-only LLM. We conducted experiments using both the standard and the extended thinking variants of the Claude model to evaluate the impact of extended reasoning on performance.

Model	Claude
Version	3.7
Variant	Sonnet
Output modalities	text
Extended thinking	disabled
Model name	anthropic.claude-3-7- -sonnet-20250219-v1:0

Parameter	**Value**
temperature	0.0
max-tokens	1024

Table 5.1: Model and inference parameters for the Stanford Image Paragraph Description dataset.

Model	Claude
Version	3.7
Variant	Sonnet
Output modalities	text
Extended thinking	disabled
Model name	anthropic.claude-3-7- -sonnet-20250219-v1:0

Parameter	**Value**
temperature	0.0
max-tokens	1024

Table 5.2: Model and inference parameters for the OK-VQA image descriptions.

Model	Claude
Version	3.7
Variant	Sonnet
Output modalities	text
Extended thinking	disabled
Model name	anthropic.claude-3-7--sonnet-20250219-v1:0

Parameter	**Value**
temperature	0.0
max-tokens	1024

Table 5.3: Model and inference parameters for OK-VQA baseline and multimodal Analyser.

Model	Claude
Version	3.7
Variant	Sonnet
Output modalities	text
Extended thinking	enabled
Model name	anthropic.claude-3-7--sonnet-20250219-v1:0

Parameter	**Value**
temperature	1.0
max-tokens	24,000
budget-tokens	16,000

Table 5.4: Model and inference parameters for OK-VQA, multimodal Analyser with CoT.

Model	Llama
Version	3.3
Parameters	70B
Input modalities	text
Output modalities	text
Model name	meta.llama3-3-70b-instruct-v1:0

Parameter	**Value**
temperature	0.0
top-p	0.9
max-gen-len	1024

Table 5.5: Model and inference parameters for OK-VQA LLM Analyser.

Model	Llama
Version	3.3
Parameters	70B
Input modalities	text
Output modalities	text
Model name	meta.llama3-3-70b-instruct-v1:0

Parameter	**Value**
temperature	0.0
top-p	0.9
max-gen-len	512

Table 5.6: Model and Inference Parameters for OK-VQA Resolver.

5.2 Results

5.2.1 Image Paragraph Captioning with Multimodal LLMs

Stanford Image Paragraph Description

We present evaluation results for Claude 3.7 Sonnet on our randomly selected 5,000-image subset of the Stanford Image Paragraph Captioning dataset, using METEOR and BLEU-{1,2,3,4} as our evaluation metrics. While most prior work has relied on the full dataset, we conduct our experiments on a representative subset to balance computational efficiency with robustness. We also provide a comparison against human-written paragraph descriptions, using the evaluation conducted by Krause et al. [65] on a 500-image subset of the Stanford Image Paragraph Captioning dataset.

Clearly, Claude 3.7 generates semantically relevant paragraphs, as evidenced by its high METEOR score, which exceeds that of human-written descriptions. However, the model's outputs differ significantly from the exact wording of human references, resulting in low BLEU scores, especially for longer n-grams. Human-written paragraphs, on the other hand, match more closely with reference paragraphs in terms of word choice, reflecting greater consistency in phrasing across humans. This suggests that while Claude is effective at capturing the meaning of the image, it diverges from the ground-truth paragraphs in terms of surface-level phrasing.

We hypothesise that Claude generates semantically rich paragraphs, but has a distinct style of delivery than humans that is not adequately captured by narrow metrics such as BLEU. To validate this, we employ SBERT cosine similarity, a more profound metric that does not rely on exact word matching. For instance, SBERT can acknowledge that ‘a man riding a bicycle’ and ’someone cycling down the street’ are semantically similar, even if phrasing differs. As expected, Claude achieves a high score of 70.18%, indicating that the generated paragraphs are very similar to the reference paragraphs in terms of actual context.

Model	**METEOR**	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**
Claude 3.7 - Sonnet	24.35	21.15	9.30	4.01	1.84
Human	19.22	42.88	25.68	15.55	9.66

Table 5.7: Performance comparison between Claude 3.7 - Sonnet and human references using METEOR and BLEU scores.

Figure 5.2.1: Word cloud of human-annotated paragraphs from the 5,000-image Stanford Image Paragraph Captioning subset, showing a more object-centric and literal focus.

Figure 5.2.2: Word cloud of paragraphs generated by Claude 3.7 Sonnet for the 5,000-image Stanford Image Paragraph Captioning subset, showing a more abstract and interpretive focus.

To shed more light on why the obtained BLEU scores are so low, we use word clouds to visualise the differences in commonly used terms between the generated and reference paragraphs (Figure 5.2.1, Figure 5.2.2). Indeed, it is easily observable that the phrasing differs significantly between the generated and reference paragraphs. On the one hand, human annotators frequently reference objects, colours, and spatial relationships, suggesting a focus that may be more object-centric and literal. On the other hand, Claude uses more abstract and interpretive language (e.g., shows, suggesting, possibly, characteristic, capture, typical), as well as scene-level vocabulary (e.g. foreground, background, scene, landscape). Because fewer specific words (e.g., red, tree) appear prominently in its word cloud, we hypothesise that Claude may employ a broader vocabulary than human annotators, which we explore further in the linguistic analysis in the following section. However, due to the lack of specificity, we also suspect that Claude may be sticking to templates when generating paragraphs, which we also investigate further in the next section.

Linguistic Analysis

To gain deeper insight into Claude's true capabilities in image paragraph generation, we perform a linguistic analysis as described in the Methods section 4.2.1. We measure average paragraph length, standard deviation of length, lexical diversity (TTR), the proportions of nouns, verbs, and pronouns, vocabulary size, subordinating conjunctions frequency, and grammar error rate for both the reference paragraphs and those generated by Claude. The results clearly indicate that Claude exhibits superior linguistic performance compared to the ground-truth paragraphs in nearly every measured aspect. Indeed, Claude-generated paragraphs are nearly twice as long as the reference paragraphs, exhibiting greater lexical diversity and a significantly lower grammar error rate. Notably, their vocabulary size is nearly double, and they use more than three times as many subordinating conjunctions.

	Avg. Length	**St. Dev. Length**	**Diversity (TTR)**	**Nouns (%)**	**Verbs (%)**	**Pronouns (%)**	**Vocab Size**	**Subord. Conj.**	**Grammar Error Rate**
Claude 3.7 Sonnet	118.96	11.38	0.6788	29.36	12.17	3.83	11,739	2,831	0.06
Stanford Dataset	61.36	23.54	0.6172	27.92	9.81	5.82	7,034	903	0.21

Table 5.8: Linguistic and grammatical feature comparison between Claude 3.7 Sonnet and the Stanford Image Paragraph Captioning dataset.

Notes: Grammar Error Rate is calculated as errors per sentence. Vocabulary Size reflects the number of unique words. Subord. Conj. refers to subordinating conjunctions.

These findings imply that Claude-generated descriptions are not only more detailed, as seen by their average length and vocabulary size, but also more contextually rich and structurally complex, as reflected in their frequent use of subordinating conjunctions. In addition, Claude-generated paragraphs also contain more nouns and verbs overall and make use of a significantly larger vocabulary. This suggests that they not only identify objects and actions more frequently but also do so with greater specificity. This interpretation also aligns with the word cloud, where relatively few nouns and verbs appear prominently in Claude's outputs, suggesting greater lexical variety and specificity. We also note that the lower standard deviation in Claude's paragraph lengths suggests a preference for consistency, potentially at the expense of adapting to the varying informational content of each image. As for the lower proportion of pronouns, it remains unclear whether this reflects a reduced presence of coreference or a deliberate choice by the model in order to prioritise readability and avoid ambiguity.

To illustrate how these linguistic features manifest in actual paragraphs, we include images from our subset alongside the references and the generated descriptions. Clearly, Claude 3.7 produces paragraphs that are more detailed, coherent, and information-rich than those written by humans in the Stanford subset. Indeed, Claude's outputs feature low- to high-level semantics, with most falling into the mid- and high-level range. On the other hand, as expected from our linguistic analysis, the ground truth paragraphs produced by humans are generally more literal, primarily conveying low- to mid-level semantics. Claude is also more specific in naming entities, drawing on knowledge encoded in its parameters.

In the example shown in Table 5.9, while human annotators refer to the structure simply as a 'building', Claude not only recognises it as an art museum but also identifies the specific institution and wing, referring to it as the 'Art Institute of Chicago's Modern Wing'. It then situates the building within an architectural movement and interprets the meaning of the image through this lens, treating the museum as a symbol. Symbolism and 'aboutness' sit at the top of the high-level semantic hierarchy (Figure 4.1.1), showcasing the model's capacity for advanced conceptual interpretation. Even the low- and mid-level semantics the model chooses to present are relevant to the overall story it conveys. Rather than using generic terms like 'large' or 'cars', the model selects elements that are more relevant to the architectural movement at hand and the narrative it is trying to convey, such as 'glass', 'canopy', and 'panels'. This kind of abstraction also reflects the highest level of semantic reasoning. Further examples are depicted in Table 5.10 to Table 5.15.

Image ID: 2356347 **Source**: Stanford Image Captioning Dataset

Human Annotators: A large building with bars on the windows in front of it. There is people walking in front of the building. There is a street in front of the building with many cars on it.

Claude 3.7 Sonnet: This image shows the modern architecture of what appears to be the Art Institute of Chicago's Modern Wing. The building features a striking design with clean lines, extensive glass panels, and a distinctive overhanging roof structure that creates a sheltering canopy. The exterior combines light-colored stone or concrete with steel and glass elements. The street view shows pedestrian crossings marked with yellow warning signs, ornate street lamps, and several vehicles parked along the curb. The scene is captured on what looks like a sunny day with blue skies, and there's some greenery visible around the building's perimeter. The overall design reflects contemporary architectural principles with its emphasis on transparency, geometric forms, and the integration of natural light.

Table 5.9: Human annotators and Claude 3.7 Sonnet: Image ID 2356346.
Notes: high-level semantics
mid-level semantics
low-level semantics

Image ID: 2365091 **Source**: Stanford Image Captioning Dataset

Human Annotators: A large red and white train is traveling on tracks in a what looks to be a rural area. There are trees and hills in the background and the ground looks dry. The train has many large windows for the passengers to look out of. The train is mostly white with red on the front upper part of the train and red stripes and trim on the sides. The roof of the train is grey.

Claude 3.7 Sonnet: This image shows a Renfe regional train in Spain. The train unit appears to be a diesel multiple unit (DMU) in Renfe's distinctive white livery with red and gray stripes along its sides. The train is traveling on a rural or semi-rural track section, with rocky terrain and some trees visible in the background. There's a kilometer post marker visible beside the tracks marked "261." The rolling stock appears to be one of Renfe's Series 596 (or similar) trains commonly used for regional services on non-electrified lines. The train's front car is numbered "026M." The surrounding landscape suggests this is likely somewhere in Spain's interior, with characteristic Mediterranean vegetation visible along the railway line.

Table 5.10: Human annotators and Claude 3.7 Sonnet: Image ID 2365091

Notes: high-level semantics
mid-level semantics
low-level semantics

Image ID: 2388203 **Source**: Stanford Image Captioning Dataset

Human Annotators: The man is taking a photo in the round mirror. He is bald. He is wearing an orange jacket. His camera is black. There is a train in the mirror too.

Claude 3.5 Sonnet: This image shows a convex security mirror mounted on a wall in a subway or underground train station. The mirror reflects the station environment, including a train with blue and white cars that has stopped at the platform. There's a "TAKE THIRTY" sign visible in the reflection. The station has typical subway infrastructure with industrial ceiling elements visible. The mirror is mounted on what appears to be a textured concrete or stone wall, providing surveillance capability for the corner area. The setting has the characteristic utilitarian design of public transportation systems, with the mirror serving as a safety feature to help people see around the corner.

Claude 3.7 Sonnet: The image shows a curved security mirror mounted on a wall in what appears to be a London Underground station. The mirror reflects a person taking a photograph while wearing an orange/brown jacket. In the background, a subway train is visible at the platform with its distinctive blue and white colors. The mirror is positioned at a corner for safety and visibility purposes, allowing people to see around the bend. The station has a typical underground architecture with a curved ceiling and industrial design elements. There's some text visible in the mirror that appears to be a sign or station information. The overall scene captures the urban transit environment with its utilitarian safety features.

ChatGPT 4o: In a bustling London Underground station, a man in a red jacket stands in front of a convex safety mirror mounted on a concrete pillar, capturing a photograph of his reflection with a DSLR camera. The mirror distorts the surrounding scene, curving the tiled floor, fluorescent lights, and adjacent pillars into a fisheye perspective. Behind him, a white and blue train is parked along the platform, its windows revealing seated passengers and the familiar roundel logo of the London Underground. A bold black-and-white sticker reading "THE THIRST" is partially affixed to the mirror's surface, adding an urban, slightly gritty aesthetic. The scene blends functionality and artistry, offering a unique glimpse into the quiet, reflective moments within a normally fast-paced transit environment.

Table 5.11: Human annotators, Claude 3.7 Sonnet, Claude 3.5 Sonnet, and ChatGPT 4 descriptions: Image ID 2388203

Image ID: 2331342 **Source**: Stanford Image Captioning Dataset

Human Annotators: A silver train is on the ground. There is graffiti on the bottom of the train. There are rocks and grass next to the train.

Claude 3.7 Sonnet: The image depicts an old, abandoned train car sitting on a stretch of railroad tracks. The train car appears to be in a state of disrepair, with peeling paint and graffiti covering its exterior. The surrounding landscape is barren, with a clear blue sky overhead. The train car stands alone, a remnant of a bygone era, serving as a testament to the passage of time and the transient nature of human infrastructure. The scene evokes a sense of nostalgia and the ephemeral nature of progress, leaving the viewer to ponder the stories and histories that this lone train car has witnessed over the years.

Table 5.12: Human annotators and Claude 3.7 Sonnet descriptions: Image ID 2331342
Notes: high-level semantics
mid-level semantics
low-level semantics

Image ID: 2357480 **Source**: Stanford Image Captioning Dataset

Human Annotators: A street is flooded with brown water. The water is slightly choppy. There is a yellow and black traffic post in the water. There are traffic sign indicating the direction of cities. The sign is green. Over it is a blue street sign. Trees surround the area.

Claude 3.7 Sonnet: The image shows severe flooding on a road where water has completely covered the pavement. There are directional road signs visible showing routes to "Amiens" and "Texas," with the "Southern Hwy" sign partially visible at the top. A yellow and black striped barricade has been placed to block access to the flooded area, but the barricade itself is now partially submerged in the floodwaters. Trees can be seen in the background, also standing in water, indicating that the flooding has spread across the surrounding landscape. The brown, muddy floodwater appears to be flowing across what would normally be a road intersection, making it completely impassable to vehicles.

Table 5.13: Human annotators and Claude 3.7 Sonnet descriptions: Image ID 2357480
Notes: high-level semantics
mid-level semantics
low-level semantics

Image ID: 2354489 **Source**: Stanford Image Captioning Dataset

Human Annotators: A bunch of statues are displayed on a shelf. The shelf is a rustic wooden color. All of the statues are of animals except for the bell and empty bottle. The collection is very nice looking.

Claude 3.7 Sonnet: The image appears to be a wooden cabinet or shelving unit displaying a variety of interesting and diverse objects. On the shelves, there are several small figurines and sculptures, including a black toy monkey, a red toy bird, and a tan toy horse. There are also some glass bottles, including a clear glass bottle, as well as what looks like a vintage glass lamp or other decorative item.

The shelves also contain some plush or stuffed animal toys, including what appears to be a teddy bear or other fuzzy creature. There are also some other miscellaneous items, such as a red toy or figurine that resembles a cartoon character or superhero.

The overall impression is of a collection of eclectic and intriguing objects, each with its own unique character and story. The wooden shelving unit creates a rustic, vintage-inspired backdrop that adds to the sense of curiosity and wonder about the various items displayed. It's a visually engaging and thought-provoking image that invites the viewer to imagine the history and significance of the objects on the shelves.

Table 5.14: Human annotators and Claude 3 Haiku descriptions: Image ID 2354489

Notes: high-level semantics
mid-level semantics
low-level semantics

Image ID: 2396483 **Source**: Stanford Image Captioning Dataset

Human Annotators: People are on a play ground standing in the sand. A boy in a blue shirt and blue pants are climbing the pole. The woman in brown is holding a brown umbrella. There are people shadows on the ground.

Claude 3.7 Sonnet: The image shows a beach scene with several people enjoying a day at the sandy shore. In the foreground, someone is holding a black parasol or umbrella for shade. In the background, a person in blue shorts is climbing on what appears to be a pole or beach fitness equipment. There are several other individuals standing on the beach, and the setting includes palm trees and multi-story buildings or hotels lining the shore. The sandy beach stretches across the foreground, and the scene has a warm, vacation-like atmosphere. The urban beachfront setting suggests this is likely a popular coastal destination with residential or hotel properties facing the beach.

Table 5.15: Human annotators and Claude 3.7 Sonnet descriptions: Image ID 2396483
Notes: high-level semantics
mid-level semantics
low-level semantics

Through evaluation, linguistic analysis, and examples, it becomes clear that modern multimodal LLMs are capable of generating paragraph descriptions that surpass both our datasets and evaluation metrics. The image paragraph captioning task is not only challenging but also fundamentally important. Its value extends beyond applications, as it intersects with tasks like VQA and serves as a meaningful benchmark for evaluating reasoning and semantic capabilities in LLMs. Therefore, there is a pressing need for image-paragraph datasets that sufficiently test the semantic, reasoning, and cognitive capacities of LLMs. In the following section, we propose key considerations for designing a suitable image-paragraph dataset that meets these goals by assessing what current multimodal LLMs handle well and what remains challenging for them.

Discussion

Paragraphs are one of the primary ways we take in and express information. Naturally, this renders image paragraph description a fundamentally important task with broad practical and theoretical relevance. An automatic image paragraph description tool supports professionals such as researchers, doctors, creatives, journalists, and educators, who rely on understanding or expressing visual content in their work. In addition, it aids in archiving, search, and metadata enrichment by producing information-rich annotations for images. Another crucial application is providing access to the visually impaired by generating textual descriptions of visual content, which can be delivered audibly through assistive technologies.

Beyond its practical applications, the task holds theoretical significance and is closely related to another important and well-studied task: Visual Question Answering (VQA). Models often struggle to answer complex questions about images in a single step, particularly those requiring reasoning or outside knowledge. As CoT prompting suggests, starting with a structured description of an image, taking time to understand its content, and then attempting to answer the question can provide a strong foundation for more effective reasoning. In fact, the image paragraph captioning task is inherently similar to VQA in that it requires answering not just one question about an image, but many. For example, in order to produce a detailed description of an image, the model has to answer questions that range from *concrete*, such as 'What colour is this?', to more *abstract*, such as 'Is that detail important?'.

Moreover, the multiple-choice format commonly used in VQA datasets does not necessarily reflect the way people realistically phrase questions about images. Indeed, if an LLM chatbot user wanted to know who painted an artwork, they likely wouldn't ask 'a) Berthe Morisot, b) Claude Monet, c) Mary Cassatt, or d) Edgar Degas'. They would simply ask, 'Who painted this?' and expect a comprehensive response that includes not only the artist's name but also a description of stylistic characteristics, background information, and notable elements, conveyed in paragraph form. Another important consideration is explainability. Users would likely not trust a response that simply states 'Claude Monet', as it provides no explanation. In contrast, they are more inclined to trust a paragraph that vividly describes the artwork and highlights distinctive elements of Monet's style, such as fluid brushstrokes, glowing light, and natural scenery. Therefore, both the VQA and image captioning tasks would benefit from mutual exploration, and the possibility of integrating the two is worth further exploration.

We have demonstrated that datasets like the Stanford Image Description Dataset, which predate multimodal LLMs, do not reflect the true capabilities of these models or explicitly target the areas where they are most challenged. Given the significance of the task, there is a clear need for new datasets that truly challenge state-of-the-art MLLMs. To propose directions for creating datasets in the area of image paragraph captioning that are tailored to LLMs, we have identified the following key aspects that remain difficult for current multimodal models like Claude and GPT:

- **Optical Character Recognition (OCR)**: Recognising text within images is one of the most obvious ways in which multimodal LLMs fail when generating paragraph descriptions. Inaccurate OCR is costly not just because text conveys essential information, as in signs, signatures, diagrams, or labels, but also because humans easily recognise written language and therefore lose trust in the model when it makes clearly visible errors. A recent study [101] benchmarked Claude 3, Gemini 1.5, and GPT-4o, and found that they achieve strong OCR performance, surpassing traditional computer vision models in accuracy on certain tasks. However, we find that Claude often struggles with simple OCR when tasked with image paragraph description rather than explicit OCR tasks. We believe this may be due to the model being overwhelmed by the complexity of the task, focusing on broader aspects of the visual context and paragraph generation, and therefore failing to recognise simple text. For instance, in the example prsented in Table 5.11, Claude 3.5 Sonnet misreads the text, while Claude 3.7 Sonnet does not reference it at all. Similarly, in the case of Table 5.13, ambiguity in the signage leads Claude 3.7 Sonnet to hallucinate the content of the sign. We observe that Claude tends to include text in its descriptions, even when it is uncertain about its accuracy or has entirely fabricated the content. Because image paragraph captioning does not explicitly ask the model to recognise or reference text, its decision to include such content, even when uncertain, highlights the persistent issue of hallucinations in multimodal LLMs.

- **Spatial Reasoning**: Images with complex spatial relationships are generally more challenging for models because they require longer chains of reasoning to infer mid-level semantics within the scene. While spatial reasoning may be associated with lower-level semantics, it also requires a degree of abstraction, as the model must interpret spatial cues in an in-context manner. For example, the phrase 'on the left of the car' can be ambiguous. Depending on context, it might refer to a person standing to the left of the car from the viewer's perspective, or it could describe something physically on the left side of the car, such as a sticker, or a dent. Understanding which interpretation is correct, depending on the context, requires a higher level of linguistic capability that not all MLLMs can navigate. That said, paragraphs can be more forgiving, as what matters most is that the content presented is accurate, leaving the interpretation to the viewer. The true challenge arises when images contain both complex spatial relationships and require intensive reasoning at the same time. This is what causes state-of-the-art models, which otherwise perform well on spatial reasoning datasets, to struggle during image paragraph captioning.

- **Hallucinations**: Perhaps the most straightforward type of hallucination in LVLMs and MLLMs occurs when the visual encoder misinterprets what it sees or generates features that aren't actually present in the image. The language model then assumes the visual encoder's output is correct, constructing its descriptions based on inaccurate representations. In most architectures, the visual encoder produces the same image representation regardless of the task or prompt. This means that whether the model is generating a short caption or a detailed paragraph, it relies on the same fixed set of visual features. As a result, when the required output is longer, the model often has to fill in the gaps, which increases the risk of hallucination. In the example of Table 5.14, the hallucination of the 'vintage glass lamp' appears when the model is prompted to generate a longer paragraph.

  Another type of hallucination originates from the language model itself. In paragraph captioning, models tend to prioritise narrative flow over faithful visual description. Generating a paragraph caption of an image is not only linguistically demanding, but also requires a sense of storytelling. After all, the core strength of language models, even multimodal LLMs, still lies in their fluency. We observe that although multimodal LLMs possess a strong visual understanding of the image, the paragraph content often follows the flow of language, rather than the image itself. As a result, the model struggles not only with choosing which concepts to present, but also with hallucinating content. The man in the image, for example, has high visual weight due to his central placement, strong colour contrast, and the focal emphasis created by the skewed mirror. He also has semantic salience, as he is the one taking the photo. Yet, Claude 3.5 Sonnet still does not include him in the description of the image. The omission is likely not due to a lack of recognition, but rather an instance of the model's narrative drifting away from him. This tendency to prioritise language in paragraph captioning can cause the model not only to overlook salient elements, but also to introduce objects that aren't actually present. When a key figure like the man is excluded, the model still needs to maintain paragraph coherence, so it may compensate by inventing other elements to fill narrative gaps. Uncertainty is another factor that leads to hallucinated content. In the example presented in Table 5.13, the model misreads the sign entirely. Since the sign is not clearly visible, a human would likely avoid referencing it at all, but the model instead chooses to fabricate its content. This suggests that the language component might treat visual features as context, not constraint.

- **Outside Knowledge**: Images like the train presented in Table 5.10, which require factual knowledge about the world to be fully understood and described pose a greater challenge to multimodal models than those that can be interpreted through visual cues alone. The task becomes more difficult when models must not only access knowledge, but also reason around it, particularly to infer high-level meaning. Knowledge-based image paragraph captioning is visually, linguistically, and cognitively complex, so, as with OCR, the challenge lies in having to perform challenging functions simultaneously. Activating the correct background knowledge from the model's parameters is also more difficult when the input is strictly visual. Unlike text prompts that can probe relevant knowledge more directly, images are often ambiguous or contain cues that don't map explicitly to specific concepts in the model's training data. As a result, the model may fail to activate the correct background knowledge.

- **Commonsense Knowledge**: Among all the knowledge clusters required for generating image paragraphs, commonsense knowledge is where LLMs are most capable. State-of-the-art multimodal LLMs generally understand actions, events, object relationships, interactions, and purposes, and can use them to move up levels of abstraction. Areas where LLMs still struggle include implied actions, object purposes, causal reasoning, physical reasoning, and temporal understanding.

- **Emotional Knowledge**: We find that this is the level of abstraction at which models begin to struggle. LLMs can often identify and express the emotions of people that are either explicitly or implicitly conveyed with relative ease. However, while they can mimic emotion-laden language and, in fact, frequently discuss emotions in image paragraphs, they often don't interpret the mood and affect of images accurately. Although they accurately identify core emotions, determining the intensity with which an image evokes that emotion proves more challenging, as illustrated in the examples of the train (Table 5.12) and the collection (Table 5.14). The image of a toys' collection wouldn't typically provoke thought in a real person, nor would the viewer genuinely 'ponder the stories and histories that the lone train had witnessed over the years'. Of course, Claude is known for its prose, but the issue is not with the prose itself. Rather, it lies in its difficulty in accurately gauging the intensity of emotions evoked by images and using prose to express them effectively.

- **Aesthetic Knowledge**: When it comes to deeply understanding and articulating the aesthetic appeal of an image, LLMs often fall short. Aesthetic perception is highly subjective and can vary widely between individuals, cultures, and contexts. While LLMs do point out certain aspects of the image as 'distinctive', 'striking', or 'beautiful', it is still difficult for them to integrate subtle visual elements, such as lighting, to unified descriptions that reflect the overall aesthetic appeal of the image. As a result, they may fail to capture the essence of the image or phrase it in a way that feels unnatural to humans.

- **Inductive Interpretative Knowledge**: Inductive image interpretation is an integral component of high-quality paragraph captioning and sits at the top of high-level semantics. Surprisingly, a significant portion of LLM-generated paragraphs employs this type of knowledge, incorporating symbolism (e.g., Table 5.12), 'aboutness' (e.g., Table 5.15), and abstract concepts (e.g., Table 5.12). Most state-of-the-art MLLMs, including Claude 3, GPT-4o, and Gemini 2.0, typically include at least one sentence that refers to the meaning—or 'aboutness'—of the image when asked to describe it. It is likely that these models have learned to follow a structured template, such as 'the image shows', followed by a discussion of the image's 'aboutness'. This behaviour could have been reinforced through fine-tuning on specific datasets or human feedback during training, or the patterns could have been learned during pretraining. The meaning LLMs extract about the image is often accurate and represents the most high-level and commendable aspect of their generated paragraphs. What they struggle with at this level is **concept selection** and **name specificity**. First of all, generating an image paragraph requires a certain level of abstraction in selecting the elements to discuss. For instance, in the Table 5.11 example, Claude 3.5 Sonnet discusses the sticker in the skewed mirror but completely overlooks the man taking the picture. This is clearly a poor decision regarding what to include in the description. LLMs also struggle with specificity. In fact, we notice that LLMs tend to be more specific than humans in their descriptions, which can make them sound unnatural and lead to hallucinations. For example, it is impressive that Claude 3.7 mentions the rolling stock type of the train (Series 596) (Table 5.10), yet it is highly unlikely that a human would include such a detail in their description. Given that the LLM must perform many diverse tasks to generate a paragraph, focusing too much on overly specific descriptions may hinder its ability to effectively carry out other functions, such as high-level reasoning.

Subordinating Conjunctions:
Time after, as, as long as, as soon as, before, once, since, still, until, when, whenever, while
Cause and effect as, because, even though, since, so that, though, although, now that
Condition if, even if, only if, unless, provided that, in case, in the event that
Contrast / Concession although, even though, though, whereas, while, rather than
Purpose in order that, so that
Comparison / Manner as if, as though, the way, than

Table 5.16: List of Subordinating Conjunctions categorized by function

5.2.2 Multimodal Collaboration for Knowledge-based VQA

Baseline

We use Claude 3.7 Sonnet as our baseline, providing it with images and questions from the OK-VQA validation split, using the inference parameters detailed in Table 5.3. The model is prompted to produce one- or two-word answers and to always attempt to answer the question, in a zero-shot fashion. We measure accuracy using the soft accuracy metric described in Section 4.1.2. The average accuracy is 43.59%, which is established as our baseline. The results across categories are shown in Table 5.17. The highest result is obtained in the Cooking and Food category, while the lowest is observed in the Sports and Recreation category.

Avg. (%)	**VT**	**BCP**	**OMC**	**SR**	**CF**	**GHLC**	**PEL**	**PA**	**ST**	**WC**	**Other**
43.59	41.43	44.88	45.70	40.63	47.53	44.54	41.12	41.42	44.05	46.67	46.01

Table 5.17: Claude 3.7 Sonnet baseline accuracies for the OK-VQA dataset
Notes: VT = Vehicles and Transportation, BCP = Brands, Companies and Products, OMC = Objects, Material and Clothing, SR = Sports and Recreation, CF = Cooking and Food, GHLC = Geography, History, Language and Culture, PEL = People and Everyday Life, PA = Plants and Animals, ST = Science and Technology, WC = Weather and Climate.

We observe that, despite our prompt, we are unable to steer the model toward producing predominantly one- or two-word answers. In fact, since the dataset contains no answers longer than three words and includes only specific answers, the model loses at least 12.84% of potential accuracy by generating responses containing four or more words, and at least 1.19% by not answering the question at all (Figure 5.2.3). This suggests that LLMs cannot be effectively steered through prompting to adhere to a specific format while performing reasoning- and knowledge-intensive tasks.

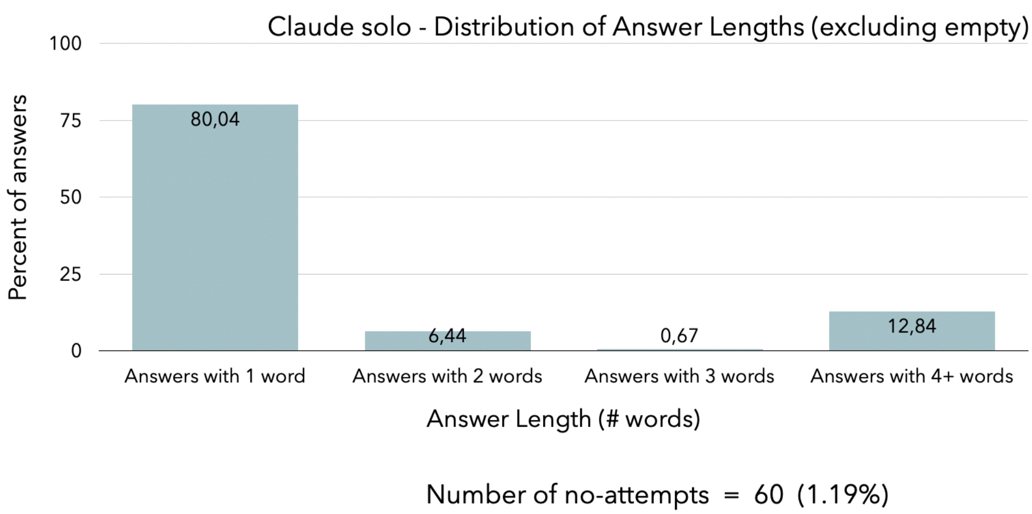

Figure 5.2.3: Word length distribution and number of no-attempts for the Claude 3.7 Sonnet baseline.

The Scout, Analyser, and Resolver Framework

For our **Scout** model, which generates *greedy descriptions* for each image, we use the inference parameters detailed in Table 5.2. We conduct a linguistic analysis, as described in Section 4.2.1, to ensure the quality of our descriptions Table 5.18. We also present a word cloud of the descriptions to gain insight into the most frequently occurring terms in the obtained paragraphs (Figure 5.2.4).

Avg. Length	St. Dev. Length	Diversity (TTR)	Nouns (%)	Verbs (%)	Pronouns (%)	Vocab Size	Subord. Conj.	Grammar Error Rate
115.65	18.15	0.6774	29.29	12.00	4.26	11,904	2,357	0.06

Table 5.18: Linguistic analysis for Scout-generated image paragraphs for the OK-VQA dataset.

Notes: Grammar Error Rate is calculated as errors per sentence. Vocabulary Size reflects the number of unique words. Subord. Conj. refers to subordinating conjunctions.

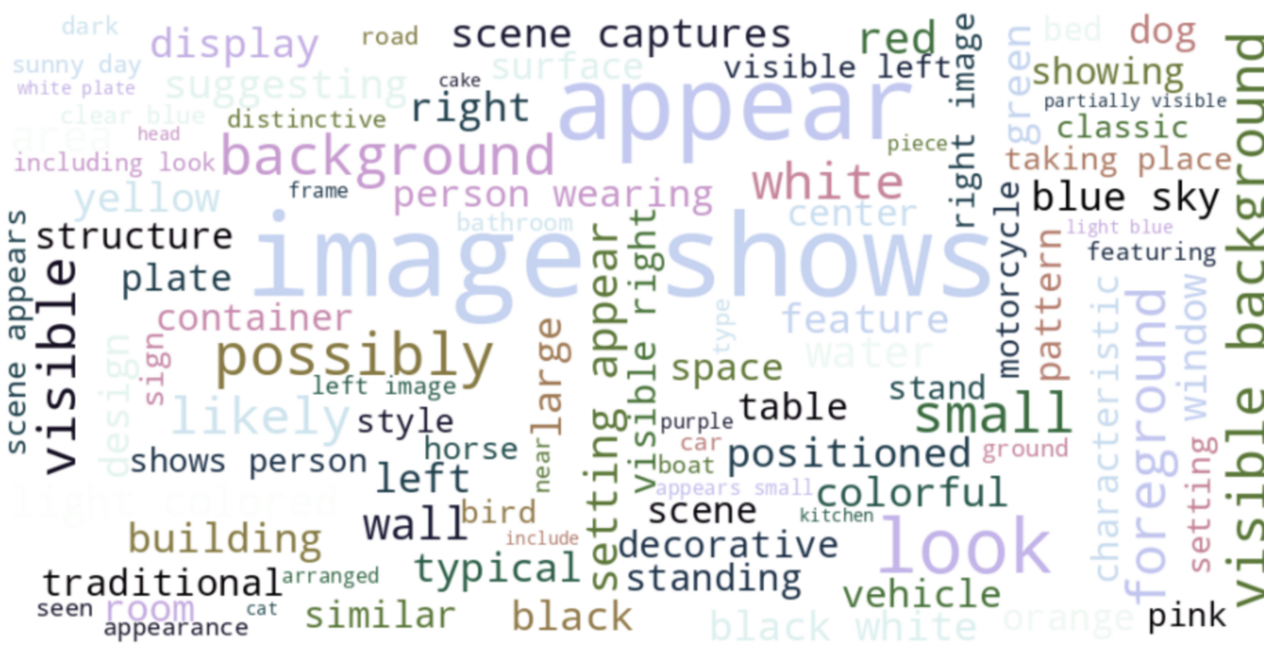

Figure 5.2.4: Word cloud for Scout-generated image paragraphs for the OK-VQA dataset.

To find salient sentences in the generated paragraphs, we split the model's answer into individual sentences and evaluate each one separately. We use a Sentence-BERT model to create embeddings that capture the meaning of each sentence. We compare each sentence to the question at hand using cosine similarity. Finally, we rank the sentences by similarity and select the top ones to extract the most relevant part of the description. Clearly, our Scout model produces semantically rich and relevant descriptions. For example, the Table 5.19 shows the salient sentences generated by our Scout model for the motorcycle image.

Salient sentences in Scout-generated image descriptions

Question: What sport can you use this for?

Image Description (Scout): The image shows a black Honda motorcycle, likely a Varadero or similar adventure touring model, parked on a sandy/paved surface. The motorcycle features a prominent windshield, comfortable seat, and silver accents on the rear section. It has the characteristic dual-purpose design with wire-spoke wheels that are suitable for both on and off-road riding. In the background, there's a white metal container or storage unit, and what appears to be an orange flag or banner visible on the left side. The setting looks like an outdoor area with some dirt/sand terrain and a paved section where the motorcycle is standing on its kickstand. The scene suggests this might be at a motorcycle event, test riding area, or off-road riding location.

Table 5.19: Salient sentences in Scout-generated image descriptions.
Notes: most salient sentence
second most salient sentence
third most salient sentence

For our **Analyser** model, which generates the initial answer based on the image description, we use the Llama 3.3 model with the inference parameters detailed in Table 5.5. For our **Resolver model**, which extracts the final answer and formats it to meet the requirements, we use the Llama 3.3 model with the inference parameters described in Table 5.6. We explore both zero-shot and few-shot approaches for the Resolver's predefined formatting rules and observe a slight improvement in accuracy with the few-shot approach (Table 5.20).

Our collaborative Scout-Analyser-Resolver with the few-shot resolver approach achieves an accuracy of **48.73%**, which is 5.14% higher than the baseline, suggesting the effectiveness of a collaborative approach for the OK-VQA task. The complete results for each category are shown in Table 5.20.

Method	Avg. (%)	VT	BCP	OMC	SR	CF	GHLC	PEL	PA	ST	WC	Other
Scout-Analyser-Resolver Zero-Shot	47.05	44.38	52.21	43.13	47.51	47.70	53.76	46.26	47.77	35.00	48.84	49.82
Scout-Analyser-Resolver Few-Shot	48.73	46.03	54.42	43.97	48.78	49.89	52.62	46.82	51.00	35.71	50.39	51.15

Table 5.20: Scout, Analyser, and Resolver accuracies for the OK-VQA dataset.
Notes: VT = Vehicles and Transportation, BCP = Brands, Companies and Products, OMC = Objects, Material and Clothing, SR = Sports and Recreation, CF = Cooking and Food, GHLC = Geography, History, Language and Culture, PEL = People and Everyday Life, PA = Plants and Animals, ST = Science and Technology, WC = Weather and Climate.

We validate the effectiveness of our Resolver as a formatter by analysing the word length distribution and the number of no-attempts, which were two of our main objectives for formatting the answers (Figure 5.2.5). Indeed, the proportion of answers with four or more words was reduced from 12.84% to almost zero, compared to the baseline. Additionally, the number of no-attempts dropped from 60 to 2. However, the increase in accuracy (5.14%) is not as significant as we would expect given the dramatic improvements in formatting. We attribute this to the fact that when the model diverges from the format or chooses not to answer the question, it often indicates a deeper lack of knowledge about the answer.

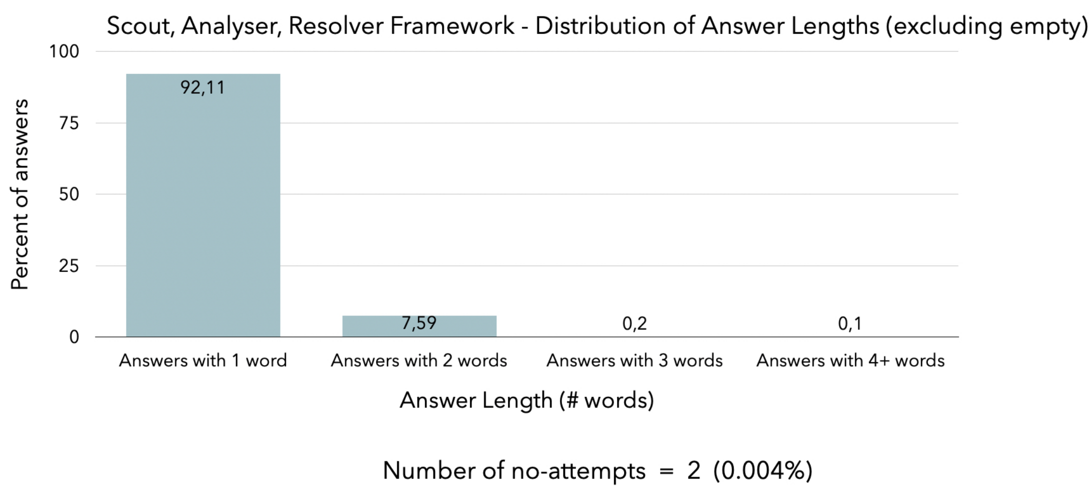

Figure 5.2.5: Word length distribution and number of no-attempts for the Scout, Analyser, Resolver framework.

We identify six underlying reasons for the shortcomings of our system.

1. **Inaccurate Information**: This error occurs when the necessary visual information to answer the question is conveyed in the image description but inaccurately. Occasionally, the Scout may hallucinate objects, attributes, or relationships and deliver them to the Analyser with fluency. This causes the subsequent models to propagate the error and produce incorrect answers with confidence.

2. **Missing Information**: Here, the necessary visual information to answer the question is not conveyed at all in the image description. This may occur because the Scout fails to reference important aspects of the image, and because the descriptions are generated in a greedy manner. As a result, the descriptions do not necessarily prioritise transferring the information needed to answer the question.

3. **Reasoning Error**: Despite the information being present in the description, the Analyser might still produce an incorrect answer. This highlights a reasoning error or a failure to correctly extract and infer the answer from the external knowledge that these questions typically require.

4. **Selection Error**: The Resolver selects the wrong final answer, either by failing to comply with the formatting instructions or by phrasing the answer in a way that does not sound natural to humans.

5. **Evaluation Error**: The framework produces a valid answer but does not receive credit because it is not formatted identically to the ground truth answers, even after post-processing. This is a common issue in open-ended datasets and represents a key reason why multimodal LLMs underperform relative to their true capabilities on the OK-VQA dataset. It also explains why fine-tuning on this dataset yields significantly better results.

6. **Abstraction Bias**: The framework produces a valid answer that differs from the ground truth in terms of specificity. Questions about images can generally be categorised into three levels of abstraction: **specific**, **intermediate**, and **abstract**. The abstraction level of a question is specific when it expects entity recognition or terminal hyponym like a brand, date, or species (e.g., 'What brand is this? → Coca-cola' or ' What singer is associated with this hairstyle? → Robert Smith'). In contrast, questions that ask for classification into a general type or category have an intermediate level of abstraction (e.g., 'What animal is this? → monkey' or 'What drink is this? → Wine'). Finally, abstract questions might ask for purpose, use, or general real-world reasoning and expect more general concepts (e.g., 'What is the purpose of this device? → Entertainment' or 'Where is the boy standing? → Ground'). LLMs have a specificity bias: they tend to respond to questions with a more specific level of abstraction than humans typically would. At the same time, they struggle to transition between abstraction levels while reasoning. Some questions can be answered correctly at three different levels of abstraction. For example, the answer to the question 'Who designed the statues?' could be *Jackie McKenna* (specific), *an artist* (intermediate), or *a human* (abstract). All answers are technically valid but differ in their level of abstraction. When seeing the question, humans might look for the name of the person who designed the statues. However, when it becomes clear that this cannot be determined from looking at the image, they naturally generalise upward, providing a more general answer like 'an artist'. Conversely, LLMs struggle with this kind of flexibility when it comes to abstraction levels. Upon seeing the question, they look for a precise answer and when they don't find it they don't generalise upward but rather choose to not answer the question at all. The abstraction bias presents a challenge in the OK-VQA dataset and other open-ended datasets and is a key reason why LLMs lose credit despite knowing the answer or why they might fail to answer the question altogether.

Example 1: Correct Answer

Question: Which part of this animal would be in use if it was playing the game that is played with the items the man is holding?

Image Description (Scout): The image shows a person playing frisbee in a grassy yard while a dog is perched on their back. The person is wearing a white t-shirt, dark pants, and a black cap, and is bent forward holding two white frisbees - one in each hand. The dog appears to be a black and white spotted breed, possibly an Australian Cattle Dog or similar mix, and has impressively balanced itself on the person's back while they're in motion. In the background, there's a wooden fence. This looks like a fun moment of dog training or freestyle disc dog practice, where the dog has learned to perform tricks while the handler manipulates the frisbees.

Initial Answer (Analyser): If the dog was playing the game that is played with the items the man is holding (frisbees), the part of the animal that would be in use is its mouth, as dogs typically catch frisbees in their mouths.

Final Answer (Resolver): Mouth ✓

Ground Truth: mouth

Table 5.21: Example 1: Correct Answer.
Notes: most salient sentence
second most salient sentence

Example 2: Inaccurate Information

Question: Which phase is that moon in?

Image Description (Scout): The image shows a clock tower with two visible clock faces on adjacent sides. The tower appears to be made of dark wooden shingles or similar material and has a small cupola or bell housing at the top with vented openings. The clock faces have Roman numerals, with one clock showing approximately 4:35. In the background, there's a crescent moon visible in the gray sky, and the silhouette of an evergreen tree can be seen next to the tower. The photo has a moody, atmospheric quality with its muted colors and twilight setting. ✗

Initial Answer (Analyser): The moon is in the crescent phase.

Final Answer (Resolver): crescent

Ground Truth: half

Error Type: Inaccurate Information

Table 5.22: Example 2: Inaccurate Information
Notes: most salient sentence
second most salient sentence

Example 3: Missing Information

Question: Name the type of plant this is.

Image Description (Scout): This image shows a modern bathroom with a striking biophilic design element. The bathroom features dark gray tile walls and a glass shower enclosure on the right side. What makes this space unique is a dramatic vertical garden or living wall that runs from floor to ceiling, bringing lush greenery into the bathroom environment. The plants create a vibrant contrast against the dark walls. The bathroom has a white vanity with a sink on the left side, some open shelving with towels, and recessed lighting in a wooden ceiling panel. Through an opening in the background, you can glimpse part of another room that appears to be a living area. The overall design combines contemporary bathroom fixtures with natural elements, creating a spa-like atmosphere where nature is integrated directly into the interior architecture. ✗

Initial Answer (Analyser): Based on the description of the "dramatic vertical garden or living wall" with "lush greenery", it's likely that the type of plant is a species commonly used in indoor vertical gardens or living walls, such as a Fern or an English Ivy, but given the general description, a more encompassing answer would be that the plants are likely some type of Tropical Plant or Greenery, possibly including species like Pothos or Peace Lily, which are popular for indoor use due to their ability to thrive in a variety of lighting conditions and their aesthetic appeal. However, without more specific details, the most accurate general category would be Tropical Plants or Greenery.

Final Answer (Resolver): fern

Ground Truth: vine

Error Type: Missing Information

Table 5.23: Example 3: Missing Information
Notes: most salient sentence
second most salient sentence

Example 4: Reasoning Error

Question: How fast is the vehicle pictured capable of reaching?

Image Description (Scout): The image shows a scenic railway journey through a mountainous landscape. There's an open-air train car or carriage traveling along tracks that wind through rolling hills and valleys. The railway appears to be a heritage or tourist line, with passengers seated on what looks like an observation platform or open wagon. The surrounding landscape is quite picturesque, featuring a mix of forested hills with evergreen trees covering the higher elevations, while the lower areas show a patchwork of fields and meadows. In the background, mountains rise up with their tops partially obscured by clouds, suggesting this might be in a highland region. The scene has a somewhat overcast sky, giving the landscape a moody, atmospheric quality. The railway track is built along a stone embankment that curves through the countryside, allowing travelers to enjoy panoramic views of the beautiful rural scenery as they journey through this mountainous terrain.

Initial Answer (Analyser): The vehicle, being a train on a heritage or tourist line, is likely capable of reaching moderate speeds, but not extremely high speeds due to the winding nature of the track and the scenic, tourist-oriented purpose of the railway. A reasonable guess would be that it can reach speeds of around 30-50 km/h (18-31 mph), allowing passengers to enjoy the views without rushing through the landscape.

Final Answer (Resolver): 30 km/h ✗

Ground Truth: 80 mph

Error Type: Reasoning Error

Table 5.24: Example 4: Reasoning Error
Notes: most salient sentence
second most salient sentence

Example 5: Selection Error

Question: What type of plane is that?

Image Description (Scout): The image shows an airport tarmac with multiple aircraft. In the foreground is a Qantas aircraft with the distinctive red kangaroo logo on its tail, labeled "Spirit of Australia." The plane is connected to a jetway or boarding bridge. In the background, another Qantas aircraft with the same red and white livery can be seen on the runway, and there's also a plane taking off in the distance against a clear blue sky. The scene appears to be taken during daytime at what is likely an Australian airport, showing the typical airport operations with planes parked at gates and others in motion on the taxiways or runways.

Initial Answer (Analyser): The plane is a Qantas aircraft, likely a commercial airliner such as a Boeing 737 or Airbus A330, given the context of the airport setting and the presence of a jetway.

Final Answer (Resolver): Airliner ✗

Ground Truth: commercial

Error Type: Selection Error

Table 5.25: Example 5: Selection Error
Notes: most salient sentence
second most salient sentence

Example 6: Evaluation Error

Question: What toy is this?

Image Description (Scout): The image shows a person in a red sweater sitting in what appears to be a church or chapel pew. They have a small teddy bear or stuffed animal tucked into the back of their backpack or bag The setting has wooden pews with white panels, and there are a few other people visible in the background, seated in the rows ahead. The atmosphere has a warm, soft focus quality to it, creating an intimate feeling within what seems to be a place of worship or gathering. The stuffed animal peeking out from the backpack adds a touching, personal element to the scene.

Initial Answer (Analyser): The toy is a teddy bear or a stuffed animal.

Final Answer (Resolver): teddy

Ground Truth: stuffed animal, teddy bear ✗

Error Type: Evaluation Error

Table 5.26: Example 5: Evaluation Error
Notes: most salient sentence
second most salient sentence

Example 7: Abstraction Bias

Question: What does this grow from?

Image Description (Scout): This image shows a beautiful orange lily flower in full bloom, displayed in a small metallic container or vase. The lily has vibrant orange-salmon colored petals that are fully opened, revealing its distinctive protruding stamens with dark reddish-brown anthers. The flower's elegant form creates a striking contrast against the dark blue or black fabric background. The metallic container has a simple, sleek design with a silver finish that complements the warm tones of the flower. The lighting in the image highlights the delicate texture and translucent quality of the lily's petals, creating a dramatic and elegant presentation.

Initial Answer (Analyser): This lily flower grows from a bulb. ✗

Final Answer (Resolver): Bulb

Ground Truth: ground

Error Type: Abstraction Bias

Table 5.27: Example 7: Abstraction Bias
Notes: most salient sentence
second most salient sentence

Collaboration Analysis and Discussion

To examine our models' sensitivity to the design of the collaboration framework and to gain deeper insights into the system's errors, we conduct multiple experiments and report the following results (Figure 5.2.6).

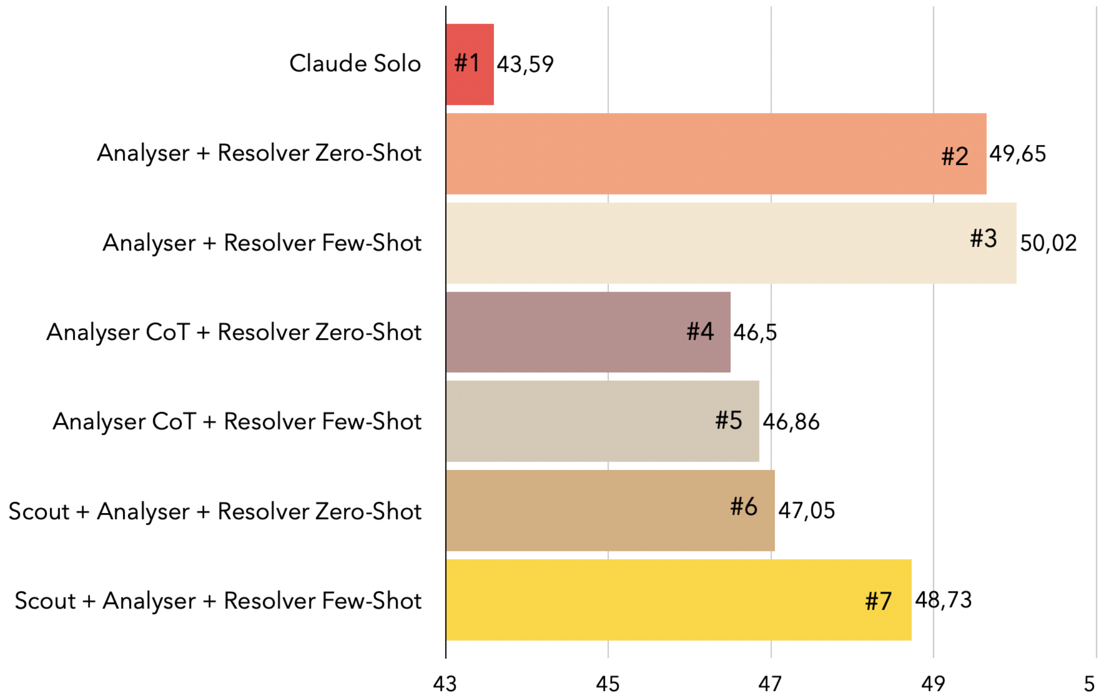

Figure 5.2.6: Word cloud for Scout-generated image paragraphs for the OK-VQA dataset.

Our baseline, shown as #1, has the lowest accuracy among the methods tested, indicating that multimodal LLMs cannot be effectively steered through prompting to perform reasoning-intensive tasks while adhering to formatting rules. The Resolver appears to be the most effective component of our framework, significantly improving results across all other methods. We experiment with using only the Analyser and the Resolver, ablating the prompting techniques used for both components. We use the inference parameters shown in Table 5.3 and the prompt detailed in Table 4.10 for our Analyser in method #2, and Table 5.6 and Table 4.8 for our Resolver. Then, for our Analyser in method #3, we apply the inference parameters shown in Table 5.3 and the prompt in Table 4.10, while for the Resolver, we use the settings and prompt detailed in Table 5.6 and Table 4.9. The third method yields the best overall average accuracy, highlighting the effectiveness of our Resolver, particularly when used in a few-shot manner. For our #4 and #5 method using CoT, we use the extended thinking version of our Analyser, as shown in Table 5.4. We then prompt our Resolver using zero-shot (Table 4.8) and few-shot prompting (Table 4.9), with the inference parameters shown in Table 5.6. Finally, for our #6 and #7 method we use all the components of our system and experiment with either zero-shot (#6) and few-shot (#7) prompting. The results of all methods accross caregories are presented in Table 5.29.

We observe that all of our methods yield better results than using Claude alone, highlighting the effectiveness of a collaborative framework for this task. We believe that Claude significantly underperforms on this dataset relative to its true capabilities due to two main reasons: (a) its inability to be steered to adhere to specific formatting rules (such as providing single-word answers or using singular form), and (b) the limitations of the evaluation metric used in the dataset. Indeed, a simple post-processing step of resolving and formatting the final answer (#3) boosts results by almost 7%, indicating that much of the knowledge being tested is already present but cannot be probed in a single-step through prompting. The Scout-Analyser-Resolver pipeline (#7) proves effective in handling the OK-VQA task and achieves the best results across three categories. It is also transparent, allowing us to observe the reasoning behind the LLMs' answers across the different stages of the system. The average performance of methods #6 and #7 is significantly impacted by lower accuracies in the Science and Technology, as well as the Objects, Material, and Clothing categories. We attribute this to questions in these categories requiring finer-grained details that are typically not captured in the greedy image-paragraph descriptions generated by the Scout model. To address this, controlled paragraph captioning could be explored in future work to better account for the missing information error type that compromises performance.

We believe that the vast majority of errors in our system arise from answer selection, formatting, and evaluation issues and do not necessarily indicate a lack of knowledge on the model's part. Because none of our methods are fine-tuned on the dataset, the model does not learn dataset-specific patterns, such as preferring 'horn' over 'horns' or 'swim' over 'swimming'. To compensate for this disadvantage, we introduced a Resolver model responsible for choosing the final answer and applying formatting rules, which significantly improved results. We find that LLMs are powerful formatters and can be effectively used in automated tasks involving the extraction and post-processing of other models' outputs. However, as statistical models, LLMs can only be guided toward a specific format to a certain extent. It goes against their learned patterns to answer the question 'Is he catching or throwing?' with 'throw' instead of 'throwing'. Yet the ground truth answer is in fact 'throw', and because the proposed evaluation metric does not normalise variations like the -ing form, the model receives zero accuracy for this answer.

We also believe that the established accuracy metric for the OK-VQA dataset is too rigid, excluding many answers that are actually correct while rewarding others that are not. For example, in response to the question 'What South American country usually has this climate?' several answers could be correct, including 'Brazil', 'Argentina', and 'Ecuador'. However, the model's answer 'Argentina' receives zero points because it is not among the ten ground-truth answers, while an incorrect answer like 'Africa'—which is neither a country nor located in South America—would receive 60% accuracy because 2 out of 10 annotators gave that answer. In the lily example (Table 5.27), although the model's answer 'bulb' is perfectly valid, it also receives zero accuracy because 'ground' was the most common human response. The problem here is not incorrectness, but rather providing an answer at a different level of abstraction compared to human responses. Finally, 'Giorgio Armani' and 'Armani' are both commonly used to refer to the same brand, with 'Armani' being the official legal name of the brand and 'Giorgio Armani' referring to the designer. However, the model's answer 'Armani' gets zero points when answering the question 'What brand of suit is the man in the image wearing?'.

While the OK-VQA dataset sets out to test models' ability to leverage external knowledge to answer challenging questions about images, this goal is undermined by shortcomings in the evaluation metric and biases present in the dataset. The A-OKVQA dataset addresses these limitations by adopting a multiple-choice format, which accounts for the significantly improved performance of multimodal LLMs when tested on it. Nevertheless, an open-ended dataset presents a much greater challenge for current LLMs, testing not only their reasoning and cognitive abilities but also their capacity to respond at the same level of abstraction as humans, while also reducing the possibility of guessing. This is why we propose a human survey to more accurately assess the true performance of multimodal LLMs on open-ended OK-VQA, along with new evaluation metrics that could potentially leverage LLMs as automatic evaluators. Evaluating LLMs' ability to respond at the same level of abstraction as humans represents another important direction for future work, which could be explored through datasets with multiple correct answers across varying abstraction levels.

Method	Word distribution*		Accuracy (%)	# of no attempts
Claude Solo	1	80.04	43.59	60
	2	6.44		
	3	0.67		
	4+	12.84		
Analyser + Resolver Zero-Shot	1	96.27	49.65	61
	2	3.47		
	3	0.10		
	4+	0.16		
Analyser + Resolver Few-Shot	1	96.21	50.02	128
	2	3.09		
	3	0.10		
	4+	0.59		
Analyser CoT + Resolver Zero-Shot	1	81.81	46.50	130
	2	17.90		
	3	0.26		
	4+	0.04		
Analyser CoT + Resolver Few-Shot	1	87.44	46.86	48
	2	12.33		
	3	0.24		
	4+	0.00		
Scout + Analyser + Resolver Zero-Shot	1	92.63	47.05	1
	2	7.09		
	3	0.22		
	4+	0.06		
Scout + Analyser + Resolver Few-Shot	1	92.11	48.73	2
	2	7.59		
	3	0.20		
	4+	0.10		

*Word distribution refers to the percentages of answers with 1 word, 2 words, 3 words, and 4+ words respectively.

Table 5.28: Comparative results of all methods on the OK-VQA dataset.

Method	**Avg. (%)**	**VT**	**BCP**	**OMC**	**SR**	**CF**	**GHLC**	**PEL**	**PA**	**ST**	**WC**	**Other**
Claude Solo	43.59	41.43	44.88	45.70	40.63	47.53	44.54	41.12	41.42	44.05	46.67	46.01
Analyser-Resolver Zero-Shot	49.65	47.06	47.67	51.68	46.74	52.94	46.81	**48.36**	50.33	**50.71**	52.56	50.44
Analyser-Resolver Few-Shot	**50.02**	**47.38**	48.26	**51.92**	47.62	**53.14**	46.10	**48.36**	50.53	**50.71**	53.80	**51.24**
Analyser CoT-Resolver Zero-Shot	46.50	40.54	46.63	46.40	46.07	51.25	44.96	44.53	46.16	43.10	54.11	49.66
Analyser CoT-Resolver Few-Shot	46.86	41.60	46.98	47.20	48.29	49.84	46.67	44.63	45.47	44.29	**55.66**	50.37
Scout-Analyser-Resolver Zero-Shot	47.05	44.38	52.21	43.13	47.51	47.70	**53.76**	46.26	47.77	35.00	48.84	49.82
Scout-Analyser-Resolver Few-Shot	48.73	46.03	**54.42**	43.97	**48.78**	49.89	52.62	46.82	**51.00**	35.71	50.39	51.15

Table 5.29: Comparative results of all methods on the OK-VQA dataset across all categories.
Notes: VT = Vehicles and Transportation, BCP = Brands, Companies and Products, OMC = Objects, Material and Clothing, SR = Sports and Recreation, CF = Cooking and Food, GHLC = Geography, History, Language and Culture, PEL = People and Everyday Life, PA = Plants and Animals, ST = Science and Technology, WC = Weather and Climate.

Chapter 6

Conclusion and Future Work

Multimodal LLMs represent some of the most technologically advanced systems developed to date. As a result, there are high expectations for them to excel across every modality, field, or application. While their strong performance across diverse tasks leaves open the possibility of a true generalist model in the future, their shortcomings suggest that fine-tuning or model collaboration may still offer more promising alternatives for downstream applications. We engage in this dialogue by benchmarking Claude as a multimodal proxy model on two tasks: image paragraph captioning and knowledge-based VQA. In addition, we develop a multi-agent framework to address the latter, drawing comparisons between single-agent and multi-agent approaches.

For our first task, we explored image paragraph captioning, which involves generating paragraph-length descriptions for images. We prompted a large-scale, multimodal LLM to generate paragraph descriptions for the images in a subset of the prominent Stanford Image Paragraph Captioning Dataset in a zero-shot setting. The linguistic metrics proposed by the creators of the dataset did not align with the detailed, rich, and flowing paragraphs that we empirically observed being generated by the model. Therefore, we conducted a linguistic analysis to compare the human-written paragraphs from the dataset annotations with the model's outputs. Our analysis indicates that the model-generated paragraphs are greater in length, lexical and semantic richness, diversity, and grammatical correctness compared to the human-annotated paragraphs. These findings suggest that modern multimodal language models have surpassed datasets and metrics that predate them, underscoring the need for new benchmarks in the area.

For our second task, we focused on knowledge-based visual question answering and specifically, the OK-VQA dataset, an open-ended benchmark which features questions about images that rely on external knowledge. We evaluated a current MLLM in a zero-shot setting and found that, despite its apparent knowledge of many of the tested facts, it falls short of reaching state-of-the-art performance levels. This is largely due to the format of the expected answers, which the model struggles to adhere to through prompting alone. To unlock the full potential of the MLLM, we propose a multi-agent system that establishes collaboration between the MLLM and a text-only LLM. Our framework comprises three agents: the **Scout**, an MLLM that takes an image as input and describes it in a paragraph caption; the **Analyser**, an LLM that produces an initial answer to the question based on the image description; and the **Resolver**, an LLM that selects and formats the final answer according to a set of predefined rules. The system improves performance over the baseline, indicating that a collaboration framework is effective for this task. Our findings indicate that multimodal LLMs can produce salient image descriptions that are sufficient to answer the question in the majority of cases.

Furthermore, LLMs can find answers to questions about images by leveraging the image paragraph description as context, provided that the visual information required to answer the question is present in the caption. Finally, we validate that LLMs can be effectively used as formatters to assist in extracting the final answer and post-processing it to align with the required format.

This thesis lays the groundwork for future exploration at the intersection of image paragraph captioning and knowledge-based visual question answering, as we believe the two tasks are not only closely related, but that advancing the former will significantly contribute to the latter. Moving forward, we plan to conduct ablation studies to better understand which component of our system is the most lacking and to assess which error types constitute the majority of our failures. We also aim to conduct experiments with other models to investigate how scalability affects performance. We hypothesise that smaller LLMs could be more efficient for this task, as they may exhibit less abstraction bias compared to larger models. To minimise errors occurring because of missing information in the description, we will explore the task of controlled image paragraph captioning, where the paragraph caption is grounded on the question or specific constraints. Finally, we plan to conduct a human evaluation to assess our system's performance on the OK-VQA task, based on the belief that the existing metric does not fully capture the models' true capabilities. We also intend to explore alternative automatic evaluation metrics, such as using LLMs as evaluators, which could better support open-ended tasks. Another important proposed direction involves the construction of new datasets that are on par with the growing capabilities of multimodal systems. Specifically, we emphasise the need for datasets in the image paragraph captioning domain that present meaningful challenges to state-of-the-art MLLMs by specifically targeting the key issues we have identified.

Chapter 7

Bibliography

[1] Agrawal, Aishwarya et al. *VQA: Visual Question Answering*. 2016. arXiv: 1505.00468 [cs.CL].

[2] Agrawal, Ayush et al. *Do Language Models Know When They're Hallucinating References?* 2024. arXiv: 2305.18248 [cs.CL].

[3] Ait, Adem, Izquierdo, Javier Luis Cánovas, and Cabot, Jordi. *Towards Modeling Human-Agentic Collaborative Workflows: A BPMN Extension*. 2024. arXiv: 2412.05958 [cs.SE].

[4] Alayrac, Jean-Baptiste et al. *Flamingo: a Visual Language Model for Few-Shot Learning*. 2022. arXiv: 2204.14198 [cs.CV].

[5] AlKhamissi, Badr et al. "A Review on Language Models as Knowledge Bases". In: *ArXiv* (Apr. 2022). eprint:

[6] AlphaCodeTeam. "AlphaCode 2 Technical Report". In: 2023. URL:

[7] Anisuzzaman, D.M. et al. "Fine-Tuning Large Language Models for Specialized Use Cases". In: *Mayo Clinic Proceedings: Digital Health* 3.1 (2025), p. 100184. ISSN: 2949-7612. DOI: https://doi.org/10.1016/j.mcpdig.2024.11.005.

[8] Argyrou, Georgia et al. *Automatic Generation of Fashion Images using Prompting in Generative Machine Learning Models*. 2024. arXiv: 2407.14944 [cs.CV].

[9] Azaria, Amos and Mitchell, Tom. *The Internal State of an LLM Knows When It's Lying*. 2023. arXiv: 2304.13734 [cs.CL].

[10] Barkur, Sudarshan Kamath, Schacht, Sigurd, and Scholl, Johannes. *Deception in LLMs: Self-Preservation and Autonomous Goals in Large Language Models*. 2025. arXiv: 2501.16513 [cs.CL].

[11] Bertolero, Maxwell A., Yeo, B. T. Thomas, and D'Esposito, Mark. "The modular and integrative functional architecture of the human brain". In: *Proceedings of the National Academy of Sciences* 112.49 (2015), E6798–E6807. DOI: 10.1073/pnas.1510619112. eprint:

[12] Bonilla-Salvador, Diego et al. *PixLore: A Dataset-driven Approach to Rich Image Captioning*. 2024. arXiv: 2312.05349 [cs.CV].

[13] Brown, Tom B. et al. *Language Models are Few-Shot Learners*. 2020. arXiv: 2005.14165 [cs.CL].

[14] Castelo, Noah, Bos, Maarten W., and Lehmann, Donald R. "Task-Dependent Algorithm Aversion". In: *Journal of Marketing Research* 56.5 (2019), pp. 809–825. DOI: 10.1177/0022243719851788. eprint:

[15] Chan, Alan et al. "Harms from Increasingly Agentic Algorithmic Systems". In: *2023 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '23. ACM, June 2023, pp. 651–666. DOI: 10.1145/3593013.3594033. URL:

[16] Chan, Chi-Min et al. *ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate*. 2023. arXiv: 2308.07201 [cs.CL].

[17] Chang, Trenton et al. "Measuring Steerability in Large Language Models". In: *Neurips Safe Generative AI Workshop 2024*. 2024.

[18] Chase, Harrison. *What is an agent?* Accessed: 2024-12-06. June 2024.

[19] Chatterjee, Moitreya and Schwing, Alexander G. *Diverse and Coherent Paragraph Generation from Images*. 2018. arXiv: 1809.00681 [cs.CV].

[20] Chen, Junzhe et al. *LLMArena: Assessing Capabilities of Large Language Models in Dynamic Multi-Agent Environments*. 2024. arXiv: 2402.16499 [cs.CL].

[21] Chen, Liang et al. "Beyond Factuality: A Comprehensive Evaluation of Large Language Models as Knowledge Generators". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 6325–6341. DOI: 10.18653/v1/2023.emnlp-main.390. URL:

[22] Chen, Liangyu et al. *Large Language Models are Visual Reasoning Coordinators*. 2023. arXiv: 2310.15166 [cs.CV].

[23] Chen, Liangyu et al. "Large Language Models are Visual Reasoning Coordinators". In: *ArXiv* (Oct. 2023). eprint:

[24] Chen, Pei, Han, Boran, and Zhang, Shuai. *CoMM: Collaborative Multi-Agent, Multi-Reasoning-Path Prompting for Complex Problem Solving*. 2024. arXiv: 2404.17729 [cs.CL].

[25] Chen, Weize et al. *AgentVerse: Facilitating Multi-Agent Collaboration and Exploring Emergent Behaviors*. 2023. arXiv: 2308.10848 [cs.CL].

[26] Chen, Xi et al. *PaLI-X: On Scaling up a Multilingual Vision and Language Model*. 2023. arXiv: 2305.18565 [cs.CV].

[27] Chen, Xi et al. *PaLI: A Jointly-Scaled Multilingual Language-Image Model*. 2023. arXiv: 2209.06794 [cs.CV].

[28] Das, Ayushman et al. "Enabling Synergistic Knowledge Sharing and Reasoning in Large Language Models with Collaborative Multi-Agents". In: *2023 IEEE 9th International Conference on Collaboration and Internet Computing (CIC)*. 2023, pp. 92–98. DOI: 10.1109/CIC58953.2023.00021.

[29] Dietvorst, Berkeley J., Simmons, Joseph P., and Massey, Cade. "Algorithm aversion: People erroneously avoid algorithms after seeing them err." In: *Journal of Experimental Psychology: General* 144.1 (2015), pp. 114–126. ISSN: 0096-3445. DOI: 10.1037/xge0000033.

[30] Driess, Danny et al. *PaLM-E: An Embodied Multimodal Language Model*. 2023. arXiv: 2303.03378 [cs.LG].

[31] Du, Yilun et al. *Improving Factuality and Reasoning in Language Models through Multiagent Debate*. 2023. arXiv: 2305.14325 [cs.CL].

[32] Duan, Nan, Tang, Duyu, and Zhou, Ming. "Machine Reasoning: Technology, Dilemma and Future". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*. Ed. by Aline Villavicencio and Benjamin Van Durme. Online: Association for Computational Linguistics, Nov. 2020, pp. 1–6. DOI: 10.18653/v1/2020.emnlp-tutorials.1.

[33] Farmer, J. Doyne and Axtell, Robert L. *Agent-based modeling in Economics and Finance: Past, Present, and Future*. INET Oxford Working Papers 2022-10. Institute for New Economic Thinking at the Oxford Martin School, University of Oxford, June 2022.

[34] Ferber, Jacques. *Multi-Agent System: An Introduction to Distributed Artificial Intelligence*. Harlow: Addison Wesley Longman, 1999. ISBN: 0-201-36048-9.

[35] Ferrara, Emilio. "Should ChatGPT be biased? Challenges and risks of bias in large language models". In: *First Monday* (Nov. 2023). ISSN: 1396-0466. DOI: 10.5210/fm.v28i11.13346. URL:

[36] Gabriel, Adrian Garret, Ahmad, Alaa Alameer, and Jeyakumar, Shankar Kumar. *Advancing Agentic Systems: Dynamic Task Decomposition, Tool Integration and Evaluation using Novel Metrics and Dataset*. 2024. arXiv: 2410.22457 [cs.AI].

[37] Gabriel, Iason et al. *The Ethics of Advanced AI Assistants*. 2024. arXiv: 2404.16244 [cs.CY].

[38] Gallegos, Isabel O. et al. *Bias and Fairness in Large Language Models: A Survey*. 2024. arXiv: 2309.00770 [cs.CL].

[39] Gao, Yunfan et al. *Retrieval-Augmented Generation for Large Language Models: A Survey*. 2024. arXiv: 2312.10997 [cs.CL].

[40] Garg, Roopal et al. *ImageInWords: Unlocking Hyper-Detailed Image Descriptions*. 2024. arXiv: 2405.02793 [cs.CV].

[41] Gauthier, Jon et al. "SyntaxGym: An Online Platform for Targeted Evaluation of Language Models". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Ed. by Asli Celikyilmaz and Tsung-Hsien Wen. Online: Association for Computational Linguistics, July 2020, pp. 70–76. DOI: 10.18653/v1/2020.acl-demos.10.

[42] Ghaffarzadegan, Navid et al. "Generative agent-based modeling: an introduction and tutorial". In: *System Dynamics Review* 40.1 (Jan. 2024). ISSN: 1099-1727. DOI: 10.1002/sdr.1761.

[43] Giadikiaroglou, Panagiotis et al. *Puzzle Solving using Reasoning of Large Language Models: A Survey*. 2024. arXiv: 2402.11291 [cs.CL].

[44] Grünebaum, Amos et al. "The exciting potential for ChatGPT in obstetrics and gynecology." eng. In: *American Journal of Obstetrics and Gynecology* 228.6 (June 2023), pp. 696–705. ISSN: 1097-6868 (Electronic); 0002-9378 (Linking). DOI: 10.1016/j.ajog.2023.03.009.

[45] Guidotti, Riccardo et al. "A Survey of Methods for Explaining Black Box Models". In: *ACM Comput. Surv.* 51.5 (Aug. 2018). ISSN: 0360-0300. DOI: 10.1145/3236009.

[46] Gunjal, Anisha, Yin, Jihan, and Bas, Erhan. *Detecting and Preventing Hallucinations in Large Vision Language Models*. 2024. arXiv: 2308.06394 [cs.CV].

[47] Guo, Taicheng et al. *Large Language Model based Multi-Agents: A Survey of Progress and Challenges*. 2024. arXiv: 2402.01680 [cs.CL].

[48] Guo, Zhijiang, Schlichtkrull, Michael, and Vlachos, Andreas. "A Survey on Automated Fact-Checking". In: *Transactions of the Association for Computational Linguistics* 10 (2022). Ed. by Brian Roark and Ani Nenkova, pp. 178–206. DOI: 10.1162/tacl_a_00454.

[49] He, Zhitao et al. "LEGO: A Multi-agent Collaborative Framework with Role-playing and Iterative Feedback for Causality Explanation Generation". In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 9142–9163. DOI: 10.18653/v1/2023.findings-emnlp.613.

[50] Hu, Jennifer et al. *A fine-grained comparison of pragmatic language understanding in humans and language models*. 2023. arXiv: 2212.06801 [cs.CL].

[51] Hu, Zhongjian et al. *Multi-Agents Based on Large Language Models for Knowledge-based Visual Question Answering*. 2024. arXiv: 2412.18351 [cs.CL].

[52] Huang, Minlie, Zhu, Xiaoyan, and Gao, Jianfeng. "Challenges in Building Intelligent Open-domain Dialog Systems". In: *ACM Transactions on Information Systems* 38.3 (Apr. 2020). ISSN: 1046-8188. DOI: 10.1145/3383123.

[53] Islam, Md. Ashraful, Ali, Mohammed Eunus, and Parvez, Md Rizwan. *MapCoder: Multi-Agent Code Generation for Competitive Problem Solving*. 2024. arXiv: 2405.11403 [cs.CL].

[54] Jain, Aman et al. "Select, Substitute, Search: A New Benchmark for Knowledge-Augmented Visual Question Answering". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '21. ACM, July 2021, pp. 2491–2498.

[55] Janis, Irving L. *Victims of Groupthink: A Psychological Study of Foreign-Policy Decisions and Fiascoes*. Boston: Houghton Mifflin, 1972.

[56] Jeong, Cheonsu. "Domain-specialized LLM: Financial fine-tuning and utilization method using Mistral 7B". In: *Journal of Intelligence and Information Systems* 30.1 (Mar. 2024), pp. 93–120. ISSN: 2288-4882. DOI: 10.13088/jiis.2024.30.1.093. URL:

[57] Ji, Ziwei et al. "Survey of Hallucination in Natural Language Generation". In: *ACM Computing Surveys* 55.12 (Mar. 2023), pp. 1–38. ISSN: 1557-7341. DOI: 10.1145/3571730.

[58] Jia, Chao et al. *Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision*. 2021. arXiv: 2102.05918 [cs.CV].

[59] Jiang, Bowen et al. *Multi-Agent VQA: Exploring Multi-Agent Foundation Models in Zero-Shot Visual Question Answering*. 2024. arXiv: 2403.14783 [cs.CV].

[60] Jiao, Junfeng et al. *Navigating LLM Ethics: Advancements, Challenges, and Future Directions*. 2025. arXiv: 2406.18841 [cs.CY].

[61] Jussupow, Ekaterina, Benbasat, Izak, and Heinzl, Armin. "WHY ARE WE AVERSE TOWARDS ALGORITHMS? A COMPREHENSIVE LITERATURE REVIEW ON ALGORITHM AVERSION". In: *ECIS 2020 Research Papers* (2020).

[62] Kapoor, Sayash et al. *AI Agents That Matter*. 2024. arXiv: 2407.01502 [cs.LG].

[63] Khattab, Omar et al. "DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines". In: *ArXiv* (Oct. 2023). eprint:

[64] Kim, Minchan et al. *ESREAL: Exploiting Semantic Reconstruction to Mitigate Hallucinations in Vision-Language Models*. 2024. arXiv: 2403.16167 [cs.CV].

[65] Krause, Jonathan et al. *A Hierarchical Approach for Generating Descriptive Image Paragraphs*. 2017. arXiv: 1611.06607 [cs.CV].

[66] Kritharoula, Anastasia, Lymperaiou, Maria, and Stamou, Giorgos. *Language Models as Knowledge Bases for Visual Word Sense Disambiguation*. 2023. arXiv: 2310.01960 [cs.CL].

[67] Kritharoula, Anastasia, Lymperaiou, Maria, and Stamou, Giorgos. "Large Language Models and Multimodal Retrieval for Visual Word Sense Disambiguation". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2023, pp. 13053–13077. DOI: 10.18653/v1/2023.emnlp-main.807.

[68] "KVQA: Knowledge-Aware Visual Question Answering". In: 33 (July 2019), pp. 8876–8884. DOI: 10.1609/aaai.v33i01.33018876.

[69] "Language and thought are not the same thing: evidence from neuroimaging and neurological patients". In: *Annals of the New York Academy of Sciences* 1369 (1), pp. 132–153.

[70] Lewis, Patrick et al. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. 2021. arXiv: 2005.11401 [cs.CL].

[71] Li, Guohao et al. *CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society*. 2023. arXiv: 2303.17760 [cs.AI].

[72] Li, Huao et al. "Theory of Mind for Multi-Agent Collaboration via Large Language Models". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2023. DOI: 10.18653/v1/2023.emnlp-main.13.

[73] Li, Junnan et al. *BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models*. 2023. arXiv: 2301.12597 [cs.CV].

[74] Li, Junyi et al. *HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models*. 2023. arXiv: 2305.11747 [cs.CL].

[75] Li, Nian et al. *EconAgent: Large Language Model-Empowered Agents for Simulating Macroeconomic Activities*. 2024. arXiv: 2310.10436 [cs.AI].

[76] Li, Xiang Lisa and Liang, Percy. "Prefix-Tuning: Optimizing Continuous Prompts for Generation". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by Chengqing Zong et al. Online: Association for Computational Linguistics, Aug. 2021, pp. 4582–4597. DOI: 10.18653/v1/2021.acl-long.353.

[77] Li, Xinyi et al. "A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges". In: *Vicinagearth* 1.1 (2024), p. 9. DOI: 10.1007/s44336-024-00009-2.

[78] Li, Yang et al. *TradingGPT: Multi-Agent System with Layered Memory and Distinct Characters for Enhanced Financial Trading Performance*. 2023. arXiv: 2309.03736 [q-fin.PM].

[79] Li, Yifan et al. "Evaluating Object Hallucination in Large Vision-Language Models". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 292–305. DOI: 10.18653/v1/2023.emnlp-main.20. URL:

[80] Li, Zongxia et al. *A Survey of State of the Art Large Vision Language Models: Alignment, Benchmark, Evaluations and Challenges*. 2025. arXiv: 2501.02189 [cs.CV].

[81] Liang, Tian et al. *Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate*. 2024. arXiv: 2305.19118 [cs.CL].

[82] Lin, Matthieu et al. *LLM-based Optimization of Compound AI Systems: A Survey*. 2024. arXiv: 2410.16392 [cs.CL].

[83] Ling, Chen et al. *Domain Specialization as the Key to Make Large Language Models Disruptive: A Comprehensive Survey*. 2024. arXiv: 2305.18703 [cs.CL].

[84] Liu, Pengfei et al. *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*. 2021. arXiv: 2107.13586 [cs.CL].

[85] Liu, Ruibo et al. *Training Socially Aligned Language Models on Simulated Social Interactions*. 2023. arXiv: 2305.16960 [cs.CL].

[86] Liu, Ruibo et al. *Training Socially Aligned Language Models on Simulated Social Interactions*. 2023. arXiv: 2305.16960 [cs.CL].

[87] Long, Jieyi. *Large Language Model Guided Tree-of-Thought*. 2023. arXiv: 2305.08291 [cs.AI].

[88] Lou, Siyu et al. *Quantifying In-Context Reasoning Effects and Memorization Effects in LLMs*. 2024. arXiv: 2405.11880 [cs.LG].

[89] Lu, Jiaying et al. *Evaluation and Enhancement of Semantic Grounding in Large Vision-Language Models*. 2024. arXiv: 2309.04041 [cs.CV].

[90] Lu, Pan et al. "Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering". In: *The 36th Conference on Neural Information Processing Systems (NeurIPS)*. 2022.
[91] Lymperaiou, Maria and Stamou, Giorgos. *A survey on knowledge-enhanced multimodal learning*. 2024. arXiv: 2211.12328 [cs.LG].
[92] Lymperaiou, Maria and Stamou, Giorgos. *The Contribution of Knowledge in Visiolinguistic Learning: A Survey on Tasks and Challenges*. 2023. arXiv: 2303.02411 [cs.CL].
[93] Lymperaiou, Maria et al. *HalCECE: A Framework for Explainable Hallucination Detection through Conceptual Counterfactuals in Image Captioning*. 2025. arXiv: 2503.00436 [cs.CV].
[94] Ma, Yukun et al. "A survey on empathetic Dialogue Systems". In: *Information Fusion* 64 (2020), pp. 50–70. ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2020.06.011.
[95] Mahowald, Kyle et al. "Dissociating language and thought in large language models". In: *Trends in Cognitive Sciences* 28.6 (2024), pp. 517–540. ISSN: 1364-6613. DOI: https://doi.org/10.1016/j.tics.2024.01.011. URL:
[96] Marino, Kenneth et al. *OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge*. 2019. arXiv: 1906.00067 [cs.CV].
[97] Melas-Kyriazi, Luke, Rush, Alexander, and Han, George. "Training for Diversity in Image Paragraph Captioning". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Ed. by Ellen Riloff et al. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 757–761. DOI: 10.18653/v1/D18-1084.
[98] Mialon, Grégoire et al. *Augmented Language Models: a Survey*. 2023. arXiv: 2302.07842 [cs.CL].
[99] Miehling, Erik et al. *Evaluating the Prompt Steerability of Large Language Models*. 2025. arXiv: 2411.12405 [cs.CL].
[100] Mündler, Niels et al. *Self-contradictory Hallucinations of Large Language Models: Evaluation, Detection and Mitigation*. 2024. arXiv: 2305.15852 [cs.CL].
[101] Nagaonkar, Sankalp et al. *Benchmarking Vision-Language Models on Optical Character Recognition in Dynamic Video Environments*. 2025. arXiv: 2502.06445 [cs.CV].
[102] Ng, Andrew. *Welcoming diverse approaches keeps machine learning strong*. Accessed: 2024-12-04. June 2024. URL:
[103] Nickerson, Raymond S., Butler, Susan F., and Carlin, Michael. "43Empathy and Knowledge Projection". In: *The Social Neuroscience of Empathy*. The MIT Press, Mar. 2009. ISBN: 9780262012973. DOI: 10.7551/mitpress/9780262012973.003.0005. eprint:
[104] Nori, Harsha et al. *Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine*. 2023. arXiv: 2311.16452 [cs.CL].
[105] Onoe, Yasumasa et al. *DOCCI: Descriptions of Connected and Contrasting Images*. 2024. arXiv: 2404.19753 [cs.CV].
[106] OpenAI et al. *GPT-4 Technical Report*. 2024. arXiv: 2303.08774 [cs.CL].
[107] Ou, Jiao et al. "DialogBench: Evaluating LLMs as Human-like Dialogue Systems". In: *ArXiv* (Nov. 2023). eprint:
[108] Ou, Jiao et al. *DialogBench: Evaluating LLMs as Human-like Dialogue Systems*. 2024. arXiv: 2311.01677 [cs.CL].
[109] Pandiani, Delfina Sol Martinez and Presutti, Valentina. *Seeing the Intangible: Survey of Image Classification into High-Level and Abstract Categories*. 2024. arXiv: 2308.10562 [cs.CV].
[110] Papadimitriou, Christos et al. *Masked Generative Story Transformer with Character Guidance and Caption Augmentation*. 2024. arXiv: 2403.08502 [cs.CV].
[111] Park, Joon Sung et al. *Generative Agents: Interactive Simulacra of Human Behavior*. 2023. arXiv: 2304.03442 [cs.HC].
[112] Petroni, Fabio et al. *Language Models as Knowledge Bases?* 2019. arXiv: 1909.01066 [cs.CL].
[113] Prasad, Archiki et al. *ADaPT: As-Needed Decomposition and Planning with Language Models*. 2024. arXiv: 2311.05772 [cs.AI].
[114] Prompting Guide. *Prompt Chaining*. Accessed: 2025-04-09. n.d. URL:
[115] Puerto, Haritz, Şahin, Gözde Gül, and Gurevych, Iryna. *MetaQA: Combining Expert Agents for Multi-Skill Question Answering*. 2023. arXiv: 2112.01922 [cs.CL].
[116] Qiao, Shuofei et al. *AutoAct: Automatic Agent Learning from Scratch for QA via Self-Planning*. 2024. arXiv: 2401.05268 [cs.CL].

[117] Qiao, Shuofei et al. "Reasoning with Language Model Prompting: A Survey". In: *ArXiv* (Dec. 2022). eprint:

[118] Radford, Alec et al. "Language Models are Unsupervised Multitask Learners". In: 2019.

[119] Radford, Alec et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: 2103.00020 [cs.CV].

[120] Rawte, Vipula, Sheth, Amit, and Das, Amitava. "A Survey of Hallucination in Large Foundation Models". In: *ArXiv* (Sept. 2023). eprint:

[121] Sahoo, Pranab et al. *A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications*. 2025. arXiv: 2402.07927 [cs.AI].

[122] Salecha, Aades et al. "Large language models display human-like social desirability biases in Big Five personality surveys". In: *PNAS Nexus* (12 2024). DOI: 10.1093/pnasnexus/pgae533.

[123] Sap, Maarten et al. *Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs*. 2023. arXiv: 2210.13312 [cs.CL].

[124] Schuster, Sebastian and Linzen, Tal. "When a sentence does not introduce a discourse entity, Transformer-based models still sometimes refer to it". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 969–982. DOI: 10.18653/v1/2022.naacl-main.71.

[125] Schwenk, Dustin et al. *A-OKVQA: A Benchmark for Visual Question Answering using World Knowledge*. 2022. arXiv: 2206.01718 [cs.CV].

[126] Shavit, Yonadav et al. "Practices for governing agentic AI systems". In: *Research Paper, OpenAI, December* (2023).

[127] Shen, Yongliang et al. *HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face*. 2023. arXiv: 2303.17580 [cs.CL].

[128] Shinn, Noah et al. *Reflexion: Language Agents with Verbal Reinforcement Learning*. 2023. arXiv: 2303.11366 [cs.AI].

[129] Singh, Aditi et al. *A Survey of Sustainability in Large Language Models: Applications, Economics, and Challenges*. 2025. arXiv: 2412.04782 [cs.AI].

[130] Singhal, Karan et al. *Towards Expert-Level Medical Question Answering with Large Language Models*. 2023. arXiv: 2305.09617 [cs.CL].

[131] Sirigu, Angela, Duhamel, Jean-Rene, and Poncet, Michel. "The role of sensorimotor experience in object recognition: A case of multimodal agnosia". In: *Brain* (6 1991), pp. 2555–2573. DOI: 10.1093/brain/114.6.2555.

[132] Soman, Karthik et al. *Biomedical knowledge graph-optimized prompt generation for large language models*. 2024. arXiv: 2311.17330 [cs.CL].

[133] Sorensen, Taylor et al. *A Roadmap to Pluralistic Alignment*. 2024. arXiv: 2402.05070 [cs.AI].

[134] Soudani, Heydar, Kanoulas, Evangelos, and Hasibi, Faegheh. "Fine Tuning vs. Retrieval Augmented Generation for Less Popular Knowledge". In: *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*. SIGIR-AP 2024. ACM, Dec. 2024, pp. 12–22. DOI: 10.1145/3673791.3698415. URL:

[135] Stoikou, Theodoti, Lymperaiou, Maria, and Stamou, Giorgos. *Knowledge-Based Counterfactual Queries for Visual Question Answering*. 2023. arXiv: 2303.02601 [cs.CL].

[136] Su, Yu et al. "Language Agents: Foundations, Prospects, and Risks". In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*. Ed. by Jessy Li and Fei Liu. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 17–24. DOI: 10.18653/v1/2024.emnlp-tutorials.3.

[137] Sumers, Theodore R. et al. *Cognitive Architectures for Language Agents*. 2024. arXiv: 2309.02427 [cs.AI].

[138] Sun, Kai et al. *Head-to-Tail: How Knowledgeable are Large Language Models (LLMs)? A.K.A. Will LLMs Replace Knowledge Graphs?* 2024. arXiv: 2308.10168 [cs.CL].

[139] Talebirad, Yashar and Nadiri, Amirhossein. *Multi-Agent Collaboration: Harnessing the Power of Intelligent LLM Agents*. 2023. arXiv: 2306.03314 [cs.AI].

[140] Talmor, Alon et al. *CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge*. 2019. arXiv: 1811.00937 [cs.CL].

[141] Tang, Xiangru et al. *MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning*. 2024. arXiv: 2311.10537 [cs.CL].
[142] Team, Gemini et al. *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context*. 2024. arXiv: 2403.05530 [cs.CL].
[143] "The Claude 3 Model Family: Opus, Sonnet, Haiku". In: URL:
[144] Trinh, Trieu H. et al. "Solving olympiad geometry without human demonstrations". In: *Nature* 625.7995 (2024), pp. 476–482. DOI: 10.1038/s41586-023-06747-5.
[145] Urbanek, Jack et al. *A Picture is Worth More Than 77 Text Tokens: Evaluating CLIP-Style Models on Dense Captions*. 2024. arXiv: 2312.08578 [cs.CV].
[146] Vafa, Keyon et al. *What's Producible May Not Be Reachable: Measuring the Steerability of Generative Models*. 2025. arXiv: 2503.17482 [cs.LG].
[147] Vaswani, Ashish et al. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL].
[148] Vulchanova, Mila et al. "Language and perception: Introduction to the Special Issue "Speakers and Listeners in the Visual World"". In: *Journal of Cultural Cognitive Science* 3.2 (2019), pp. 103–112. DOI: 10.1007/s41809-019-00047-z.
[149] Wang, Cunxiang et al. *Survey on Factuality in Large Language Models: Knowledge, Retrieval and Domain-Specificity*. 2023. arXiv: 2310.07521 [cs.CL].
[150] Wang, Jing et al. *Convolutional Auto-encoding of Sentence Topics for Image Paragraph Generation*. 2019. arXiv: 1908.00249 [cs.CV].
[151] Wang, Lei et al. "A survey on Large Language Model based autonomous agents". In: *Frontiers of Computer Science* 18.6 (Mar. 2024). ISSN: 2095-2236. DOI: 10.1007/s11704-024-40231-1.
[152] Wang, Peng et al. *Explicit Knowledge-based Reasoning for Visual Question Answering*. 2015. arXiv: 1511.02570 [cs.CV].
[153] Wang, Peng et al. *FVQA: Fact-based Visual Question Answering*. 2017. arXiv: 1606.05433 [cs.CV].
[154] Wang, Xinyi et al. *Generalization v.s. Memorization: Tracing Language Models' Capabilities Back to Pretraining Data*. 2025. arXiv: 2407.14985 [cs.CL].
[155] Wang, Xuezhi et al. *Self-Consistency Improves Chain of Thought Reasoning in Language Models*. 2023. arXiv: 2203.11171 [cs.CL].
[156] Wang, Zeqing et al. *Towards Top-Down Reasoning: An Explainable Multi-Agent Approach for Visual Question Answering*. 2024. arXiv: 2311.17331 [cs.CV].
[157] Wang, Zeqing et al. *Towards Top-Down Reasoning: An Explainable Multi-Agent Approach for Visual Question Answering*. 2025. arXiv: 2311.17331 [cs.CV].
[158] Warstadt, Alex et al. "BLiMP: The Benchmark of Linguistic Minimal Pairs for English". In: *Transactions of the Association for Computational Linguistics* 8 (2020). Ed. by Mark Johnson, Brian Roark, and Ani Nenkova, pp. 377–392. DOI: 10.1162/tacl_a_00321.
[159] Wasserkrug, Segev et al. "From Large Language Models and Optimization to Decision Optimization CoPilot: A Research Manifesto". In: *ArXiv* (Feb. 2024). eprint:
[160] Wei, Jason et al. *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. 2023. arXiv: 2201.11903 [cs.CL].
[161] Wei, Jason et al. *Emergent Abilities of Large Language Models*. 2022. arXiv: 2206.07682 [cs.CL].
[162] Wei, Jason et al. "Finetuned Language Models Are Zero-Shot Learners". In: (Sept. 2021). eprint:
[163] Weiss, Martin et al. *Rethinking the Buyer's Inspection Paradox in Information Markets with Language Agents*. 2024.
[164] Weng, Lilian. *LLM Powered Autonomous Agents*. Section: posts. Accessed on 2024-12-11. June 2023.
[165] Williams, Ross et al. *Epidemic Modeling with Generative Agents*. 2023. arXiv: 2307.04986 [cs.AI].
[166] Wu, Zhaofeng et al. "Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks". In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Ed. by Kevin Duh, Helena Gomez, and Steven Bethard. Mexico City, Mexico: Association for Computational Linguistics, June 2024, pp. 1819–1862. DOI: 10.18653/v1/2024.naacl-long.102.
[167] Wysocka, Magdalena et al. "Large Language Models, scientific knowledge and factuality: A framework to streamline human expert evaluation." eng. In: *J Biomed Inform* 158 (Oct. 2024), p. 104724. ISSN: 1532-0480 (Electronic); 1532-0464 (Linking). DOI: 10.1016/j.jbi.2024.104724.

[168] Xi, Zhiheng et al. *The Rise and Potential of Large Language Model Based Agents: A Survey*. 2023. arXiv: 2309.07864 [cs.AI].

[169] Xiao, Bushi, Yin, Ziyuan, and Shan, Zixuan. *Simulating Public Administration Crisis: A Novel Generative Agent-Based Simulation System to Lower Technology Barriers in Social Science Research*. 2023. arXiv: 2311.06957 [cs.CY].

[170] Xie, Chulin et al. *On Memorization of Large Language Models in Logical Reasoning*. 2025. arXiv: 2410.23123 [cs.CL].

[171] Xiong, Haoyi et al. *When Search Engine Services meet Large Language Models: Visions and Challenges*. 2024. arXiv: 2407.00128 [cs.IR].

[172] Xiong, Kai et al. "Examining Inter-Consistency of Large Language Models Collaboration: An In-depth Analysis via Debate". In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, 2023, pp. 7572–7590. DOI: 10.18653/v1/2023.findings-emnlp.508.

[173] Xu, Chunpu et al. "Interactive Key-Value Memory-augmented Attention for Image Paragraph Captioning". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Ed. by Donia Scott, Nuria Bel, and Chengqing Zong. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 3132–3142. DOI: 10.18653/v1/2020.coling-main.279.

[174] Xu, Zhenran et al. *Towards Reasoning in Large Language Models via Multi-Agent Peer Review Collaboration*. 2023. arXiv: 2311.08152 [cs.CL].

[175] Yang, Pu and Dong, Bin. *MoColl: Agent-Based Specific and General Model Collaboration for Image Captioning*. 2025. arXiv: 2501.01834 [cs.CV].

[176] Yang, Xu et al. "Hierarchical Scene Graph Encoder-Decoder for Image Paragraph Captioning". In: *Proceedings of the 28th ACM International Conference on Multimedia*. MM '20. Seattle, WA, USA: Association for Computing Machinery, 2020, pp. 4181–4189. ISBN: 9781450379885. DOI: 10.1145/3394171.3413859. URL:

[177] Yao, Shunyu et al. *ReAct: Synergizing Reasoning and Acting in Language Models*. 2023. arXiv: 2210.03629 [cs.CL].

[178] Yao, Shunyu et al. *Tree of Thoughts: Deliberate Problem Solving with Large Language Models*. 2023. arXiv: 2305.10601 [cs.CL].

[179] Yu, Zhou et al. *Prophet: Prompting Large Language Models with Complementary Answer Heuristics for Knowledge-based Visual Question Answering*. 2023. arXiv: 2303.01903 [cs.CV].

[180] Yuan, Quan et al. *TaskLAMA: Probing the Complex Task Understanding of Language Models*. 2023. arXiv: 2308.15299 [cs.CL].

[181] Yue, Xiang et al. *MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI*. 2024. arXiv: 2311.16502 [cs.CL].

[182] Zaharia, Matei et al. *The Shift from Models to Compound AI Systems*. 2024.

[183] Zellers, Rowan et al. *From Recognition to Cognition: Visual Commonsense Reasoning*. 2019. arXiv: 1811.10830 [cs.CV].

[184] Zhang, Honghua et al. *On the Paradox of Learning to Reason from Data*. 2022. arXiv: 2205.11502 [cs.CL].

[185] Zhang, Hongxin et al. *Building Cooperative Embodied Agents Modularly with Large Language Models*. 2024. arXiv: 2307.02485 [cs.AI].

[186] Zhang, Jintian et al. *Exploring Collaboration Mechanisms for LLM Agents: A Social Psychology View*. 2024. arXiv: 2310.02124 [cs.CL].

[187] Zhang, Yue et al. *Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models*. 2023. arXiv: 2309.01219 [cs.CL].

[188] Zhao, Haiyan et al. "Explainability for Large Language Models: A Survey". In: *ACM Trans. Intell. Syst. Technol.* 15.2 (Feb. 2024). ISSN: 2157-6904. DOI: 10.1145/3639372. URL:

[189] Zhao, Qinlin et al. *CompeteAI: Understanding the Competition Dynamics in Large Language Model-based Agents*. 2024. arXiv: 2310.17512 [cs.AI].

[190] Zheng, Zhiling et al. "ChatGPT Research Group for Optimizing the Crystallinity of MOFs and COFs". In: *ACS Central Science* 9.11 (2023), pp. 2161–2170. DOI: 10.1021/acscentsci.3c01087. eprint:

[191] Zhou, Yiyang et al. "Analyzing and Mitigating Object Hallucination in Large Vision-Language Models". In: *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*. 2023.