Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Σηματων, Ελεγχου και Ρομποτικης

Multimodal Approaches to Automatic Lyric Generation

Διπλωματικη Εργασια

της

ΟΛΓΑΣ Α. ΜΠΑΡΛΟΥ

Επιβλέπων: Αλέξανδρος Ποταμιάνος
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2024

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Σηματων, Ελεγχου και Ρομποτικης

Multimodal Approaches to Automatic Lyric Generation

Διπλωματικη Εργασια

της

ΟΛΓΑΣ Α. ΜΠΑΡΛΟΥ

Επιβλέπων: Αλέξανδρος Ποταμιάνος
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 18η Οκτωβρίου 2024.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
Αλέξανδρος Ποταμιάνος	Αθανάσιος Ροντογιάννης	Αθανάσιος Βουλόδημος
Αναπληρωτής Καθηγητής Ε.Μ.Π.	Αναπληρωτής Καθηγητής Ε.Μ.Π.	Επίκουρος Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2024

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Σηματων, Ελεγχου και Ρομποτικης

.....
Όλγα Μπάρλου

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

Περίληψη

Η μουσική διαδραματίζει θεμελιώδη ρόλο στον ανθρώπινο πολιτισμό, λειτουργώντας ως παγκόσμια γλώσσα που ξεπερνά τα εμπόδια και έχει βαθιά απήχηση στα συναισθήματα και τις εμπειρίες των ανθρώπων. Καθώς η τεχνητή νοημοσύνη συνεχίζει να εξελίσσεται, υπάρχει αυξανόμενο ενδιαφέρον για την εφαρμογή αυτών των τεχνολογιών σε δημιουργικούς τομείς, συμπεριλαμβανομένης της συγγραφής στίχων. Ακόμα και με αυτές τις εξελίξεις, τα σύγχρονα πολυτροπικά μοντέλα ήχου δεν έχουν εκπαιδευτεί επαρκώς σε εργασίες ανάκτησης μουσικής πληροφορίας, και ειδικά σε δημιουργικές εργασίες όπως η παραγωγή στίχων. Επιπλέον, μόνο ορισμένα Μεγάλα Γλωσσικά Μοντέλα έχει αποδειχθεί ότι εμφανίζουν την ικανότητα για λεπτομερή και συναισθηματικά φορτισμένη γραφή, γεγονός που υπογραμμίζει την ανάγκη για πιο εξελιγμένες προσεγγίσεις σε αυτόν τον τομέα.

Στην παρούσα διπλωματική εργασία, διεξάγουμε μια ολοκληρωμένη αξιολόγηση τεσσάρων διαφορετικών προσεγγίσεων για την παραγωγή στίχων, ενσωματώνοντας σταδιακά διαφορετικές τροπικότητες. Ξεκινάμε με την παραδοσιακή παραγωγή στίχων από κείμενο σε κείμενο με τη χρήση σύγχρονων Μεγάλων Γλωσσικών Μοντέλων. Στη συνέχεια, διερευνούμε την παραγωγή κειμένου ενισχυμένου με ήχο μέσω δύο προσεγγίσεων: ενός μοντέλου Variational Autoencoder με αρχιτεκτονική που μοιάζει με Transformer και ενός μοντέλου που ευθυγραμμίζει της αναπαραστάσεις μουσικής και κειμένου μεταξύ των μοντέλων Whisper και OpenOrca. Στη συνέχεια, υλοποιούμε μια διαδικασία δύο σταδίων με τη χρήση του SALMONN για την εξαγωγή μουσικών ετικετών και στη συνέχεια του Claude για την παραγωγή στίχων. Τέλος, προτείνουμε μια νέα πολυτροπική διάταξη που συνδυάζει το SALMONN για την περιγραφή της σκηνής της ταινίας, το Stable Diffusion για την οπτικοποίηση και το LLaVA για την τελική παραγωγή στίχων.

Η αξιολόγησή μας, που βασίζεται τόσο σε μετρικές που βασίζονται σε LLM όσο και σε ανθρώπινη αξιολόγηση, αποκαλύπτει διάφορα βασικά ευρήματα. Πρώτον, τα instruction-tuned LLM επιδεικνύουν ισχυρές baseline επιδόσεις ακόμη και χωρίς περαιτέρω εκπαίδευση στον συγκεκριμένο τομέα. Δεύτερον, η προσθήκη της τροπικότητας ήχου μέσω της εξαγωγής μουσικών ετικετών ενισχύει σημαντικά τη συσχέτιση μεταξύ των παραγόμενων στίχων και της μουσικής. Τρίτον, η νέα μας προσέγγιση που ενσωματώνει οπτικές αναπαραστάσεις επιτυγχάνει την καλύτερη ισορροπία μεταξύ συνοχής των στίχων και της συσχέτισής τους με την μουσική. Είναι ενδιαφέρον ότι, ενώ το few-shot prompting βελτίωσε τα σκορ ομοιότητας, παρουσίασε μειωμένη απόδοση στις αξιολογήσεις της ποιότητας των στίχων. Τα ευρήματα αυτά υποδηλώνουν ότι οι πολυτροπικές προσεγγίσεις μπορούν να βελτιώσουν την παραγωγή στίχων διατηρώντας παράλληλα τη δημιουργική έκφραση, ωστόσο υπάρχει μια λεπτή ισορροπία μεταξύ της καθοδηγούμενης παραγωγής και της δημιουργικής ελευθερίας.

Αυτή η έρευνα συνεισφέρει νέες μεθοδολογίες σε εργασίες ανάκτησης μουσικής πληροφορίας και ανοίγει δρόμους για μελλοντική εξερεύνηση πολυτροπικών προσεγγίσεων σε δημιουργικές εφαρμογές τεχνητής νοημοσύνης.

Λέξεις Κλειδιά

Πολυτροπικά Μοντέλα, Μετασχηματιστής, Κωδικοποιητής, Αποκωδικοποιητής, Μεγάλα Γλωσσικά Μοντέλα, Επεξεργασία Φυσικής Γλώσσας, Ανάκτηση Μουσικής Πληροφορίας

Abstract

Music plays a fundamental role in human culture, serving as a universal language that transcends barriers and resonates deeply with people's emotions and experiences. As artificial intelligence continues to advance, there is growing interest in applying these technologies to creative domains, including lyric generation. While Large Language Models (LLMs) have shown promise in creative writing tasks, the potential of multimodal approaches in lyric generation remains largely unexplored.

In this diploma thesis, we conduct a comprehensive evaluation of four distinct approaches to lyric generation, progressively incorporating different modalities. We begin with traditional text-to-text lyric generation using state-of-the-art LLMs. We then explore audio-enhanced text generation through two approaches: a Variational Autoencoder model with Transformer-like architecture, and a model with a projection layer to align music and text representations between the Whisper and OpenOrca models. Following this, we implement a two-stage process using SALMONN for music tag extraction followed by Claude for lyric generation. Finally, we propose a novel multimodal pipeline combining SALMONN for movie scene description, Stable Diffusion for visualization, and LLaVA for final lyric generation.

Our evaluation, based on both LLM-based metrics and human assessment, reveals several key findings. First, instruction-tuned LLMs demonstrate strong baseline performance even without domain-specific training. Second, the addition of audio modality through music tag extraction significantly enhances the correlation between generated lyrics and music. Third, our novel approach incorporating visual representations achieves the best balance between lyrical coherence and musical correlation. Interestingly, while few-shot prompting improved similarity metrics, it showed decreased performance in creative quality assessments. These findings suggest that thoughtfully integrated multimodal approaches can enhance lyric generation while maintaining creative expression, though there exists a delicate balance between guided generation and creative freedom.

This research contributes new methodologies to music information retrieval tasks and opens avenues for future exploration in multimodal approaches to creative AI applications.

Keywords

Multimodal Models, Transformer, Encoder, Decoder, Large Language Models, Natural Language Processing, Music Information Retrieval

Ευχαριστίες

Κατ' αρχάς, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου Αλέξανδρο Ποταμιάνο για την πολύ χρήσιμη βοήθεια και καθοδήγησή του, η οποία ήταν καθοριστική για τη βελτίωση και τη διαμόρφωση αυτής της διπλωματικής στην τελική της μορφή.

Θα ήθελα επίσης να ευχαριστήσω τον υποψήφιο διδάκτορα Χαρίλαο Παπαϊωάννου για την παροχή των γνώσεων και της εμπειρίας του στον τομέα της Ανάκτησης Μουσικής Πληροφορίας, αλλά και την βοήθεια που μου παρείχε κατά τη διάρκεια των δύσκολων τμημάτων αυτής της διπλωματικής.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου, τους φίλους μου και το αγόρι μου, που ήταν δίπλα μου κατά τη διάρκεια ολοκλήρωσης αυτής της εργασίας, αλλά και κατά τη διάρκεια όλων αυτών των πέντε ετών των σπουδών μου.

Αθήνα, Οκτώβριος 2024

Όλγα Μπάρλου

Περιεχόμενα

List of Figures

List of Tables

Κεφάλαιο 0

Εκτεταμένη Ελληνική Περίληψη

0.1 Εισαγωγή

0.1.1 Κίνητρο

Η μουσική και το τραγούδι υπήρξαν εδώ και καιρό ισχυρά μέσα αυτοέκφρασης και μια κύρια μορφή ψυχαγωγίας. Για να είναι ένα τραγούδι ποιοτικό, οι στίχοι πρέπει να εναρμονίζονται με τη μουσική, επιτυγχάνοντας μια ισορροπία μεταξύ δημιουργικότητας, συνοχής και φυσικής ροής—κριτήρια που είναι δύσκολο τόσο να μοντελοποιηθούν όσο και να αξιολογηθούν. Αυτή η πολυπλοκότητα έχει οδηγήσει σε ένα αισθητό κενό στην έρευνα, όσον αφορά στη δημιουργία στίχων με μουσική επίβλεψη. Παρά τις πρόσφατες καινοτομίες στα πολυτροπικά μοντέλα και στα μεγάλα γλωσσικά μοντέλα (LLMs), οι συγκεκριμένες προκλήσεις της ευθυγράμμισης της μουσικής με τους στίχους παραμένουν ανεξερεύνητες.

Αυτή η διπλωματική στοχεύει να γεφυρώσει αυτό το κενό αξιοποιώντας τις προόδους της τεχνητής νοημοσύνης για τη βελτίωση της δημιουργίας στίχων. Αναπτύσσοντας μοντέλα που μπορούν να κατανοούν και να παράγουν στίχους σύμφωνα με το μουσικό περιεχόμενο, επιδιώκουμε να συμβάλουμε τόσο στον τομέα της τεχνητής νοημοσύνης όσο και στον τομέα της μουσικής. Αυτά τα μοντέλα έχουν πρακτικές εφαρμογές, παρέχοντας στους καλλιτέχνες εργαλεία που διευκολύνουν τη δημιουργία στίχων.

0.1.2 Συνεισφορά

Αυτή η διπλωματική εργασία συνεισφέρει στον τομέα Ανάκτησης Μουσικής Πληροφορίας (MIR) όσον αφορά τη δημιουργία στίχων. Δεδομένου ότι το συγκεκριμένο θέμα της διπλωματικής - η δημιουργία στίχων με βάση την μουσική συνοδεία - δεν έχει μελετηθεί τόσο εκτενώς όσο άλλες παρόμοια generative tasks, όπως η δημιουργία μουσικής από στίχους ή η δημιουργία στίχων με βάση τη μελωδία του τραγουδιού, η διπλωματική εξερευνά αρκετές αρχιτεκτονικές και την απόδοσή τους στη δημιουργία στίχων.

Πιο συγκεκριμένα, οι κύριες συνεισφορές της παρούσας διπλωματικής είναι:

- Είναι η πρώτη ενδελεχής συγκριτική μελέτη πολλαπλών αρχιτεκτονικών LLM (Claude, GPT-2, Mistral OpenOrca και Vicuna) για την παραγωγή στίχων από κείμενο σε κείμενο.
- Υλοποίηση και αξιολόγηση μιας αρχιτεκτονικής παραγωγής στίχων με μουσική επίβλε-

ψη, η οποία έχει χρησιμοποιηθεί στην βιβλιογραφία, αλλά χωρίς εστίαση στην παραγωγή στίχων: το μοντέλο Whisper-OpenOrca, το οποίο χρησιμοποιεί ένα projection layer για την ευθυγράμμιση μουσικών και κειμενικών αναπαραστάσεων μεταξύ του προ-εκπαιδευμένου παγωμένου κωδικοποιητή ήχου Whisper και του παγωμένου LLM OpenOrca.

- Πρώτη πειραματική μελέτη που συγκρίνει διαφορετικές πολυτροπικές προσεγγίσεις για την παραγωγή στίχων.
- Ανάπτυξη νέων πολυτροπικών διατάξεων που ενσωματώνουν πρόσθετες μορφές στη διαδικασία παραγωγής στίχων:
  - SALMONN-Claude: μια διάταξη που χρησιμοποιεί το SALMONN για την εξαγωγή μουσικών ετικετών από τη μουσική είσοδο και στη συνέχεια χρησιμοποιεί το Claude για τη δημιουργία των τελικών στίχων με βάση αυτές τις ετικέτες.
  - SALMONN-Stable Diffusion-LLaVA: μια διάταξη που χρησιμοποιεί το SALMONN για να παράγει μια περιγραφή σκηνής ταινίας από τη μουσική είσοδο, στη συνέχεια χρησιμοποιεί Stable Diffusion για να παράγει μια εικόνα με βάση αυτή την περιγραφή και παράγει τους τελικούς στίχους χρησιμοποιώντας το μοντέλο LLaVA.

0.2 Μηχανική Μάθηση

Αυτή η υποενότητα εξετάζει πολυτροπικά μοντέλα, μεγάλα γλωσσικά μοντέλα (LLMs), Variational Autoencoders (VAEs) και το μοντέλο Stable Diffusion, παρουσιάζοντας τη σημασία τους σε διάφορες εφαρμογές μηχανικής μάθησης.

Τα **Πολυτροπικά μοντέλα** μπορούν να επεξεργάζονται και να συνδυάζουν δεδομένα από διαφορετικές τροπικότητες (modalities), όπως κείμενο, ήχος, εικόνες και βίντεο. Τα πολυτροπικά μοντέλα ήχου ενσωματώνουν ηχητικά δεδομένα με δεδομένα άλλων modalities, βελτιώνοντας συστήματα αναγνώρισης ομιλίας και διαλόγου. Η αρχιτεκτονική τους συνήθως περιλαμβάνει έναν κωδικοποιητή ήχου, ένα LLM και μια διεπαφή που συνδέει τις δύο μορφές δεδομένων [18]. Παρά τη χρησιμότητά τους, αντιμετωπίζουν προκλήσεις όπως η ποιότητα του ήχου και η συγχρονισμένη ενσωμάτωση με άλλες μορφές δεδομένων.

Τα **Μεγάλα Γλωσσικά Μοντέλα** (LLMs) βασίζονται στην αρχιτεκτονική του transformer, η οποία παρουσιάστηκε πρώτα στο paper “Attention Is All You Need” [19], και έχουν φέρει επανάσταση στην επεξεργασία φυσικής γλώσσας (NLP). Τα LLMs κατασκευάζονται μέσω προεκπαίδευσης σε τεράστια σύνολα κειμένων και μπορούν να εκτελούν καθήκοντα με λίγα ή καθόλου παραδείγματα (zero/few-shot learning) [20]. Οι τρεις κύριες αρχιτεκτονικές είναι: μόνο κωδικοποιητής, μόνο αποκωδικοποιητής και συνδυασμός των δύο, με κάθε μοντέλο να εξυπηρετεί διαφορετικούς τύπους εφαρμογών όπως η ταξινόμηση και η δημιουργία κειμένου [20].

Οι **Variational Autoencoders (VAEs)** είναι generative μοντέλα που μαθαίνουν λανθάνουσες μεταβλητές που αντιπροσωπεύουν σύνθετες κατανομές δεδομένων. Βασίζονται στη

συνδυαστική χρήση νευρωνικών δικτύων και πιθανοθεωρητικής μοντελοποίησης για να δημιουργήσουν αναπαραστάσεις και να παράγουν δεδομένα, όπως εικόνες και κείμενα [21]. Αντιμετωπίζουν προκλήσεις, όπως η ποιότητα της παραγόμενης γλώσσας λόγω της φύσης των διακριτών λέξεων στα κείμενα [21].

Τέλος, το **Stable Diffusion** είναι ένα μοντέλο δημιουργίας εικόνων από κείμενο, που βασίζεται σε πιθανοτικά μοντέλα διάχυσης αποθορυβοποίησης (Denoising Diffusion Probabilistic Models - DDPMs). Σε αντίθεση με τα GANs, τα μοντέλα διάχυσης προσθέτουν και αφαιρούν θόρυβο στα δεδομένα, μοντελοποιώντας τις σύνθετες κατανομές εικόνων. Το Stable Diffusion χρησιμοποιεί την αρχιτεκτονική Latent Diffusion Models (LDMs), που λειτουργεί σε συμπιεσμένο χώρο, προσφέροντας υψηλής ποιότητας δημιουργίες εικόνων με χαμηλότερο υπολογιστικό κόστος [22].

0.3 Πολυτροπικά Μοντέλα Ήχου και Γλωσσικά Μοντέλα για Παραγωγή Στίχων

Σε αυτή την υποενότητα, αναλύουμε τα μοντέλα που είναι σχετικά με την δουλειά μας. Συγκεκριμένα, εξετάζουμε τη βιβλιογραφία σχετικά με τα πολυτροπικά μοντέλα ήχου και συγκεκριμένα τους κωδικοποιητές ήχου που χρησιμοποιούνται συχνά σε πολυτροπικά μοντέλα ήχου, και τα μοντέλα κατανόησης μουσικής. Εξετάζουμε επίσης, πολυτροπικά μοντέλα όρασης, συγκεκριμένα μεγάλα γλωσσικά μοντέλα τα οποία είτε χρησιμοποιούνται ρητά στην εργασία μας, είτε αποτελούν μέρος άλλων χρησιμοποιούμενων πολυτροπικών γλωσσικών μοντέλων. Τέλος, εξετάζουμε διάφορες μεθόδους που χρησιμοποιούνται στη βιβλιογραφία για την αξιολόγηση generative tasks.

0.3.1 Κωδικοποιητές ήχου στα Πολυτροπικά Μοντέλα Ήχου

Ο κωδικοποιητής ήχου **Whisper** της OpenAI, όπως παρουσιάστηκε στο άρθρο “Robust Speech Recognition via Large-Scale Weak Supervision” [23], είναι ένα σύστημα αναγνώρισης ομιλίας με μεγάλη κλίμακα, εκπαιδευμένο σε 680.000 ώρες πολυγλωσσικών δεδομένων. Χρησιμοποιεί αρχιτεκτονική κωδικοποιητή-αποκωδικοποιητή βασισμένη σε transformers για να μετατρέπει τον ήχο σε υψηλής ποιότητας μεταγραφές. Το Whisper ξεχωρίζει για τις δυνατότητές του στην αναγνώριση σε πολλαπλές γλώσσες και εργασίες, και τη δυνατότητα αποδοτικής λειτουργίας σε συνθήκες χωρίς προσαρμογή (zero-shot).

Ο **Audio Spectrogram Transformer** (AST) [24] είναι το πρώτο μοντέλο που χρησιμοποιεί μια πλήρως attention-based αρχιτεκτονική για ταξινόμηση ήχου, χωρίς τη χρήση συνελικτικών στρωμάτων (CNN). Βασίζεται στον transformer για την επεξεργασία ηχητικών φασματογραμμάτων και επιτυγχάνει καλύτερα αποτελέσματα από τα CNN μοντέλα, ιδίως σε εργασίες που απαιτούν ανάλυση χρονικών εξαρτήσεων μεγάλης εμβέλειας.

Το **MERT** (Music undERstanding model with large-scale self-supervised Training) [9] είναι ένα μοντέλο μεγάλης κλίμακας αυτοεποπτευόμενης μάθησης (self-supervised learning), σχεδιασμένο για την κατανόηση ακουστικής μουσικής. Χρησιμοποιεί μοντέλα-δασκάλους (teacher models), έναν ακουστικό και έναν μουσικό, για να εκπαιδεύεται σε χαρακτηριστικά που αφορούν τον ήχο και τη μουσική. Το MERT έχει αποδειχθεί εξαιρετικά αποδοτικό

σε διάφορες εργασίες κατανόησης μουσικής, όπως μουσική σήμανση (tagging), εντοπισμός ρυθμού και ταξινόμηση μουσικών ειδών.

0.3.2 Μοντέλα Κατανόησης Μουσικής

Στην υποενότητα αυτή εξετάζονται μοντέλα κατανόησης μουσικής σε πολυτροπικά μοντέλα ήχου που χρησιμοποιούνται για εργασίες όπως η σήμανση μουσικής (music tagging), η περιγραφή (captioning) και η απάντηση σε ερωτήσεις (Q&A) σχετικά με μουσικά δεδομένα.

Το **SALMONN** (Speech Audio Language Music Open Neural Network) [25] είναι ένα καινοτόμο πολυτροπικό μοντέλο που επεξεργάζεται ήχο, λόγο και μουσική. Η αρχιτεκτονική του χρησιμοποιεί δύο κωδικοποιητές ήχου (Whisper και BEATs) που συνδυάζονται μέσω ενός Q-Former, επιτρέποντας στο μοντέλο να κατανοεί και να αναπαριστά διαφορετικούς τύπους ήχου. Διαθέτει δυνατότητες γενίκευσης σε tasks τα οποία δεν έχει εκπαιδευτεί, και παρουσιάζει εξαιρετικές επιδόσεις σε εργασίες κατανόησης ήχου.

Το **MusiLingo** [10] γεφυρώνει το χάσμα μεταξύ μουσικής και φυσικής γλώσσας. Βασισμένο στον κωδικοποιητή MERT και στο LLM Vicuna, το μοντέλο εξάγει ακουστικά χαρακτηριστικά από μουσικά κομμάτια και τα μετατρέπει σε κατανοητά κείμενα. Εκπαιδευμένο σε δεδομένα περιγραφών μουσικής και ερωτήσεων-απαντήσεων, παρέχει ακριβείς περιγραφές και απαντήσεις σε μουσικές ερωτήσεις, προσφέροντας εξαιρετικές επιδόσεις σε MIR εργασίες.

Το **MU-LLaMA** [26], ένα μοντέλο κατανόησης μουσικής και δημιουργίας caption, βασίζεται στο MERT και το γλωσσικό μοντέλο LLaMA. Στόχος του είναι να ξεπεράσει τους περιορισμούς στην έλλειψη δεδομένων για τη δημιουργία μουσικής από κείμενο. Το μοντέλο επιτυγχάνει υψηλές επιδόσεις σε καθήκοντα όπως η περιγραφή και η απάντηση σε μουσικές ερωτήσεις, υπερέχοντας σε πολλά αξιολογικά κριτήρια.

Το **M^{2}UGen** [27] ενσωματώνει την κατανόηση και τη δημιουργία μουσικής χρησιμοποιώντας πολυτροπικά δεδομένα όπως εικόνες και βίντεο, επιπροσθέτως της μουσικής. Συνδυάζει προεκπαιδευμένους κωδικοποιητές και χρησιμοποιεί το LLaMA 2 για να επεξεργάζεται και να δημιουργεί μουσική με βάση εισαγωγές από κείμενο, εικόνες και βίντεο. Το M^{2}UGen ξεχωρίζει για τις επιδόσεις του σε εργασίες πολυτροπικής κατανόησης και δημιουργίας μουσικής, παρουσιάζοντας εξαιρετικά αποτελέσματα στην καλλιτεχνική δημιουργία μουσικής μέσω τεχνητής νοημοσύνης.

0.3.3 Πολυτροπικά Μοντέλα Όρασης

Τα πολυτροπικά μοντέλα όρασης είναι από τα πιο μελετημένα στον τομέα της τεχνητής νοημοσύνης, καθώς συνδυάζουν δεδομένα όρασης με γλωσσικά μοντέλα, επιτρέποντας εφαρμογές όπως η περιγραφή εικόνων (image captioning), η απάντηση σε ερωτήσεις βασισμένες σε εικόνες (visual question answering - VQA) και η πολυτροπική λογική και διάλογοι. Αυτά τα μοντέλα ενσωματώνουν πληροφορίες από εικόνες και βίντεο με κείμενο για να παρέχουν σύνθετες ερμηνείες και να διαχειρίζονται δεδομένα σε πολυτροπικά πλαίσια [28].

Η αρχιτεκτονική τους περιλαμβάνει συνήθως έναν κωδικοποιητή εικόνας ή βίντεο, ένα προεκπαιδευμένο μεγάλο γλωσσικό μοντέλο (LLM) και μια διεπαφή πολυτροπικών δεδομένων που συγχρονίζει τα οπτικά χαρακτηριστικά με τις γλωσσικές πληροφορίες. Προηγμένες τεχνικές όπως τα cross-attention layers έχουν βελτιώσει σημαντικά την αλληλεπίδραση

μεταξύ εικόνας και γλώσσας, επιτρέποντας βαθύτερη κατανόηση και λογική [18].

Τα μοντέλα αυτά βρίσκουν εφαρμογή σε ποικίλα πεδία, όπως η απάντηση σε ερωτήσεις βασισμένες σε εικόνες, η περιγραφή εικόνων, και η καθοδήγηση ρομπότ ή εικονικών πρακτόρων στην αλληλεπίδραση με τον πραγματικό κόσμο. Παρ' όλα αυτά, εξακολουθούν να υπάρχουν προκλήσεις όπως οι 'παραισθήσεις πολυτροπικών μοντέλων' (multimodal hallucinations), όπου το μοντέλο προσθέτει ή παρερμηνεύει πληροφορίες που δεν υπάρχουν στην εικόνα. Η διαχείριση των πολυτροπικών δεδομένων σε μεγάλης ανάλυσης εικόνες και βίντεο αποτελεί επίσης μια δύσκολη πρόκληση [18].

Το **LLaVA** (Large Language and Vision Assistant) [29] είναι ένα προηγμένο πολυτροπικό μοντέλο που συνδυάζει τον κωδικοποιητή εικόνας CLIP με το γλωσσικό μοντέλο Vicuna, επιτρέποντας στο μοντέλο να χειρίζεται σύνθετες οπτικο-γλωσσικές εργασίες. Μέσω της τεχνικής Visual Instruction Tuning, το μοντέλο εκπαιδεύεται σε δεδομένα που μιμούνται την ανθρώπινη συμπεριφορά εκτέλεσης οδηγιών (instruction following). Το LLaVA επιτυγχάνει εξαιρετικά αποτελέσματα σε καθήκοντα όπως η απάντηση σε ερωτήσεις που βασίζονται σε εικόνες και η λογική βασισμένη σε εικόνες, ξεπερνώντας μοντέλα όπως το BLIP-2 και το Qwen-VL-Chat.

Η μέθοδος visual instruction tuning παίζει καθοριστικό ρόλο στην επιτυχία του LLaVA, επιτρέποντάς του να γενικεύει σε διαφορετικούς τομείς όρασης και να εκτελεί σύνθετες εργασίες λογικής που απαιτούν κατανόηση τόσο της εικόνας όσο και της γλώσσας.

0.3.4 Μεγάλα Γλωσσικά Μοντέλα

Αυτή η ενότητα αναλύει μεγάλα γλωσσικά μοντέλα (LLMs) όπως το GPT-2, το LLaMA, το Vicuna, το Mistral, το Mistral-OpenOrca και το Claude 3, τα οποία αντιπροσωπεύουν την εξέλιξη των γλωσσικών μοντέλων με στόχο την ενίσχυση της επίδοσης σε διάφορες εργασίες επεξεργασίας φυσικής γλώσσας (NLP).

Το **GPT-2** [30] της OpenAI ήταν ένα από τα πρώτα σημαντικά βήματα στην εξέλιξη των γλωσσικών μοντέλων, με 1.5 δισεκατομμύρια παραμέτρους. Η αρχιτεκτονική του βασίζεται σε ένα μόνο αποκωδικοποιητή και είναι αυτοπαλινδρομικό (autoregressive), προβλέποντας το επόμενο token σε μια ακολουθία με βάση τα προηγούμενα. Το GPT-2 έδειξε τις δυνατότητες των μεγάλων μοντέλων χωρίς την ανάγκη ειδικής προσαρμογής (fine-tuning), αν και έχει περιορισμούς, όπως τη δυσκολία με μεγάλες ακολουθίες και θέματα προκατάληψης.

Το **LLaMA** [31] της Meta, ένα ανοιχτού κώδικα μοντέλο, απέδειξε ότι τα μικρότερα μοντέλα μπορούν να αποδίδουν εξίσου καλά με τα μεγαλύτερα. Το LLaMA χρησιμοποίησε τεχνικές όπως τα rotary positional embeddings και τα στρώματα ενεργοποίησης SwiGLU για βελτίωση της αποδοτικότητας. Έχει χρησιμοποιηθεί ως βάση για μοντέλα όπως το Vicuna και το Alpaca, αλλά αντιμετωπίζει προκλήσεις όπως η τοξικότητα και το hallucination στα αποτελέσματα [20].

Το **Vicuna** [32], με 13 δισεκατομμύρια παραμέτρους, είναι ένα finetuned μοντέλο που βασίζεται στο LLaMA και έχει προσαρμοστεί για διαλόγους. Παρά τις ανταγωνιστικές επιδόσεις του σε διαλογικές εφαρμογές, παρουσιάζει προκλήσεις στη διατήρηση της συνοχής σε εκτενείς διαλόγους.

Το **Mistral** [15] είναι ένα από τα νεότερα LLMs, το οποίο, με 7 δισεκατομμύρια πα-

ραμέτρους, υπερέχει σε εργασίες όπως η δημιουργία κώδικα και η επίλυση μαθηματικών προβλημάτων. Διαθέτει προηγμένες τεχνικές όπως το grouped-query attention, βελτιώνοντας την αποδοτικότητά του, αλλά εξακολουθεί να αντιμετωπίζει προκλήσεις στην κατανόηση γλώσσας γενικού σκοπού.

Το μοντέλο **Mistral-7B-OpenOrca** [33] είναι μια παραλλαγή του μοντέλου Mistral-7B που έχει προσαρμοστεί με εκπαίδευση (fine-tuning). Εκπαιδεύτηκε σε ένα προσεκτικά επιλεγμένο υποσύνολο δεδομένων από το σύνολο OpenOrca, το οποίο έχει ενισχυθεί με δεδομένα από το GPT-4 και σχεδιάστηκε για να αναπαράγει το σύνολο δεδομένων που χρησιμοποιήθηκε στην έρευνα Orca της Microsoft. Το Mistral-7B-OpenOrca υπερέχει έναντι άλλων μοντέλων στην κατηγορία του μεγέθους του, καταλαμβάνοντας την πρώτη θέση στο Leaderboard του Hugging Face για μοντέλα μικρότερα από 30B παραμέτρους κατά την κυκλοφορία του. Το μοντέλο πέτυχε σημαντική αύξηση στην επίδοση σε διάφορα benchmarks, όπως MMLU, ARC και HellaSwag, με εξαιρετικά αποτελέσματα σε tasks λογικής, μαθηματικών και παραγωγής κώδικα. Η διαδικασία fine-tuning του μοντέλου περιλάμβανε 4 εποχές εκπαίδευσης σε 8 A6000 GPUs, επιτυγχάνοντας αξιόλογα αποτελέσματα με χαμηλό κόστος.

Τέλος, το **Claude 3** [34] της Anthropic αποτελεί σημαντική εξέλιξη στα γλωσσικά μοντέλα, με έμφαση στη δημιουργική γραφή και την ανάλυση. Τα μοντέλα της σειράς Claude 3, και κυρίως τα Opus και Sonnet, παρουσιάζουν αυξημένες ικανότητες σε εργασίες που αφορούν τη δημιουργική γραφή, την λεπτομερή ανάλυση και την παραγωγή δομημένων κειμένων. Σύμφωνα με την Anthropic, το Claude 3 είναι καλύτερο στην δημιουργική γραφή σε σύγκριση με το Claude 2.1. Συγκεκριμένα, το Claude 3 καταγράφει ποσοστό επιτυχίας 63% σε σχέση με το βασικό μοντέλο Claude Instant στην δημιουργική γραφή. Οι συγγραφείς της εργασίας "A Confederacy of Models: a Comprehensive Evaluation of LLMs on Creative Writing" [35] αναφέρουν ότι το Claude Instant 1.2 κατέλαβε την τρίτη ή υψηλότερη θέση σε όλα τα κριτήρια αξιολόγησης τους, κατακτώντας τη δεύτερη θέση στην συνοχή και τη δεύτερη θέση συνολικά, μετά το GPT-4. Αυτό δείχνει ότι το Claude 3, λαμβάνοντας υπόψη την καλύτερη επίδοσή του στη δημιουργική γραφή, μπορεί να θεωρηθεί πολύ υποσχόμενο για τη δημιουργία δημιουργικών και συνεκτικών στίχων. Επιπλέον, όπως αναφέρεται από την Anthropic, το πιο πρόσφατο μοντέλο, το Claude 3.5 Sonnet, δείχνει ακόμη καλύτερη επίδοση σε όλες τις αξιολογημένες εργασίες σε σχέση με το Claude 3 [36].

0.3.5 Μέθοδοι αξιολόγησης σε Εργασίες Παραγωγής Λόγου

Οι μέθοδοι αξιολόγησης σε εργασίες παραγωγής λόγου, όπως η δημιουργία στίχων, είναι πιο σύνθετες από ότι στις εργασίες ταξινόμησης. Σε εργασίες ταξινόμησης, η επίδοση ενός μοντέλου μετριέται εύκολα συγκρίνοντας την έξοδο με τις αληθινές ετικέτες του dataset. Ωστόσο, αυτή η προσέγγιση δεν μπορεί πάντα να εφαρμοστεί σε generative tasks, όπου οι εξόδοι είναι δημιουργικές και πιο ελεύθερες. Στην ενότητα αυτή, εξετάζονται μερικές από τις μεθόδους που χρησιμοποιούνται στη βιβλιογραφία και είναι κατάλληλες για την αξιολόγηση μοντέλων παραγωγής στίχων.

0.3.5.1 Αντικειμενικές Μετρικές

Στο paper MusicJam [1], χρησιμοποιήθηκαν τέσσερις βασικές αντικειμενικές μετρικές για την αξιολόγηση της ποιότητας των παραγόμενων στίχων: BLEU, Distinct/Diversity, Novelty, και Coherence.

- BLEU [37]: Αυτή η μετρική χρησιμοποιείται κυρίως στην αυτόματη μετάφραση και μετρά την επικάλυψη n-grams μεταξύ των αληθινών στίχων και των παραγόμενων στίχων. Η μετρική BLEU βασίζεται στην εξίσωση:

$$BLEU_N = BP \cdot \sum_{n=1}^{N} exp(w_n log p_n),$$

όπου το BP είναι μια ποινή που επιβάλλεται στα σύντομα αποτελέσματα, ώστε να μην ευνοούνται οι μικρές προτάσεις. Το BLEU παρέχει μια εικόνα για το πόσο κοντά είναι οι παραγόμενοι στίχοι στους αληθινούς στίχους, αλλά συχνά προάγει την ακριβή αντιστοιχία λέξεων, παραμελώντας τη δημιουργικότητα.

- Distinct/Diversity [38]: Αυτή η μετρική μετρά την ποικιλία των παραγόμενων στίχων, υπολογίζοντας την αναλογία των μοναδικών n-grams σε σχέση με τον συνολικό αριθμό n-grams. Η εξίσωση της μετρικής είναι:

$$Distinct_N = \frac{\mid unique(Ngrams) \mid}{\mid Ngrams \mid}.$$

Η μετρική αυτή προάγει τη λεξιλογική ποικιλία, αλλά υπάρχει ο κίνδυνος να οδηγήσει σε στίχους που είναι υπερβολικά ασύνδετοι και μη φυσικοί.

- Novelty [39]: Η μετρική αυτή μετρά τον λόγο των σπάνιων n-grams σε σχέση με το συνολικό αριθμό των n-grams. Σπάνιες θεωρούνται οι φράσεις που δεν βρίσκονται ανάμεσα στις 2000 πιο συχνές. Η εξίσωση της μετρικής είναι:

$$Novelty_N = \frac{\mid infrequent(Ngrams) \mid}{\mid Ngrams \mid}.$$

Το Novelty αξιολογεί την πρωτοτυπία των στίχων, συγκρίνοντας τους με τις πιο κοινές φράσεις, προσδιορίζοντας έτσι την καινοτομία στη γλώσσα που χρησιμοποιείται.

- Coherence [40]: Η μετρική αυτή μετρά τη συνοχή των στίχων, υπολογίζοντας τον αριθμό των λέξεων που επαναλαμβάνονται σε έναν στίχο και παίρνοντας τον μέσο όρο των επιμέρους αποτελεσμάτων. Η εξίσωση είναι:

$$Coherence = \frac{1}{M} \cdot \sum_{k=1}^{M} \sum_{i=1}^{n_k} \mathbb{1}(count(w_i) > 1),$$

όπου M είναι ο αριθμός των τραγουδιών και n_k ο αριθμός των λέξεων που παράχθηκαν για το τραγούδι k. Αν και η συνοχή είναι κρίσιμη για την ποιότητα ενός τραγουδιού, αυτή η προσέγγιση δεν μετρά επαρκώς τη σημασιολογική συνοχή, καθώς η επανάληψη

λέξεων μπορεί να οδηγήσει σε υψηλή βαθμολογία χωρίς οι στίχοι να είναι πραγματικά συνεκτικοί ή φυσικοί.

Ενώ οι παραπάνω μετρικές παρέχουν χρήσιμες πληροφορίες, καθεμία έχει τους περιορισμούς της. Το BLEU και το Coherence τονίζουν τη γλωσσική ακρίβεια εις βάρος της δημιουργικότητας, ενώ οι μετρικές Distinct και Novelty μπορεί να προωθήσουν την μοναδικότητα εις βάρος της σαφήνειας. Μια πιο ολοκληρωμένη αξιολόγηση θα μπορούσε να περιλαμβάνει ανθρώπινες κρίσεις, οι οποίες λαμβάνουν υπόψη τη θεματική συνέπεια, το συναισθηματικό βάθος και τη συνολική μουσικότητα των παραγόμενων στίχων.

0.3.5.2 Βαθμολογία Ομοιότητας με Cross Encoder

Οι Cross-encoders είναι μοντέλα βασισμένα σε μετασχηματιστές (transformers), σχεδιασμένοι για να συλλαμβάνουν τη σχέση μεταξύ δύο εισόδων. Ειδικότερα, οι cross-encoders λαμβάνουν δύο εισόδους και τις κωδικοποιούν μαζί σε μια κοινή αναπαράσταση. Αυτό διαφέρει από τους bi-encoders, όπου οι δύο είσοδοι περνούν ανεξάρτητα στο μοντέλο BERT. Παρόλο που η χρήση cross-encoder είναι πιο απαιτητική υπολογιστικά, μπορεί να αποτυπώσει με μεγαλύτερ ακρίβεια την ομοιότητα μεταξύ δύο κειμένων [41].

Αυτή η μέθοδος βαθμολόγησης της ομοιότητας βοηθά να αποφευχθούν οι αδυναμίες μετρικών όπως το BLEU, το οποίο υπολογίζει μόνο την επικάλυψη των n-grams. Το BLEU αγνοεί περιπτώσεις όπου δύο κείμενα έχουν το ίδιο νόημα αλλά χρησιμοποιούν συνώνυμες λέξεις ή διαφορετική φρασεολογία. Οι cross-encoders μπορούν να αναγνωρίσουν αυτές τις περιπτώσεις και να αποδώσουν μια πιο ακριβή εκτίμηση της ομοιότητας των κειμένων.

Αυτό καθιστά τη χρήση cross-encoders κατάλληλη για την αξιολόγηση της δημιουργίας στίχων, καθώς μπορεί να μετρήσει την ομοιότητα σε επίπεδο νοήματος και όχι μόνο γλωσσικής ακρίβειας, κάτι που είναι ζωτικής σημασίας για τα generative tasks όπου η δημιουργικότητα και η παραλλαγή στη φρασεολογία μπορεί να είναι εξίσου σημαντικές με την ακρίβεια.

0.3.5.3 JudgeLM: Κλιμακούμενο LLM για Αξιολόγηση

Το JudgeLM είναι ένα εκπαιδευμένο μεγάλο γλωσσικό μοντέλο σχεδιασμένο να αξιολογεί την επίδοση άλλων LLMs σε ανοικτού τύπου εργασίες [17]. Οι παραδοσιακές μετρικές αξιολόγησης συχνά αποτυγχάνουν να αξιολογήσουν τα LLMs με ακρίβεια, λόγω της πολυπλοκότητας και της μεταβλητότητας των εξόδων τους. Το JudgeLM αντιμετωπίζει αυτό το πρόβλημα μέσω της εκπαίδευσης open-source μοντέλων, όπως το Vicuna, χρησιμοποιώντας ένα μεγάλης κλίμακας σύνολο δεδομένων που περιλαμβάνει 105.000 tasks και αξιολογήσεις παραγόμενες από το GPT-4. Το μοντέλο έχει σχεδιαστεί για να λειτουργεί ως κλιμακούμενος και αποτελεσματικός αξιολογητής, επιτυγχάνοντας υψηλότερα επίπεδα συμφωνίας από ότι οι ανθρώπινοι αξιολογητές.

Το JudgeLM διαθέτει αρχιτεκτονική που είναι διαθέσιμη σε μεγέθη που κυμαίνονται από 7B έως 33B παραμέτρους. Για να μετριάσει τις έμφυτες προκαταλήψεις (biases), όπως η προκατάληψη θέσης (position bias), γνώσης (knowledge bias) και μορφής (format bias), χρησιμοποιεί τεχνικές όπως swap augmentation, reference support και reference drop.

Η τεχνική swap augmentation εκπαιδεύει το μοντέλο να κρίνει το περιεχόμενο και όχι τη θέση των απαντήσεων, αλλάζοντας τη σειρά των απαντήσεων στα δεδομένα εκπαίδευσης. Το reference support βοηθά το μοντέλο να βελτιώσει την ακρίβεια σε εργασίες βασισμένες σε γεγονότα, ενώ το reference drop το καθιστά ικανό να αξιολογεί τόσο με, όσο και χωρίς αναφορές, προσφέροντας μεγαλύτερη ευελιξία.

Το JudgeLM μπορεί να αξιολογήσει διάφορα μοντέλα σε πολλές εργασίες, όπως αξιολόγηση μίας ή πολλαπλών απαντήσεων, πολυτροπικών μοντέλων και διαλόγους πολλαπλών γύρων. Συγκρίνει τις εξόδους των μοντέλων με αναφορές ή με τα αποτελέσματα άλλων μοντέλων, παρέχοντας λεπτομερείς βαθμολογίες και επεξηγήσεις. Σε πολλές αξιολογήσεις, επιτυγχάνει πάνω από 90% συμφωνία με το GPT-4, ξεπερνώντας ακόμη και το GPT-3.5. Επιπλέον, το JudgeLM μπορεί να αξιολογήσει 5.000 δείγματα σε μόλις τρία λεπτά, χρησιμοποιώντας 8 A100 GPUs, καθιστώντας το μια οικονομικά αποδοτική και κλιμακούμενη λύση σε σύγκριση με παραδοσιακές μεθόδους αξιολόγησης από ανθρώπους ή το GPT-4.

0.3.5.4 Αξιολόγηση με βάση τα LLMs

Πρόσφατες μελέτες δείχνουν ότι οι αξιολογήσεις που πραγματοποιούνται από μεγάλα γλωσσικά μοντέλα (LLMs) μπορεί να είναι πιο αποτελεσματικές από τις αντικειμενικές μετρικές, ειδικά σε εργασίες που απαιτούν δημιουργικότητα και ποικιλομορφία [42]. Δίνοντας συγκεκριμένα κριτήρια, όπως η δημιουργικότητα, η συνοχή, η φυσικότητα και το πόσο εύκολα μπορεί να τραγουδηθεί, τα ισχυρά LLMs μπορούν να αξιολογήσουν αποτελεσματικά τους παραγόμενους στίχους.

Ένα πρόβλημα που συχνά προκύπτει με αυτή τη μέθοδο είναι η προκατάληψη που μπορεί να ενσωματώνουν τα LLMs όταν κάνουν αυτές τις κρίσεις. Έχουν παρατηρηθεί διάφορες μορφές προκατάληψης, όπως η προκατάληψη θέσης, όπου το μοντέλο τείνει να προτιμά συγκεκριμένες θέσεις, η προκατάληψη υπερβολικής φλυαρίας, όπου ευνοούνται μακροσκελείς απαντήσεις ακόμα και αν είναι χαμηλότερης ποιότητας, και η προκατάληψη αυτοενίσχυσης, όπου τα μοντέλα προτιμούν τις απαντήσεις που έχουν δημιουργήσει τα ίδια [43].

Στο paper “Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena” [43], αναφέρεται ότι το GPT-4 εμφανίζει ποσοστό συμφωνίας 80% με τους ανθρώπινους αναλυτές, ποσοστό που είναι ίσο με τη συμφωνία μεταξύ των ανθρώπινων αξιολογητών.

Ωστόσο, στο paper “Reference-Guided Verdict: LLMs-as-Judges in Automatic Evaluation of Free-Form Text” [44], διαπιστώθηκε ότι η χρήση πολλαπλών και ποικίλων LLMs ως κριτές μειώνει τις προκαταλήψεις των μοντέλων και βελτιώνει σημαντικά την ευθυγράμμιση με τις ανθρώπινες κρίσεις, ειδικά σε δύσκολα tasks που περιλαμβάνουν ελεύθερο κείμενο. Στην συγκεκριμένη μέθοδο χρησιμοποιήθηκαν δύο μοντέλα ανοιχτού κώδικα και ένα κλειστού κώδικα: Mistral-Instruct-7B-v0.3, Llama-3.1-70B και GPT-3.5-turbo.

0.4 Το σύνολο δεδομένων

Το σύνολο δεδομένων που χρησιμοποιήσαμε είναι το DALI, το οποίο περιλαμβάνει τραγούδια με συγχρονισμένο ήχο, στίχους και νότες [45]. Το DALI αποτελείται από 7756 τραγούδια και περιέχει αρχεία MP3 που συνοδεύονται από στίχους σε διαφορετικά επίπεδα

λεπτομέρειας, όπως παραγράφους, γραμμές, λέξεις και νότες. Κάθε τραγούδι συνοδεύεται από μεταδεδομένα, όπως είδος μουσικής, γλώσσα, καλλιτέχνη, τίτλο τραγουδιού και άλμπουμ.

Τα αρχεία MP3 του συνόλου δεδομένων δεν είναι διαχωρισμένα σε φωνή και συνοδεία. Για να αποκτήσουμε τη μουσική συνοδεία, χρησιμοποιήσαμε τη βιβλιοθήκη Spleeter [46], η οποία πραγματοποιεί διαχωρισμό σε δύο μέρη: φωνή και συνοδεία. Αυτός ο διαχωρισμός ήταν απαραίτητος για να αποτρέψουμε τα μοντέλα να δίνουν εξόδους βάσει της φωνής.

Επιπλέον, για να αυξήσουμε την επίδοση των εκπαιδευόμενων μοντέλων μας, χρειάστηκε να κάνουμε μια πιο ενδελεχή προεπεξεργασία του συνόλου δεδομένων. Συγκεκριμένα, το σύνολο δεδομένων περιέχει τα εξής χαρακτηριστικά:

- Λανθασμένη ορθογραφία μερικών λέξεων των στίχων.
- Διαχωρισμό λέξεων σε μικρότερα τμήματα λόγω του τρόπου προφοράς του τραγουδιστή (π.χ. tomo rrow αντί για tomorrow).
- Ασυνήθιστη χρήση αποστρόφων (π.χ. ev'rything αντί για everything).
- Λανθασμένος χαρακτηρισμός μερικών τραγουδιών ως αγγλικά, ενώ ήταν σε άλλη γλώσσα.

Για την αντιμετώπιση αυτών των προβλημάτων, πραγματοποιήσαμε την εξής προεπεξεργασία:

- Αφαιρέσαμε τα τραγούδια που είχαν λέξεις που επαναλαμβάνονταν πάνω από τρεις φορές, για να αποτρέψουμε την επανάληψη λέξεων στην έξοδο των εκπαιδευόμενων μοντέλων.
- Αφαιρέσαμε τραγούδια που είχαν κατηγοριοποιηθεί λανθασμένα ως αγγλικά με τη βοήθεια μοντέλου ανίχνευσης γλώσσας [47]. Μετά, με χειροκίνητο έλεγχο, αφαιρέσαμε και τραγούδια που, ενώ ήταν στα αγγλικά, περιείχαν μερικές φράσεις σε άλλες γλώσσες.
- Αφαιρέσαμε τα τραγούδια που είχαν στίχους με παύλες, και αντικαταστήσαμε τους ειδικούς χαρακτήρες, είτε με κενό χαρακτήρα είτε, στην περίπτωση ψηφίων, γράφοντας ολογράφως τον αριθμό.
- Αντικαταστήσαμε τη συνένωση λέξεων με αποστρόφους, με την πλήρη μορφή τους.
- Διορθώσαμε τη λανθασμένη ορθογραφία λέξεων, χρησιμοποιώντας την βιβλιοθήκη Spellchecker [48] για την ανίχνευση των λέξεων με ορθογραφικά λάθη, και μετά με χειροκίνητη διόρθωσή τους.

Μετά από αυτή την προεπεξεργασία, καταλήξαμε σε ένα σύνολο 3111 τραγουδιών, το οποίο χωρίστηκε σε training set και test set σε ποσοστό 80-20.

0.5 Τα μοντέλα που δοκιμάστηκαν για κάθε συνδυασμό από modalities

0.5.1 Από κείμενο σε κείμενο

Αρχικά, στη διαδικασία παραγωγή στίχων από κείμενο, δοκιμάσαμε τα LLMs που χρησιμοποιούνται και με άλλες μεθοδολογίες αργότερα, κάνοντας αυτά τα πειράματα να χρησιμεύουν ως μελέτες απαλοιφής (ablation studies) για τα προτεινόμενα pipelines παρακάτω. Τα LLMs που δοκιμάστηκαν είναι τα Claude 3.5 Sonnet, GPT-2, Mistral OpenOrca και Vicuna. Αυτή η δοκιμή πραγματοποιήθηκε αρχικά χωρίς περαιτέρω εκπαίδευση (out-of-the-box), δηλαδή χωρίς προσαρμογή στο σύνολο δεδομένων που επιλέξαμε. Επειδή το GPT-2 είναι completion μοντέλο και όχι instruction μοντέλο, παρέχουμε απλώς την πρώτη γραμμή του τραγουδιού χωρίς καμία εντολή.

Δοκιμάσαμε επίσης κατά πόσο το finetuning των GPT-2 και OpenOrca στο σύνολο δεδομένων DALI θα βελτίωνε την ποιότητα των παραγόμενων αποτελεσμάτων. Όσον αφορά τη μορφή του συνόλου δεδομένων για την εκπαίδευση, κάθε δείγμα εκπαίδευσης είναι ένα τραγούδι.

Χρησιμοποιήσαμε την τεχνική LoRA (Low-Rank Adaptation) για την προσαρμογή, η οποία επιτρέπει αποδοτική προσαρμογή των μεγάλων γλωσσικών μοντέλων.

0.5.2 Από κείμενο & ήχο σε κείμενο

Η διαδικασία δημιουργίας κειμένου από κείμενο & ήχο περιλαμβάνει την χρήση της μουσικής συνοδείας και των προηγούμενων στίχων ως είσοδο. Τα δύο μοντέλα που δοκιμάστηκαν σε αυτή την κατηγορία είναι το μοντέλο VAE AST-GPT2 και το μοντέλο Whisper-OpenOrca.

0.5.2.1 Μοντέλο VAE AST-GPT2

Στοχεύοντας στην αναπαραγωγή της δουλειάς των Chuer Chen et al. [1], ακολουθήσαμε μια παρόμοια προσέγγιση με αυτή που περιγράφεται στο paper τους. Καθώς ο κώδικας και η μέθοδος προεπεξεργασίας που χρησιμοποίησαν δεν είναι διαθέσιμα, προσπαθήσαμε να ακολουθήσουμε όσο το δυνατόν περισσότερο την δουλειά τους με βάση τις πληροφορίες που δίνονται στο paper.

Το μοντέλο δημιουργίας στίχων που προτείνουν περιλαμβάνει έναν Variational Autoencoder βασισμένο στο GPT-2. Η αρχιτεκτονική του μοντέλου μοιάζει με ένα Transformer, όπου ο κωδικοποιητής είναι ο κωδικοποιητής ήχου AST και ο αποκωδικοποιητής είναι το GPT-2. Το μοντέλο λειτουργεί ως εξής: το mel-spectorgram της μουσικής συνοδείας 5 δευτερολέπτων εισάγεται στον κωδικοποιητή AST, από τον οποίο εξάγεται η λανθάνουσα αναπαράσταση H_{music}. Αυτή η αναπαράσταση μετασχηματίζεται σε δύο διανύσματα που περιγράφουν την κατανομή της μουσικής, και υπολογίζονται σύμφωνα με τους εξής τύπους:

$$\mu = W_{\mu} H_{music}$$

$$\sigma = exp(\frac{W_{\sigma} H_{music}}{2})$$

Στη συνέχεια, πραγματοποιείται επαναπαραμετροποίηση, όπου γίνεται δειγματοληψία από την κανονική κατανομή, ώστε το λανθάνον διάνυσμα να είναι μη ντετερμινιστικό:

$$z = \mu + \sigma \odot \epsilon, \epsilon \sim \mathcal{N}(0, I)$$

Το λανθάνον διάνυσμα εισάγεται στον πολυτροπικό αποκωδικοποιητή μέσω ενός cross-attention layer. Ο αποκωδικοποιητής λαμβάνει την προηγούμενη γραμμή στίχων ως είσοδο, η οποία χρησιμοποιείται ως context για τη δημιουργία της επόμενης γραμμής στίχων. Η αρχιτεκτονική του μοντέλου φαίνεται στο Σχήμα 0.1

Το paper χρησιμοποίησε το σύνολο δεδομένων DALI, από το οποίο διατήρησαν μόνο τη μουσική συνοδεία και χρησιμοποίησαν γραμμές διάρκειας 5 δευτερολέπτων. Διατήρησαν 2590 τραγούδια, χωρίζοντάς τα σε 2072 και 518 για το training set και test set αντίστοιχα.

Ακολουθήσαμε την ίδια μέθοδο εκπαίδευσης, χρησιμοποιώντας ένα συνδυασμό του reconstruction loss και του KL divergence ως loss function, όπως περιγράφεται από την παρακάτω συνάρτηση:

$$L_{\theta}(x, y, z, \hat{y}) = L_{reconstr}(y, \hat{y}) + \beta KL(q_{\varphi}(z \mid x) \parallel p(z)),$$

όπου το $L_{reconstr}$ υπολογίζει τη διαφορά μεταξύ των παραγόμενων στίχων y και των αληθινών στίχων $\hat{y}$, το $KL(q_{\varphi}(z \mid x) \parallel p(z))$ αξιολογεί τη διαφορά μεταξύ του $p(z)$ και της κατανομής που παράγεται από τον κωδικοποιητή, και το β είναι η υπερπαράμετρος που ελέγχει τη συνεισφορά της απόκλισης KL στο loss.

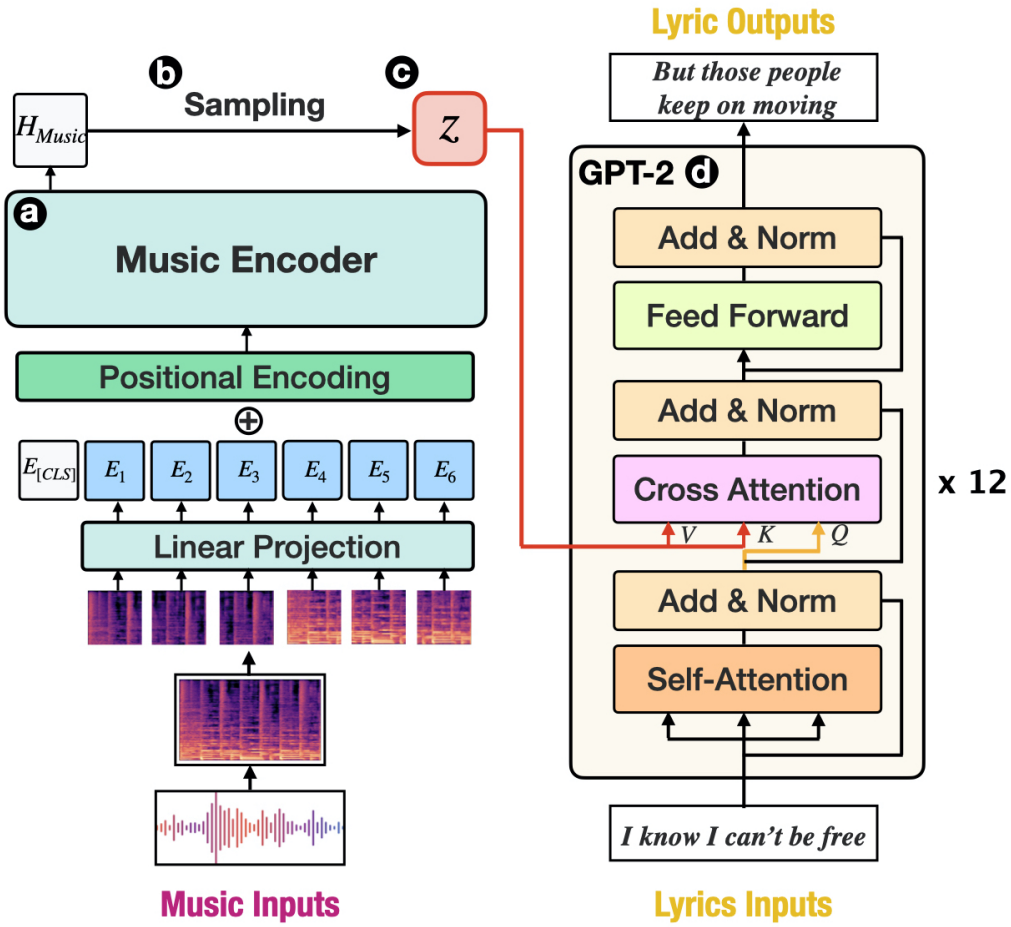
0.5.2.2 Whisper-OpenOrca

Εξετάζοντας τα αποτελέσματα του προηγούμενου μοντέλου, παρατηρήσαμε ότι η ποιότητά τους δεν ήταν επαρκής. Λόγω του περιορισμένου context στο οποίο βασιζόταν η προηγούμενη μέθοδος, οι παραγόμενοι στίχοι δεν ήταν αρκετά ουσιαστικοί και συνεκτικοί. Αυτό μας οδήγησε στην εξερεύνηση άλλων μοντέλων με πιο σύγχρονες αρχιτεκτονικές και επιπλέον προεκπαίδευση.

Σε αυτό το μοντέλο, χρησιμοποιούμε το μοντέλο Whisper ως τον audio encoder και το Mistral OpenOrca ως το LLM. Αυτά τα μοντέλα φορτώνονται με τα προεκπαιδευμένα τους βάρη, και η αντιστοίχιση των ηχητικών αναπαραστάσεων με τους στίχους γίνεται μέσω ενός projection layer, το οποίο είναι το μοναδικό τμήμα του μοντέλου που εκπαιδεύεται (η αρχιτεκτονική του μοντέλου φαίνεται στο Σχήμα 0.2). Η συνάρτηση απώλειας που χρησιμοποιείται είναι το cross-entropy loss μεταξύ των αληθινών στίχων ($\hat{y}$) και των παραγόμενων στίχων (y), με τον τύπο να δίνεται παρακάτω:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

Για να αντιμετωπίσουμε το πρόβλημα του περιορισμένου context του προηγούμενου μοντέλου, το οποίο μείωνε σημαντικά τη συνοχή των παραγόμενων στίχων, εκπαιδεύσαμε το μοντέλο δίνοντας ολόκληρο το προηγούμενο σύνολο αληθινών στίχων ως context στο prompt του LLM. Με παρόμοιο τρόπο, κατά το inference, το εκπαιδευμένο μοντέλο λαμβάνει τη



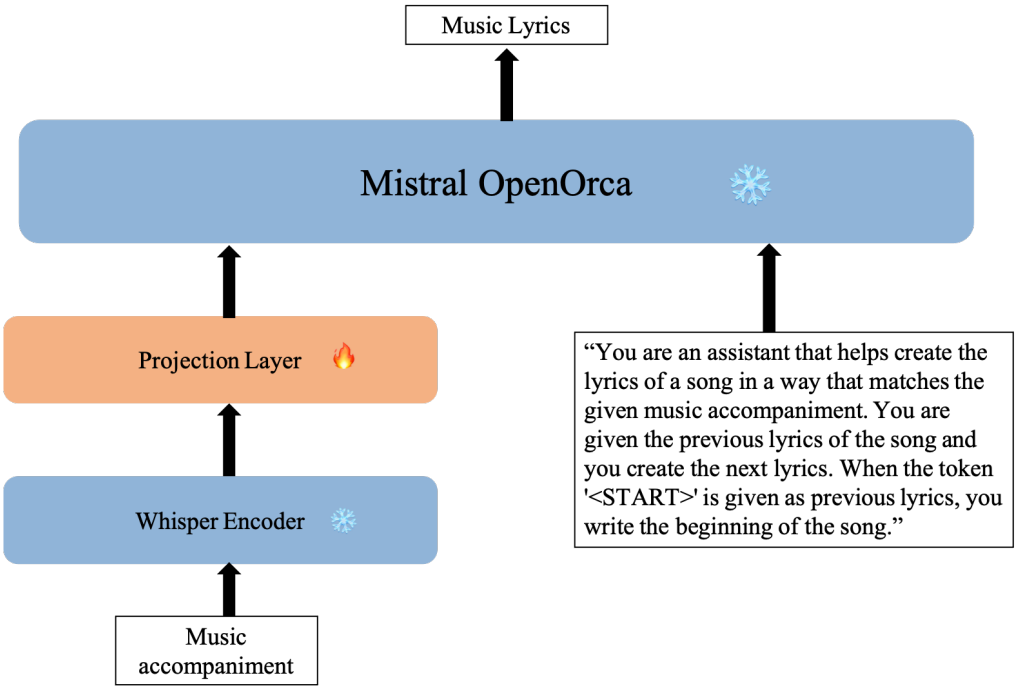
Σχήμα 0.1: *Το μοντέλο VAE AST-GPT2. Πηγή: [1]*

μουσική συνοδεία και του ζητάμε στο prompt είτε να ξεκινήσει να δημιουργεί στίχους, είτε να συνεχίσει αυτούς που έχει ήδη παράξει. Δεδομένου ότι αυτό το LLM είναι instruction μοντέλο, σε αντίθεση με το GPT-2 που είναι completion μοντέλο, μαζί με τους προηγούμενους στίχους δίνουμε και μια οδηγία στο prompt.

0.5.3 Από κείμενο & ήχο σε κείμενο σε κείμενο

Η διαδικασία δημιουργίας από ήχο σε κείμενο με επιπλέον βήμα περιλαμβάνει ένα επιπρόσθετο στάδιο δημιουργίας σε σύγκριση με την προηγούμενη μέθοδο: ενδιάμεσα παράγονται ετικέτες μουσικής (music tags), οι οποίες στη συνέχεια χρησιμοποιούνται για τη δημιουργία των τελικών στίχων του τραγουδιού. Το μοντέλο που χρησιμοποιείται σε αυτή τη διάταξη είναι το SALMONN-Claude. Με αυτό και το επόμενο μοντέλο, θέλαμε να εξετάσουμε πώς επηρεάζεται η δημιουργία στίχων όταν περνάμε από διαφορετικά modalities πριν καταλήξουμε στη διαδικασία δημιουργίας στίχων.

Με αυτό το μοντέλο, αξιοποιήσαμε τη δημιουργικότητα και την εξαιρετική επίδοση του Claude στην δημιουργική γραφή. Συγκεκριμένα, χρησιμοποιήσαμε το προεκπαιδευμένο μοντέλο SALMONN για την εξαγωγή ετικετών από τη μουσική. Χρησιμοποιήσαμε prompts που βρέθηκαν στο παράρτημα του σχετικού paper [25], καθώς είχαν καλά αποτελέσματα. Ζητήσαμε από το μοντέλο να δώσει μια λεπτομερή περιγραφή της μουσικής συνοδείας,



Σχήμα 0.2: *Το μοντέλο The Whisper-OpenOrca.*

να εξάγει το συναίσθημα της μουσικής και τα μουσικά όργανα που ακούγονται. Αυτές οι ετικέτες χρησιμοποιήθηκαν στη συνέχεια για να καθοδηγήσουν το μοντέλο Claude στη δημιουργία στίχων βασισμένων στις παραγόμενες μουσικές ετικέτες. Η αρχιτεκτονική του μοντέλου φαίνεται στο Σχήμα 0.3.

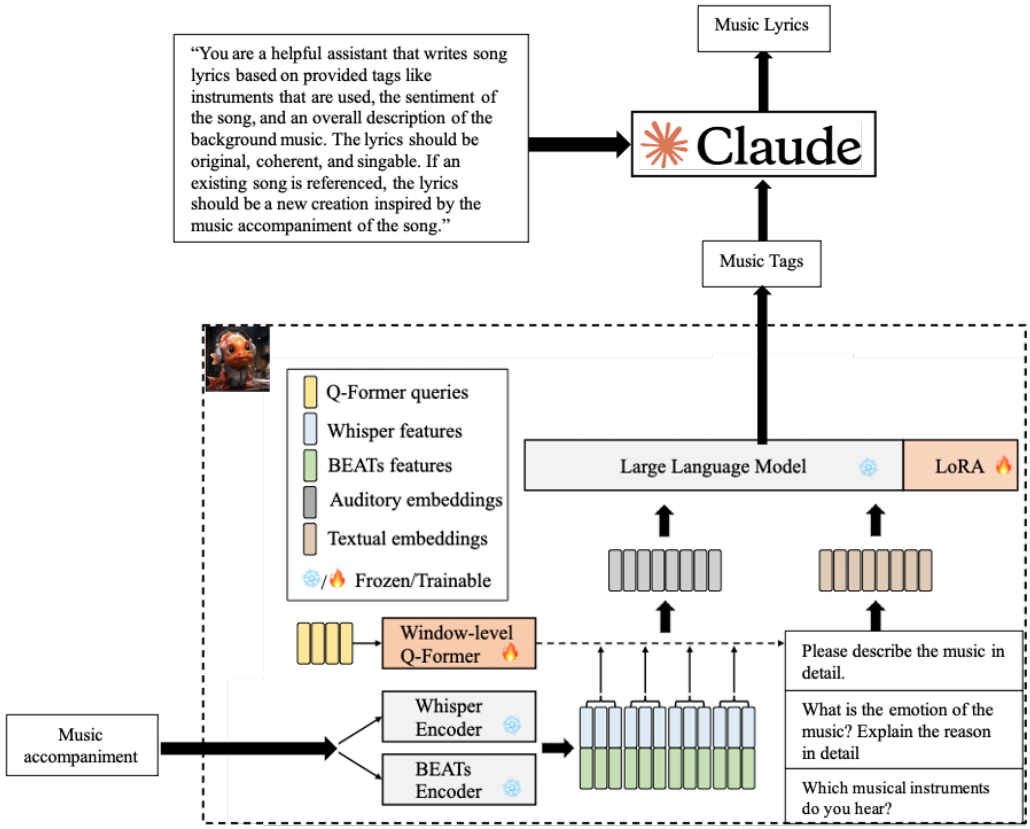
Δοκιμάσαμε επίσης αν η τεχνική few-shot learning μπορούσε να βελτιώσει ακόμα περισσότερο την ποιότητα των παραγόμενων στίχων, διαμορφώνοντας δύο διαφορετικά prompts για το μοντέλο Claude: ένα για zero-shot και ένα για few-shot.

Στην περίπτωση του few-shot prompt, δώσαμε στο μοντέλο έξι παραδείγματα, δύο από κάθε ένα από τα κυριότερα μουσικά είδη που υπάρχουν στο σύνολο δεδομένων DALI: ποπ, ροκ και εναλλακτική μουσική.

0.5.4 Από κείμενο & ήχο σε κείμενο σε εικόνα σε κείμενο

Η διαδικασία δημιουργίας από κείμενο & ήχο σε εικόνα και στη συνέχεια σε κείμενο περιλαμβάνει ένα επιπλέον βήμα σε σύγκριση με την προηγούμενη διαμόρφωση. Σε αυτό το στάδιο, δημιουργείται μια εικόνα με βάση μια περιγραφή που παράγεται από το μοντέλο κατανόησης μουσικής. Στη συνέχεια, οι στίχοι παράγονται από την εικόνα.

Με αυτό το μοντέλο, θέλαμε να δοκιμάσουμε αν η προσθήκη της τροπικότητας της όρασης (vision modality) θα μπορούσε να αυξήσει τη δημιουργικότητα στη διαδικασία δημιουργίας στίχων. Η διάταξη αυτού του μοντέλου αποτελείται από τα εξής βήματα: πρώτα, δίνουμε τη μουσική συνοδεία στο SALMONN και το προτρέπουμε να περιγράψει μια παγωμένη σκηνή ταινίας που θα μπορούσε να συνοδεύεται από τη δεδομένη μουσική. Δεδομένου ότι το μοντέλο έχει εκπαιδευτεί να ανταποκρίνεται επαρκώς σε δημιουργικές εργασίες, το αποτέλεσμα είναι ικανοποιητικό. Στη συνέχεια, το αποτέλεσμα του μοντέλου δίνεται στο Claude, το οποίο προτρέπεται να τροποποιήσει την περιγραφή έτσι ώστε να μπορεί να χρησιμοποιη-



Σχήμα 0.3: *Η διάταξη ήχος και κείμενο σε κείμενο σε κείμενο. Η διακεκομμένη γραμμή περικλείει το μοντέλο SALMONN. Προσαρμοσμένο από [2], [3] και [4]*

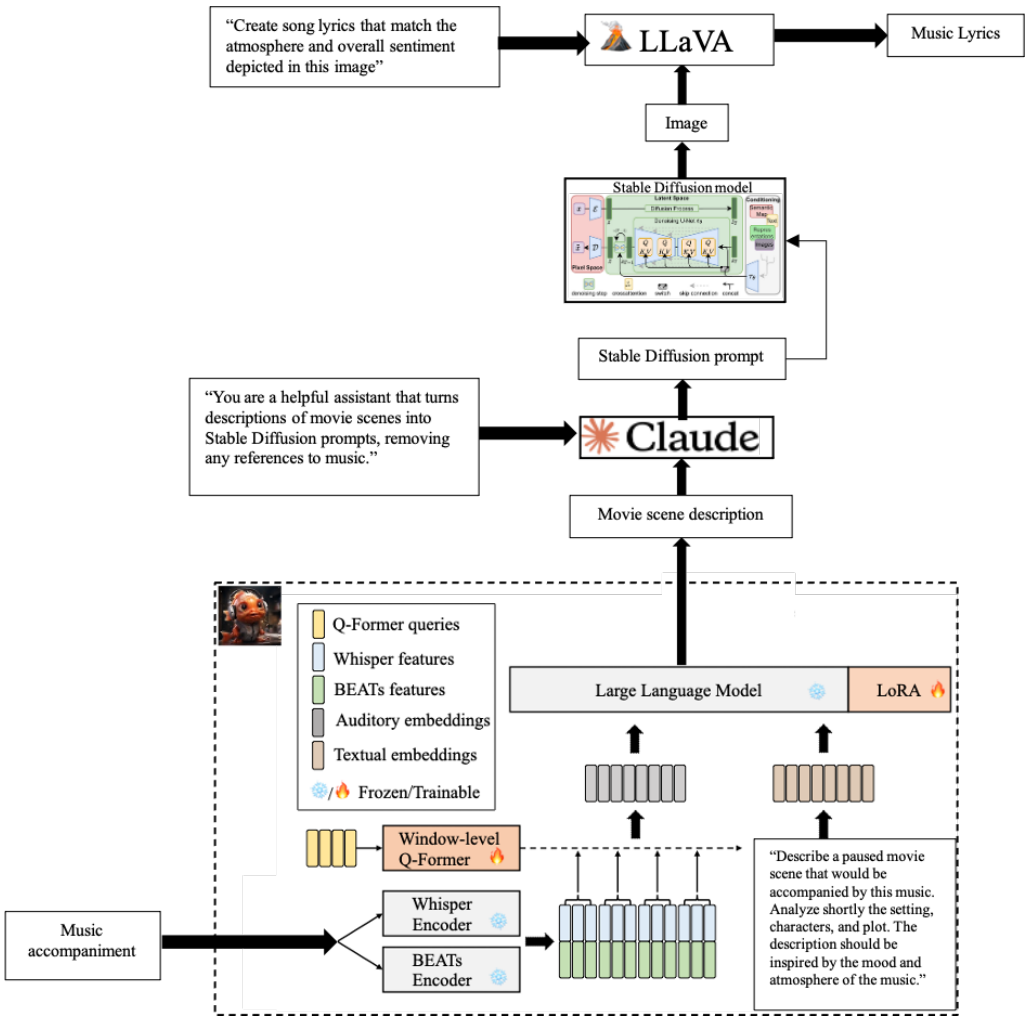
θεί ως προτροπή για το Stable Diffusion. Ακολούθως, η απάντηση του Claude δίνεται στο Stable Diffusion, το οποίο δημιουργεί μια εικόνα, και αυτή η εικόνα τελικά εισάγεται στο μοντέλο LLaVA, το οποίο προτρέπεται να δημιουργήσει στίχους βασισμένους στην εικόνα. Η αρχιτεκτονική του μοντέλου φαίνεται στο Σχήμα 0.4.

Επιπλέον, δοκιμάσαμε την απόδοση αυτής της διάταξης με τη χρήση της τεχνικής few-shot learning στο μοντέλο LLaVA, δίνοντας τα παραδείγματα τραγουδιών που δώσαμε και στην προηγούμενη διάταξη.

Επιπρόσθετα, δοκιμάσαμε μια διαμόρφωση ως ablation study για το SALMONN-Stable Diffusion-LLaVA. Δεδομένου ότι το μοντέλο LLaVA βασίζεται στο Vicuna-7b-v1, παραλείψαμε το vision modality, δημιουργώντας έτσι το μοντέλο SALMONN-Vicuna. Σε αυτή τη διάταξη, το μοντέλο SALMONN παρέχει την περιγραφή της σκηνής της ταινίας, και στη συνέχεια το μοντέλο Vicuna προτρέπεται να δημιουργήσει τους στίχους με βάση την περιγραφή που δόθηκε από το SALMONN.

0.6 Πειράματα και Αποτελέσματα

Κατά την εξαγωγή αποτελεσμάτων, αξιολογήσαμε την επίδοση των μοντέλων στο test set. Λόγω του κόστους ορισμένων υπομονάδων των δύο τελευταίων διατάξεων, η αξιολόγηση έγινε σε 100 τραγούδια. Η αξιολόγηση περιλαμβάνει, όπως αναλύεται λεπτομερέστερα στην υποενότητα 0.3.5, τον υπολογισμό της βαθμολογίας ομοιότητας με τη χρήση ενός cross-encoder,



Σχήμα 0.4: *Η διάταξη 'Ηχος και κείμενο σε κείμενο σε εικόνα σε κείμενο. Η διακεκομμένη γραμμή περικλείει το μοντέλο SALMONN. Προσαρμοσμένο από [2], [3], [4] και [5]*

και μεθόδων αξιολόγησης που βασίζονται σε LLM, όπως είναι η αξιολόγηση με το JudgeLM, και οι προτροπές σε ισχυρά και ποικίλα LLMs. Παρέχουμε περισσότερες λεπτομέρειες για κάθε μέτρηση και τα μοντέλα που χρησιμοποιήθηκαν για τις αξιολογήσεις των LLMs και, στη συνέχεια, παρουσιάζουμε τα αποτελέσματα αυτών των αξιολογήσεων. Επίσης, διεξάγουμε ένα user study για να λάβουμε τη γνώμη των χρηστών σχετικά με τα μοντέλα που πέτυχαν την καλύτερη συνολική επίδοση με τις προηγούμενες μεθόδους αξιολόγησης.

0.6.1 Σημασιολογική Κειμενική Ομοιότητα

Αναγνωρίζοντας την αδυναμία του BLEU σκορ, αποφασίσαμε να χρησιμοποιήσουμε μια μέθοδο που υπολογίζει την ομοιότητα μεταξύ των αληθινών και των παραγόμενων στίχων, χωρίς τον αυστηρό έλεγχο των επικαλύψεων n-gram. Για την επίτευξη αυτού, χρησιμοποιήσαμε ένα μοντέλο cross-encoder. Αν και οι bi-encoders παράγουν αναπαραστάσεις προτάσεων σταθερών διαστάσεων και είναι υπολογιστικά αποδοτικοί, συχνά έχουν χαμηλότερη απόδοση σε σχέση με τους cross-encoders, οι οποίοι μπορούν να αξιοποιήσουν τα επίπεδα προσοχής για να εκμεταλλευτούν τις αλληλεπιδράσεις μεταξύ προτάσεων για καλύτερη απόδοση

[49]. Το συγκεκριμένο μοντέλο που χρησιμοποιήθηκε για τον υπολογισμό των βαθμολογιών ομοιότητας είναι το ms-marco-MiniLM-L-12-v2 [50].

Η ακριβής μέθοδος υπολογισμού της βαθμολογίας ομοιότητας είναι ο υπολογισμός του μέσου όρου για όλες τις επιμέρους βαθμολογίες κάθε ζεύγους παραγόμενων-αληθινών στίχων. Η βαθμολογία ομοιότητας κυμαίνεται από 0 έως 1. Τα αποτελέσματα αυτής της μεθόδου φαίνονται στον Πίνακα 0.1.

Μοντέλο	**Σκορ ομοιότητας**
SALMONN-Stable Diffusion-LLaVA few-shot	0.89
Vicuna out-of-the-box	0.85
OpenOrca out-of-the-box	0.81
GPT-2 out-of-the-box	0.71
Claude out-of-the-box	0.69
OpenOrca finetuned	0.58
SALMONN-Claude zero-shot	0.51
GPT-2 finetuned	0.50
SALMONN-Stable Diffusion-LLaVA zero-shot	0.48
SALMONN-Vicuna	0.36
Whisper-OpenOrca	0.26
SALMONN-Claude few-shot	0.25
VAE-AST-GPT2	0.07

Πίνακας 0.1: *Τα σκορ ομοιότητας υπολογισμένα με το μοντέλο cross-encoder*

0.6.2 Χρησιμοποιώντας το LLM ως κριτή για την αξιολόγηση στίχων

Χρησιμοποιήσαμε επίσης το μοντέλο JudgeLM, το οποίο έχει αναλυθεί στην υπο-υποενότητα 0.3.5.3. Δώσαμε στο JudgeLM την πληροφορία ότι το task που έπρεπε να πραγματοποιήσουν τα μοντέλα ήταν η δημιουργία συνεκτικών και δημιουργικών στίχων. Λόγω του πλήθους των μοντέλων και των διατάξεων που δοκιμάστηκαν, δεν ήταν εφικτό να ζητήσουμε από το JudgeLM να βαθμολογήσει όλα τα αποτελέσματα ταυτόχρονα, λόγω του περιορισμού μήκους context του μοντέλου (2048). Για να το αντιμετωπίσουμε αυτό και να ενσωματώσουμε μια μέθοδο που συγκρίνει τα μοντέλα πριν τα βαθμολογήσει — καθώς η μεμονωμένη βαθμολόγηση μπορεί να μην εντοπίζει πάντα λεπτές διαφορές μεταξύ συγκεκριμένων ζευγών [43] — χρησιμοποιήσαμε το JudgeLM για βαθμολόγηση ζευγών, καταλήγοντας σε 78 συγκρίσεις. Το μοντέλο, λαμβάνοντας υπόψη και τους αληθινούς στίχους, αποδίδει μια βαθμολογία από 0 έως 10 σε κάθε μοντέλο κάθε ζεύγους, βάσει της ποιότητας των στίχων. Με αυτή τη μέθοδο, υπολογίζουμε αρχικά τη μέση βαθμολογία για κάθε ζεύγος μοντέλων. Στη συνέχεια, παρουσιάζουμε μια κατάταξη, αθροίζοντας όλες τις βαθμολογίες που έλαβε κάθε μοντέλο σε κάθε ζεύγος συγκρίσεων. Τα αποτελέσματα αυτής της αξιολόγησης φαίνονται στον Πίνακα 0.2.

0.6.3 Αξιολόγηση από LLMs με prompt

Ακολουθώντας μια μέθοδο παρόμοια με αυτή της εργασίας “Reference-Guided Verdict: LLMs-as-Judges in Automatic Evaluation of Free-Form Text” [44], όπως εξηγήθη-

Μοντέλο	Αθροισμένα Σκορ
SALMONN-Claude zero-shot	96.87
SALMONN-Claude few-shot	96.35
Claude out-of-the-box	95.91
OpenOrca out-of-the-box	92.80
SALMONN-Stable Diffusion-LLaVA zero-shot	92.53
Vicuna out-of-the-box	90.02
SALMONN-Vicuna	89.05
OpenOrca finetuned	49.76
GPT-2 finetuned	38.59
Whisper-OpenOrca	24.16
SALMONN-Stable Diffusion-LLaVA few-shot	16.26
GPT-2 out-of-the-box	15.20
VAE-AST-GPT2	13.36

Πίνακας 0.2: *Τα αθροισμένα σκορ που έδωσε το JudgeLM, σε φθίνουσα σειρά*

κε και στην υπο-υποενότητα 0.3.5.4, χρησιμοποιήσαμε τρία LLMs (ένα κλειστού κώδικα και δύο ανοιχτού κώδικα): Claude 3.5 Sonnet, Mistral-Instruct-7B-v0.3 [51], και Llama-3.1-Instruct-70B [52]. Ζητήσαμε από τα μοντέλα να αξιολογήσουν τους παραγόμενους στίχους, δίνοντας βαθμολογία από 0 έως 10 για καθένα από τα παρακάτω κριτήρια: συνοχή, δημιουργικότητα, το πόσο εύκολα μπορεί να τραγουδηθεί, και φυσικότητα. Στη συνέχεια, υπολογίστηκε προγραμματιστικά ο μέσος όρος αυτών των βαθμών, για να εξαχθεί η συνολική βαθμολογία κάθε τραγουδιού, και ακολούθως υπολογίστηκε ο μέσος όρος για να προκύψουν οι συνολικές βαθμολογίες που αποδόθηκαν από κάθε LLM-κριτή σε κάθε μοντέλο.

Τα LLMs ορίστηκαν σε χαμηλές θερμοκρασίες (0.2-0.4) για αυτή την αξιολόγηση, ώστε να αποφευχθούν πολύ απρόβλεπτες απαντήσεις. Τα τελικά αποτελέσματα αυτής της μεθόδου αξιολόγησης βρίσκονται στον Πίνακα 0.3.

Μοντέλο	Claude	LLaMA	Mistral	Μέση βαθμολογία
SALMONN-Stable Diffusion-LLaVA zero-shot	8.19	8.31	8.50	8.33
SALMONN-Claude zero-shot	8.43	8.15	8.20	8.26
Claude out-of-the-box	8.31	8.08	8.30	8.23
OpenOrca out-of-the-box	7.87	7.89	8.61	8.12
SALMONN-Claude few-shot	8.26	7.80	8.21	8.09
Vicuna out-of-the-box	7.68	7.63	8.54	7.95
SALMONN-Vicuna	7.54	7.34	8.29	7.72
OpenOrca finetuned	5.89	6.74	7.31	6.64
Whisper-OpenOrca	4.40	5.27	6.38	5.35
GPT-2 finetuned	3.66	5.56	6.65	5.29
SALMONN-Stable Diffusion-LLaVA few-shot	2.96	2.91	6.34	4.07
GPT-2 out-of-the-box	1.83	1.36	5.55	2.91
VAE-AST-GPT2	0.35	0.87	1.53	0.92

Πίνακας 0.3: *Οι αξιολογήσεις από τα LLMs, σε φθίνουσα σειρά των μέσων βαθμολογιών*

0.6.4 Μελέτη Χρηστών: Κατάταξη των Μεθόδων μας με Ανθρώπινες Αξιολογήσεις

Λόγω της δημιουργικής φύσης της παραγωγής στίχων, θεωρούμε σημαντικό να ενσωματωθεί η ανθρώπινη κρίση μαζί με τις αυτόματες μεθόδους αξιολόγησης. Αν και η μελέτη χρηστών μας διεξήχθη σε μικρότερη κλίμακα σε σύγκριση με άλλες τεχνικές αξιολόγησης, παρέχει πολύτιμες πληροφορίες σχετικά με την αντιληπτή ποιότητα και δημιουργικότητα των παραγόμενων στίχων από την προοπτική των χρηστών.

Η μελέτη δομήθηκε ως εξής: Οι συμμετέχοντες άκουσαν πρώτα 1 λεπτό από τη μουσική συνοδεία για κάθε τραγούδι. Στη συνέχεια, τους παρουσιάστηκαν δύο υποψήφια σύνολα στίχων και τους ζητήθηκε να επιλέξουν, πρώτα ως προς την ποιότητα των στίχων (συνεκτικότητα, δομή) και ύστερα ως προς την συσχέτισή τους με την μουσική. Η μελέτη περιλάμβανε συνολικά 12 τραγούδια και εξετάσαμε μόνο τα μοντέλα που είχαν τα καλύτερα αποτελέσματα με τις άλλες μεθόδους, για κάθε συνδυασμό από modalities, που είναι: Claude out-of-the-box, Whisper-OpenOrca, SALMONN-Claude zero-shot, και SALMONN-Stable Diffusion-LLaVA zero-shot. Κάθε τραγούδι είχε έξι πιθανά ζεύγη υποψηφίων στίχων για σύγκριση.

Για να αποτρέψουμε την κόπωση των συμμετεχόντων και να διατηρήσουμε το ενδιαφέρον τους, εφαρμόσαμε τυχαία ανάθεση των ζευγών αξιολόγησης. Κάθε συμμετέχοντας βαθμολόγησε 12 ζεύγη συνολικά, πράγμα που σημαίνει ότι η μελέτη χρηστών είχε 6 ομάδες ερωτήσεων που ανατίθεντο τυχαία κάθε φορά που κάποιος άνοιγε την φόρμα. Αυτή η τυχαία ανάθεση εξασφάλιζε επιπλέον ότι κάθε συμμετέχοντας αξιολογούσε δύο φορές κάθε ζεύγος μοντέλων και ένα ζεύγος μοντέλων για κάθε τραγούδι, διατηρώντας μια ισορροπία μεταξύ της πλήρους κάλυψης και ενός διαχειρίσιμου φόρτου για τους αξιολογητές.

Η μέθοδος που χρησιμοποιήθηκε για την κατάταξη των μοντέλων από τη μελέτη χρηστών είναι το μοντέλο Bradley-Terry [53], το οποίο είναι ένα μοντέλο πιθανοτήτων που χρησιμοποιείται για την πρόβλεψη του αποτελέσματος συγκρίσεων ανά ζεύγη και έχει ευρεία εφαρμογή σε διάφορους τομείς, συμπεριλαμβανομένης της κατάταξης μοντέλων τεχνητής νοημοσύνης [54]. Στο μοντέλο Bradley-Terry, σε κάθε υποψήφιο ανατίθεται μια παράμετρος ισχύος π_i, για τον υποψήφιο i. Το μοντέλο υποθέτει ότι η πιθανότητα ο υποψήφιος i να κερδίσει τον υποψήφιο j σε σύγκριση ανά ζεύγη είναι:

$$P(i \, beats \, j) = \frac{\pi_i}{\pi_i + \pi_j}.$$

Για την κατάταξη των υποψηφίων, το μοντέλο Bradley-Terry χρησιμοποιεί τη εκτίμηση μέγιστης πιθανοφάνειας (MLE) για να βρει τις καλύτερες τιμές για κάθε π_i, που μεγιστοποιούν την πιθανότητα να παρατηρηθούν τα δεδομένα των συγκρίσεων. Ο υπολογισμός των π_i απαιτεί την επίλυση της ακόλουθης εξίσωσης για κάθε υποψήφιο i:

$$log(\pi_i) = \sum_{j \neq i} \frac{W_{ij}}{\pi_i + \pi_j},$$

όπου W_{ij} είναι ο αριθμός των φορών που ο υποψήφιος i νίκησε τον υποψήφιο j. Μετά την εκτίμηση των τιμών π_i, οι υποψήφιοι με τις υψηλότερες τιμές π κατατάσσονται στην κορυφή.

Τα αποτελέσματα, τα οποία προέκυψαν από τη συμμετοχή 28 χρηστών, φαίνονται στον

Πίνακα 0.4 και στον Πίνακα 0.5.

Μοντέλο	Πιθανότητες Bradley-Terry για συνοχή και δομή
SALMONN-Stable Diffusion-LLaVA zero-shot	0.316393
Claude out-of-the-box	0.304484
SALMONN-Claude zero-shot	0.282116
Whisper-OpenOrca	0.097007

Πίνακας 0.4: *Οι πιθανότητες Bradley-Terry για τη συνοχή και τη δομή των στίχων, σε φθίνουσα σειρά*

Μοντέλο	Πιθανότητες Bradley-Terry για συσχέτιση με την μουσική
SALMONN-Claude zero-shot	0.355462
SALMONN-Stable Diffusion-LLaVA zero-shot	0.241730
Claude out-of-the-box	0.228919
Whisper-OpenOrca	0.173889

Πίνακας 0.5: *Οι πιθανότητες Bradley-Terry για την συσχέτιση των στίχων με τη μουσική, σε φθίνουσα σειρά*

Εκτελέσαμε επίσης στατιστική ανάλυση των αποτελεσμάτων της μελέτης χρηστών, χρησιμοποιώντας το z-test και υπολογίζοντας τα p-values για κάθε ζεύγος μοντέλων. Το Z-score μετρά πόσο απέχει το παρατηρούμενο ποσοστό νίκης από το αναμενόμενο ποσοστό υπό τη μηδενική υπόθεση. Η μηδενική υπόθεση υποθέτει ότι και τα δύο μοντέλα είναι εξίσου πιθανό να κερδίσουν (δηλαδή, 50-50 κατανομή των νικών). Όταν το μέγεθος των Z-scores είναι κοντά στο μηδέν (περίπου 0 έως ±1), υποδηλώνουν ότι τα αποτελέσματα δεν απέχουν πολύ από τη μηδενική υπόθεση, ενώ τα z-scores που απέχουν περισσότερο από το μηδέν (π.χ., μεγαλύτερα από ±1.96 για επίπεδο εμπιστοσύνης 95% και μεγαλύτερα από ±1.645 για επίπεδο εμπιστοσύνης 90%) υποδηλώνουν στατιστικά σημαντική απόκλιση από τη μηδενική υπόθεση [55]. Αναφορικά με την τιμή p-value, μια μικρή τιμή p-value (συνήθως $p < 0.05$) υποδηλώνει ότι η παρατηρούμενη διαφορά στις νίκες είναι στατιστικά σημαντική, που σημαίνει ότι είναι απίθανο να έχει προκύψει τυχαία, και θα μπορούσαμε να απορρίψουμε τη μηδενική υπόθεση (δηλ. τα μοντέλα δεν προτιμώνται εξίσου). Μια μεγάλη τιμή p-value ($p > 0.05$) υποδηλώνει ότι η παρατηρούμενη διαφορά στις νίκες δεν είναι στατιστικά σημαντική, πράγμα που σημαίνει ότι δεν μπορούμε να απορρίψουμε τη μηδενική υπόθεση (δηλαδή, τα μοντέλα προτιμώνται εξίσου) [56]. Η στατιστική ανάλυση των αποτελεσμάτων (z-test και p-values) για κάθε ζεύγος μοντέλων φαίνονται στον Πίνακα 0.6 και στον Πίνακα 0.7

Ζεύγος Μοντέλων	Z-score (απόλυτη τιμή)	p-value
Claude-out-of-the-box & Whisper-OpenOrca	3.21	0.00134
Claude-out-of-the-box & SALMONN-Claude	1.07	0.285
Claude-out-of-the-box & SALMONN-Stable Diffusion-LLaVA	0.27	0.789
Whisper-OpenOrca & SALMONN-Claude	4.81	1.50ε-06
Whisper-OpenOrca & SALMONN-Stable Diffusion-LLaVA	3.47	0.00051
SALMONN-Claude & SALMONN-Stable Diffusion-LLaVA	0.80	0.423

Πίνακας 0.6: *Τα z-test και p-values για κάθε ζεύγος μοντέλων, για το κριτήριο της δομής/συνοχής*

Με αυτή τη στατιστική ανάλυση, μπορούμε να πούμε ότι, όσον αφορά το κριτήριο της δο-

Ζεύγος Μοντέλων	Z-score (απόλυτη τιμή)	p-value
Claude-out-of-the-box & Whisper-OpenOrca	1.60	0.109
Claude-out-of-the-box & SALMONN-Claude	1.87	0.061
Claude-out-of-the-box & SALMONN-Stable Diffusion-LLaVA	0.53	0.593
Whisper-OpenOrca & SALMONN-Claude	2.67	0.0075
Whisper-OpenOrca & SALMONN-Stable Diffusion-LLaVA	0.53	0.593
SALMONN-Claude & SALMONN-Stable Diffusion-LLaVA	1.07	0.285

Πίνακας 0.7: *Τα z-test και p-values για κάθε ζεύγος μοντέλων, για το κριτήριο της συσχέτισης μουσικής-στίχων*

μής/συνεκτικότητας, το Whisper-OpenOrca φαίνεται να έχει χειρότερες επιδόσεις σε σύγκριση με τα άλλα μοντέλα, ιδίως σε σύγκριση με το SALMONN-Claude και το SALMONN-Stable Diffusion-LLaVA. Από την άλλη, τα μοντέλα SALMONN-Claude και SALMONN-Stable Diffusion-LLaVA έχουν παρόμοιες επιδόσεις. Για το κριτήριο της συσχέτισης μουσικής-στίχων, το SALMONN-Claude φαίνεται να υπερέχει σημαντικά έναντι του Whisper-OpenOrca. Η σύγκριση μεταξύ του Claude out-of-the-box και του SALMONN-Claude έχει στατιστική σημαντικότητα με 90% - αντί για 95% - εμπιστοσύνη, υποδεικνύοντας ότι το SALMONN-Claude υπερτερεί ελαφρώς του Claude out-of-the-box. Τα άλλα ζεύγη παρουσιάζουν παρόμοιες επιδόσεις.

0.7 Συμπεράσματα

0.7.1 Συζήτηση

Αυτή η διπλωματική εργασία εξερευνά τη διαδικασία αυτόματης δημιουργίας στίχων και πώς επηρεάζεται η ποιότητα των παραγόμενων στίχων με την προσθήκη διαφορετικών modalities. Από όσο γνωρίζουμε, η ενσωμάτωση της τροπικότητας της όρασης (vision modality), μαζί με το κείμενο και τον ήχο, δεν έχει μελετηθεί προηγουμένως.

Διερευνήσαμε τέσσερις διαφορετικούς συνδυασμούς δεδομένων. Ο πρώτος ήταν η μετατροπή από κείμενο σε κείμενο, όπου δοκιμάσαμε τρία μεγάλα γλωσσικά μοντέλα (LLMs) για τη δημιουργία στίχων. Αρχικά τα χρησιμοποιήσαμε χωρίς προσαρμογή (out-of-the-box) και στη συνέχεια τα προσαρμόσαμε στο σύνολο δεδομένων DALI. Διαπιστώθηκε ότι τα LLMs που δέχονται οδηγίες (instruction-based LLMs) είχαν καλή επίδοση ακόμα και χωρίς προσαρμογή, λόγω της εκτενούς προεκπαίδευσής τους σε κείμενα δημιουργικής γραφής, όπως τραγούδια και ποιήματα. Στη μελέτη χρηστών, το επιλεγμένο LLM για την αξιολόγηση του συνδυασμού 'από κείμενο σε κείμενο' ήταν το Claude, το οποίο παρουσίασε την καλύτερη επίδοση στις αξιολογήσεις των LLM. Αυτό το μοντέλο είχε αρκετά καλές επιδόσεις και στα δύο κριτήρια της μελέτης χρηστών, αλλά δεν υπερείχε ιδιαίτερα σε κανένα από τα δύο.

Ο δεύτερος συνδυασμός που εξερευνήσαμε ήταν η προσθήκη μουσικής επιτήρησης (κείμενο και ήχος σε κείμενο). Δοκιμάσαμε δύο μοντέλα: το πρώτο βασίστηκε στην αναπαραγωγή μιας προηγούμενης εργασίας που χρησιμοποίησε ένα μοντέλο variational autoencoder με αρχιτεκτονική τύπου transformer, που ενσωματώνει τον Audio Spectrogram Transformer ως κωδικοποιητή ήχου και το GPT-2 ως αποκωδικοποιητή. Το μοντέλο αυτό, που προσπαθήσαμε να αναπαράγουμε όσο το δυνατόν πιο πιστά, αποδείχθηκε το πιο αδύναμο από όλα

τα μοντέλα που μελετήσαμε. Αυτό το αποδίδουμε στην έλλειψη κώδικα και στη μη διαθεσιμότητα της διαδικασίας προεπεξεργασίας των δεδομένων που χρησιμοποίησαν, καθώς και στην παλαιότητα του GPT-2 σε σύγκριση με τα άλλα LLMs. Δεδομένου ότι η ποιότητα των στίχων εξαρτάται κυρίως από την επίδοση του LLM, η χρήση πιο αδύναμων μοντέλων οδηγεί σε χαμηλότερη ποιότητα στίχων.

Η άλλη διαμόρφωση που δοκιμάστηκε για τον συνδυασμό κειμένου και ήχου σε κείμενο ήταν το μοντέλο Whisper-OpenOrca, το οποίο χρησιμοποιεί τον κωδικοποιητή ήχου Whisper, το LLM Mistral OpenOrca, και ένα εκπαιδεύσιμο projection layer που ευθυγραμμίζει τις αναπαραστάσεις μουσικής με τη δημιουργία στίχων. Αυτό το μοντέλο είχε πολύ καλύτερη επίδοση, κάτι που οφείλεται στο γεγονός ότι τόσο ο κωδικοποιητής ήχου όσο και το LLM είναι πιο σύγχρονα και ισχυρά μοντέλα, εκμεταλλευόμενα την πλούσια προεκπαίδευσή τους. Στη μελέτη χρηστών, το μοντέλο Whisper-OpenOrca παρουσίασε χαμηλότερες επιδόσεις και στα δύο κριτήρια σε σύγκριση με τα άλλα μοντέλα που συμπεριλήφθηκαν στη μελέτη, γεγονός που αντικατοπτρίζει τις αξιολογήσεις των LLM.

Ο τρίτος συνδυασμός πρόσθεσε ένα ενδιάμεσο στάδιο παραγωγής κειμένου, συγκριτικά με το προηγούμενο (κείμενο & ήχος σε κείμενο). Αυτός ο συνδυασμός δοκιμάστηκε με το μοντέλο SALMONN-Claude, όπου εξάγαμε ετικέτες μουσικής από τον ήχο και τις δώσαμε στο μοντέλο Claude για να δημιουργήσει στίχους. Αυτή η προσθήκη έδειξε ελαφρώς καλύτερη επίδοση από το σενάριο κείμενο σε κείμενο, και σημαντικά καλύτερη από τους άλλους συνδυασμούς, αποδεικνύοντας ότι αυτή η ροή εργασίας ενισχύει τη διαδικασία δημιουργίας στίχων. Από τη μελέτη χρηστών, παρατηρούμε επίσης ότι αυτό το pipeline πέτυχε την καλύτερη βαθμολογία όσον αφορά τη συσχέτιση στίχων και μουσικής, γεγονός που ενισχύει περαιτέρω τον ισχυρισμό ότι η προσθήκη του βήματος της εξαγωγής μουσικών ετικετών ενισχύει τη διαδικασία παραγωγής στίχων.

Ο τέταρτος συνδυασμός πρόσθεσε την οπτική διάσταση στην προηγούμενη διαμόρφωση (κείμενο & ήχος σε εικόνα και στη συνέχεια σε κείμενο). Αυτό διερευνήθηκε με το μοντέλο SALMONN-Stable Diffusion-LLaVA, όπου το μοντέλο SALMONN δημιούργησε μια περιγραφή σκηνής ταινίας που θα μπορούσε να συνοδευτεί από τη δεδομένη μουσική. Στη συνέχεια, η περιγραφή μετατράπηκε με τη βοήθεια του Claude σε prompt για το Stable Diffusion, το οποίο παρήγαγε μια εικόνα. Η εικόνα δόθηκε στο μοντέλο LLaVA για να δημιουργήσει τους τελικούς στίχους. Η προσθήκη του vision modality αύξησε την ποιότητα των στίχων και η απόδοσή της ήταν εφάμιλλη με το προηγούμενο μοντέλο, αποδεικνύοντας ότι τα επιπλέον modalities μπορούν να ενισχύσουν τη δημιουργικότητα στη διαδικασία δημιουργίας στίχων. Επιπλέον, η μελέτη χρηστών έδειξε ότι το μοντέλο SALMONN-Stable Diffusion-LLaVA σημείωσε ελαφρώς καλύτερη βαθμολογία από τα άλλα μοντέλα όσον αφορά τη συνοχή και τη δομή των στίχων, διατηρώντας παράλληλα μια λογική συσχέτιση με τη μουσική, επιτυγχάνοντας μια ισορροπία μεταξύ αυτών των δύο κριτηρίων.

Τέλος, δοκιμάσαμε την τεχνική few-shot prompting για δύο από τα παραπάνω σενάρια. Αν και σε ορισμένες περιπτώσεις αυξήθηκε το σκορ ομοιότητας, παρατηρήθηκε μείωση στη δημιουργικότητα των στίχων, καθώς τα μοντέλα καθοδηγούνταν προς παρόμοιους στίχους με τα παραδείγματα.

Τα συμπεράσματα αυτά προέκυψαν κυρίως από τις αξιολογήσεις των LLM. Δεδομένου ότι, ακόμη και οι ανθρώπινοι συνθέτες θα επινοούσαν διαφορετικούς στίχους για μια δε-

δομένη μουσική ή θα συνέχιζαν διαφορετικά ένα τραγούδι δεδομένης της πρώτης γραμμής του, θεωρούμε ότι οι βαθμολογίες ομοιότητας και οι αντικειμενικές μετρικές είναι λιγότερο σημαντικές από τις αξιολογήσεις LLM και τη μελέτη χρηστών, οι οποίες δίνουν μια καλύτερη εικόνα της ποιότητας των στίχων. Δεδομένου ότι χρησιμοποιήσαμε μια μέθοδο που μετριάζει τη μεροληψία των LLM για τις αξιολογήσεις των LLM, εμπιστευόμαστε περισσότερο αυτά τα αποτελέσματα. Επιπλέον, η μελέτη χρηστών έδειξε συσχέτιση μεταξύ των αξιολογήσεων LLM και των ανθρώπινων κρίσεων, και συγκεκριμένα επιβεβαιώνοντας ότι το μοντέλο Whisper-Mistral είναι σημαντικά χειρότερο, ενώ τα άλλα μοντέλα της μελέτης χρηστών έχουν παρόμοιες επιδόσεις. Αυτό υποδηλώνει ότι οι αξιολογήσεις LLM χρησιμεύουν ως αξιόπιστο μέσο για την αξιολόγηση της ποιότητας των στίχων.

Συμπερασματικά, η διπλωματική εργασία αυτή συμβάλλει στην εξέλιξη των generative tasks στον τομέα της ανάκτησης μουσικής πληροφορίας, φέρνοντας νέες ιδέες και δυνατότητες στο πεδίο της αυτόματης δημιουργίας στίχων.

0.7.2 Περιορισμοί και Μελλοντικές Κατευθύνσεις

Η έρευνά μας περιορίστηκε από τους διαθέσιμους υπολογιστικούς πόρους. Για το μεγαλύτερο μέρος της εργασίας μας, βασιστήκαμε σε δωρεάν GPU πόρους από πλατφόρμες όπως το Google Colab και το Kaggle, καθώς και στον σέρβερ του εργαστηρίου SLP-NTUA, ο οποίος ήταν εξοπλισμένος με δύο GPU των 12GB (NVIDIA GeForce GTX 1080 Ti και GeForce GTX TITAN X). Για πιο απαιτητικά μοντέλα και ισχυρότερα LLMs, χρησιμοποιήσαμε επιλεκτικά on-demand πόρους από την AWS.

Αυτοί οι περιορισμοί επηρέασαν την επίδοση των εκπαιδευμένων μοντέλων μας με διάφορους τρόπους. Συγκεκριμένα, οι μέθοδοι εκπαίδευσης ήταν περιορισμένες λόγω των υπολογιστικών περιορισμών. Κυρίως χρησιμοποιήσαμε ελαφρύτερα μοντέλα και το εύρος των πειραμάτων με μεγαλύτερα, πιο απαιτητικά μοντέλα ήταν περιορισμένο. Η χρήση ισχυρότερων μοντέλων ίσως να είχε βελτιώσει περαιτέρω την ποιότητα των αποτελεσμάτων. Άλλα μοντέλα κατανόησης μουσικής, όπως το MU-LLaMA [26], το M^2UGen [27] και το MusiLingo [10], δεν μπορούσαν να χρησιμοποιηθούν στο τρίτο σύνολο συνδυασμών λόγω της υψηλής υπολογιστικής τους απαίτησης.

Ένας ακόμη περιορισμός, που δεν αφορά μόνο τη δική μας εργασία αλλά και τον τομέα της δημιουργίας στίχων γενικότερα, είναι η περιορισμένη διαθεσιμότητα συνόλων δεδομένων. Υπάρχουν πολλά σύνολα μουσικών δεδομένων με ετικέτες και μεταδεδομένα (όπως το Million Song Dataset [57]), αλλά όχι με συγχρονισμένους στίχους. Το σύνολο δεδομένων DALI είναι, από όσο γνωρίζουμε, το μόνο διαθέσιμο σύνολο αυτού του μεγέθους, με τραγούδια σε μορφή MP3 και ευθυγραμμισμένους στίχους. Άλλα διαθέσιμα ευθυγραμμισμένα σύνολα δεδομένων, όπως το Lakh MIDI dataset [58], χρησιμοποιούν αρχεία MIDI αντί για MP3. Επιπλέον, η πλειοψηφία των εγγραφών στο σύνολο δεδομένων DALI είναι ποπ τραγούδια, γεγονός που κάνει τα εκπαιδευμένα μοντέλα να περιορίζουν την έξοδό τους σε θέματα που σχετίζονται με την ποπ μουσική, όπως η αγάπη, κάτι που περιορίζει τη δημιουργικότητα και την ποικιλία των παραγόμενων στίχων.

Στην μελλοντική έρευνα, πιστεύουμε ότι θα ήταν χρήσιμο να εξερευνηθεί η ενσωμάτωση των άλλων μοντέλων κατανόησης μουσικής που αναφέρθηκαν, καθώς και άλλων ισχυρών

LLMs στη διαδικασία δημιουργίας στίχων. Επίσης, η προσαρμογή της υπομονάδας παραγωγής στίχων στη διαμόρφωση κείμενο & ήχος σε κείμενο σε εικόνα σε κείμενο θα ήταν ένα ενδιαφέρον πεδίο μελέτης, ή ακόμα και η ενσωμάτωση ενός Vision Language Model με αρχιτεκτονική που υποστηρίζει few-shot learning με εικόνες, όπως το μοντέλο Flamingo [59]. Επιπλέον, μια άλλη προέκταση της μελέτης μας θα ήταν η δημιουργία στίχων βασισμένη σε καλλιτέχνη ή μουσικό είδος, και η διερεύνηση της διαφοροποίησης των εξόδων, με δεδομένη την ίδια μουσική συνοδεία αλλά διαφορετικό καλλιτέχνη ή είδος. Τέλος, μια άλλη προέκταση αυτής της διπλωματικής εργασίας θα μπορούσε να είναι η εκπαίδευση ενός μοντέλου για την παραγωγή μελωδίας που να ταιριάζει με τους παραγόμενους στίχους, παρόμοια με την εργασία “Lyrics and Vocal Melody Generation conditioned on Accompaniment” [60], με τη διαφορά ότι αυτή η εργασία έγινε με συμβολική μουσική (MIDI) αντί για ήχο.

Κεφάλαιο 1

Introduction

1.1 Motivation

Music and singing have been integral to human culture throughout history, serving not just as entertainment but as powerful vehicles for emotional expression, cultural preservation, and social connection. For millennia, the craft of songwriting has relied on the delicate interplay between music and lyrics, where the words must not only convey meaning but also harmonize with the musical accompaniment in terms of rhythm, emotion, and style. For a song to be truly engaging, the lyrics must achieve multiple objectives simultaneously: they must harmonize with the music's mood and tempo, maintain narrative coherence, follow linguistic patterns, and possess creative originality—a complex set of criteria that are challenging both to model computationally and to evaluate objectively.

This complexity has resulted in a noticeable research gap in the field of automated lyric generation, particularly in scenarios where music serves as the primary input. While recent years have seen remarkable advances in artificial intelligence, with breakthroughs in multimodal learning and large language models (LLMs) transforming various creative domains, the specific challenges of aligning music with lyrics remain underexplored. This gap is particularly striking given the successful applications of AI in related tasks such as music generation, melody-to-lyric alignment, and pure text generation.

The challenge lies not only in the technical aspects of processing musical input but also in capturing the subtle relationships between musical elements and lyrical content. Musical features such as tempo, key, instrumentation, and emotional tone all influence lyrical choices in ways that human songwriters intuitively understand but that prove difficult to systematize. Additionally, while LLMs have demonstrated impressive capabilities in creative writing tasks, their ability to generate contextually appropriate lyrics that maintain both musical and narrative coherence represents a unique challenge.

This thesis aims to bridge these gaps by leveraging recent advancements in AI to enhance the process of lyric generation. By developing and evaluating models that can understand musical input and generate appropriate lyrics, we seek to contribute to both the artificial intelligence and music information retrieval fields. These models have practical applications beyond academic interest, potentially providing songwriters and musicians with innovative tools that can aid in the creative process, suggest lyrical directions, and help overcome creative blocks. Furthermore, this research contributes to our un-

derstanding of the relationship between music and language, potentially offering insights into how humans process and create multimodal artistic content.

1.2 Contribution

This thesis contributes to the MIR field regarding lyric generation. Given that the specific subject of the thesis - lyric generation given the music accompaniment - hasn't been as extensively studied as other similar generation tasks, such as music generation from lyrics, and lyric generation given the melody of the song, this thesis explores several architectures and their performance regarding the lyric generation task. Firstly, we explore the lyric generation without any music supervision, by testing the performance of the LLMs Claude, GPT-2, Mistral OpenOrca and Vicuna on the lyrics generation task, by providing the first line of the song. Then, we explore the effect of music supervision on the lyric generation process, by testing two architectures, the variational autoencoder AST GPT-2 model, and the Whisper-OpenOrca model. Specifically, the exploration of the variational autoencoder AST GPT-2 architecture constitutes a reproducibility study of the paper MusicJam [1], given that the authors do not provide any code or the checkpoint weights of the model presented in their work. The Whisper-OpenOrca model uses a technique that has recently been used in other music understanding models, but not with a focus on lyric generation [61],[10]. Finally, with the last two pipelines, which, to our knowledge, have not been already used in the literature, we introduce a novel approach to lyric generation by adding additional modalities to the lyric generation process.

More specifically, the main contributions of this thesis are:

- First comprehensive comparative study of multiple LLM architectures (Claude, GPT-2, Mistral OpenOrca, and Vicuna) for text-to-text lyric generation.
- Implementation and evaluation of a music-supervised lyric generation architecture, which has been used for other tasks but without a focus in lyric generation: the Whisper-OpenOrca model, which employs a projection layer to align music and text representations between the pre-trained frozen audio encoder Whisper and the frozen LLM OpenOrca.
- First experimental study comparing different multimodal approaches for lyric generation.
- Development of novel multimodal pipelines that incorporate additional modalities into the lyric generation process:
  - SALMONN-Claude: a pipeline that uses SALMONN to extract music tags from the music input, and then uses Claude to generate the final lyrics based on these tags.
  - SALMONN-Stable Diffusion-LLaVA: a pipeline that uses SALMONN to generate a movie scene description from the music input, then uses Stable Diffusion to generate an image based on that description, and generates the final lyrics using the LLaVA model.

1.3 Thesis Structure

In Chapters 2 and 3, we present the theoretical background, and analyze the specific models relevant to this thesis. In particular:

- In Chapter 2, we explore several domains of machine learning that are relevant to our work, by analyzing multimodal models, large language models, the variational autoencoder, and Stable Diffusion.
- In Chapter 3, we analyze models and methods more specifically related to the task of multimodal lyric generation, such as audio encoders, audio language models, music understanding models, specific large language models relevant to our work, vision language models, and evaluation methods for generative tasks.

In Chapters 4 through 6, we present our work regarding this task. More specifically:

- In Chapter 4, we explain the methodology of our work, by analyzing the preprocessing of the dataset that we used, and the models that we tested for our task.
- In Chapter 5, we present the experiments that we conducted, by specifying the training details and hyperparameters. We explain further our evaluation methods and present the results of the tested models.
- In Chapter 6, we talk about the observed results, and the possible future work regarding this domain.

Κεφάλαιο 2

Machine Learning

2.1 Audio Multimodal Models: An Overview

Multimodal models represent a class of machine learning systems designed to process and integrate data from multiple modalities, such as text, audio, images, and videos. Recent advancements in artificial intelligence, particularly with the development of Large Language Models (LLMs), have brought multimodal capabilities to new heights. Multimodal models possess the ability to process different types of input (e.g., images and text) and generate contextually meaningful outputs based on them, unlocking unprecedented potential for tasks such as visual storytelling, captioning, and complex reasoning across multiple data types. Audio multimodal models specialize in integrating audio data, such as speech or environmental sounds, with other modalities like text and images. These models are crucial in applications like automatic speech recognition (ASR), voice assistants, and multimedia understanding systems, where audio information complements other data types to provide richer context and understanding.

2.1.1 Overview and Architecture

The architecture of audio multimodal models typically consists of three core components: a pre-trained audio encoder, a language model (LLM), and a modality interface [18]. The audio encoder processes raw audio signals, transforming them into a more compact representation that the LLM can reason over. The most widely used audio encoders include models such as CLAP (Contrastive Language-Audio Pretraining), which are trained on large-scale audio-text pairs to create representations aligned with text and audio.

Once the audio data is encoded, it is passed to the modality interface, which connects the encoded audio to the LLM. This interface serves as a bridge, ensuring that the audio information can be understood and processed by the language model. In some models, the interface may transform the audio features into token-based representations that can be concatenated with textual data and processed in the same context as language tokens.

2.1.2 Applications

Audio multimodal models have been extensively used in several domains. They are central to improving speech-to-text systems, where audio inputs are converted into text

for tasks such as transcription and real-time translation. Additionally, these models are increasingly employed in more complex environments, such as multimodal dialogue systems, where they combine audio inputs with visual and textual data to understand user intentions in a conversational setting.

2.1.3 Challenges

One of the significant challenges facing audio multimodal models is the variability in audio quality and background noise, which can hinder performance. Additionally, the alignment between audio and other modalities is not always straightforward, as temporal synchronization and context relevance must be carefully handled to ensure accurate and coherent outputs. Multimodal hallucination, where the model generates responses inconsistent with the audio content, is another issue that still requires attention in current research.

2.2 Large Language Models

Large Language Models (LLMs) have revolutionized natural language processing (NLP) through their remarkable ability to handle a wide range of tasks, from language translation to complex reasoning. These models are built on the transformer architecture and are pre-trained on large-scale text corpora, enabling them to perform zero-shot and few-shot learning without requiring explicit fine-tuning. The versatility of LLMs is rooted in their ability to capture rich contextual information, emergent properties, and the use of attention mechanisms, allowing them to process long sequences of text. The development of LLMs follows decades of research in language modeling, transitioning from statistical approaches to deep neural networks. As models grew in size and complexity, they exhibited emergent capabilities such as instruction following, in-context learning, and multi-step reasoning.

2.2.1 Architectures of LLMs

The foundation of most modern LLMs is the transformer architecture, introduced in the paper “Attention Is All You Need” [19]. This architecture is designed to address the limitations of previous recurrent models by using self-attention mechanisms to capture dependencies between distant tokens in a sequence. Transformers can be categorized into three main architectural types: encoder-only, decoder-only, and encoder-decoder models [20].

- **Encoder-only models**: These models, such as BERT, focus on understanding input sequences by processing them in their entirety. They are suited for tasks like text classification and sentence understanding.
- **Decoder-only models**: Decoder-only models, such as GPT and its successors, predict the next token in a sequence in an autoregressive manner. They are most effective for text generation tasks.

- **Encoder-decoder models**: This architecture is primarily used for sequence-to-sequence tasks like translation and summarization, where the encoder processes the input and the decoder generates the output.

2.2.2 Pre-training and Fine-tuning Paradigms

LLMs are pre-trained on massive text datasets, often using self-supervised learning techniques such as masked language modeling (MLM) or autoregressive prediction. Pre-training enables the model to learn a general understanding of language, which can be fine-tuned on specific tasks using smaller, task-specific datasets. Fine-tuning enhances model performance by tailoring its knowledge to particular domains.

LLMs like GPT and LLaMA are pre-trained using vast datasets containing web text, books, and other publicly available sources. Techniques like Reinforcement Learning from Human Feedback (RLHF) have also been used to align LLM outputs with human preferences, improving their utility in real-world applications [20].

2.3 Variational Autoencoder

Variational Autoencoders (VAEs) are a class of deep generative models that combine deep neural networks with probabilistic modeling to learn latent variable representations of data. VAEs aim to model complex data distributions, such as images or text, by learning the underlying latent variables that generate the observed data. VAEs were introduced as a method to address the intractability of exact Bayesian inference by approximating the posterior distribution using a simpler, tractable family of distributions.

2.3.1 Architecture and Mechanism

VAEs consist of two main components: the encoder (inference model) and the decoder (generative model). The encoder approximates the posterior distribution $q(z \mid x)$, mapping input data x to latent variables z, while the decoder generates data x from these latent variables using the generative distribution $p(x \mid z)$. Instead of directly learning these distributions, VAEs optimize the Evidence Lower Bound (ELBO) using variational inference, minimizing the Kullback-Leibler (KL) divergence between the true posterior and the approximate distribution [21].

During training, VAEs use a reparameterization trick to enable backpropagation through stochastic layers. This trick ensures that the gradients can flow through the latent variable sampling process, allowing the model to learn parameters effectively through gradient-based optimization.

2.3.2 Amortized Variational Inference

In traditional variational inference, the optimization process can be slow and computationally expensive. VAEs address this with amortized variational inference, where a single function is learned to infer the latent variables for all data points, rather than

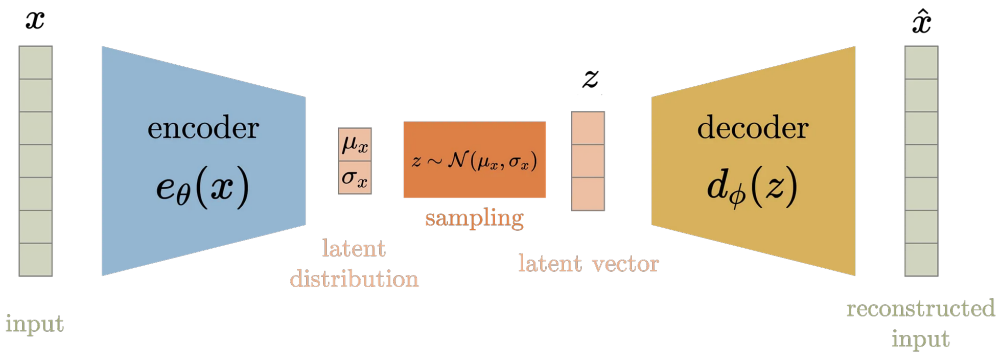


Figure 2.1: *The architecture of VAE. Adapted from [6]*

independently optimizing for each data point. This approach allows VAEs to scale to large datasets efficiently, using neural networks to estimate the variational parameters.

2.3.3 Applications and Advances

VAEs have been widely applied in tasks such as image generation, clustering, and unsupervised learning. Recent advancements have focused on improving the expressiveness of both the likelihood function and the posterior distribution. For example, PixelVAE models pixel dependencies within an image for better image generation quality, and Importance-Weighted Autoencoders (IWAE) increase the flexibility of the posterior approximation, enabling VAEs to model more complex data distributions [21].

2.3.4 Challenges and Limitations

Despite their versatility, VAEs face several challenges. One significant issue is the suboptimality in inference, where the learned approximation to the posterior distribution is not always optimal, leading to poor performance in some tasks. Another challenge is their application to text data, where the discrete nature of words makes it difficult for VAEs to generate high-quality textual outputs without significant adjustments to the model.

2.4 Stable Diffusion

Stable Diffusion is a powerful text-to-image generation model based on Denoising Diffusion Probabilistic Models (DDPMs). Unlike generative adversarial networks (GANs), diffusion models iteratively add and then remove noise from data, which allows them to model complex image distributions. Stable Diffusion uses a more efficient variant called Latent Diffusion Models (LDMs). LDMs reduce the computational load by working in a compressed latent space instead of the high-dimensional pixel space, significantly improving the efficiency of both training and inference.

2.4.1 How Stable Diffusion Works

Stable Diffusion operates in two key phases:

- Encoding: The model compresses input images into a lower-dimensional latent representation using an autoencoder. This compressed space preserves crucial semantic information while eliminating irrelevant details, reducing computational demands.
- Denoising: Once the image is transformed into this latent space, the model adds noise in a controlled manner. The process is then reversed, where the model gradually removes the noise, reconstructing the original image in its latent space. The image is then decoded back to the pixel space.

This approach, built on latent diffusion models, allows for high-quality image generation with lower computational costs compared to pixel-based diffusion models. By employing a UNet architecture combined with cross-attention mechanisms, Stable Diffusion is capable of generating highly detailed images based on text prompts [22].

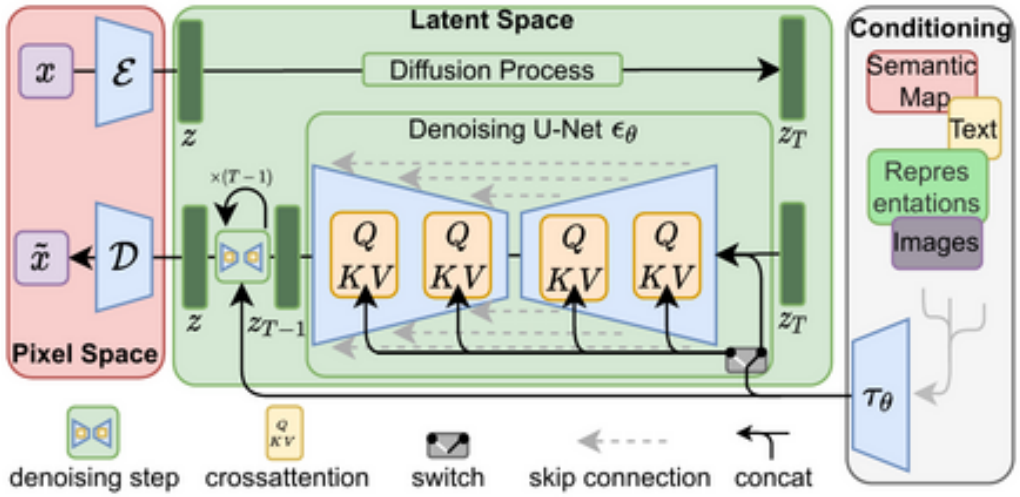


Figure 2.2: *The architecture of Stable Diffusion models. Source: [5]*

Κεφάλαιο 3

Multimodal Audio and Language Models for Lyric Generation

In this chapter, we analyze the models that are relevant to our task. Specifically, we review the literature regarding audio multimodal models and specifically audio encoders commonly used in audio multimodal models, and music understanding models. We also review, vision multimodal models, specific large language models which are either explicitly used in our work, or constitute part of other used multimodal language models. Finally, we review several methods used in the literature for evaluating generation tasks.

3.1 Audio Encoders used in Audio Multimodal Models

In this section we analyze some commonly used audio encoders in audio multimodal models, which have also been studied and used in this thesis.

3.1.1 Whisper Audio Encoder: Large-Scale Speech Recognition

Introduced in the paper "Robust Speech Recognition via Large-Scale Weak Supervision" [23], Whisper is an advanced speech recognition system developed by OpenAI, leveraging large-scale weak supervision to achieve robust performance across a wide range of environments. Trained on 680,000 hours of multilingual and multitask data sourced from the web, Whisper models are capable of generalizing well to unseen datasets in a zero-shot transfer setting without fine-tuning. Its encoder-decoder architecture, based on the transformer model, is designed to handle complex audio inputs and convert them into high-quality transcriptions.

3.1.1.1 Architecture and Design

Whisper adopts an encoder-decoder transformer framework, where raw audio is first processed into 16,000 Hz log-magnitude Mel spectrograms, followed by encoding into feature representations. The encoder uses convolutional layers and residual blocks with self-attention mechanisms to capture the underlying structure in audio sequences. The decoder then predicts the transcript tokens based on these encoded audio features.

The model is pre-trained on a highly diverse set of transcripts from the web, enabling Whisper to handle varying audio quality, speaker accents, and languages effectively. Additionally, it supports multitask training for speech transcription, translation, and voice activity detection [23].

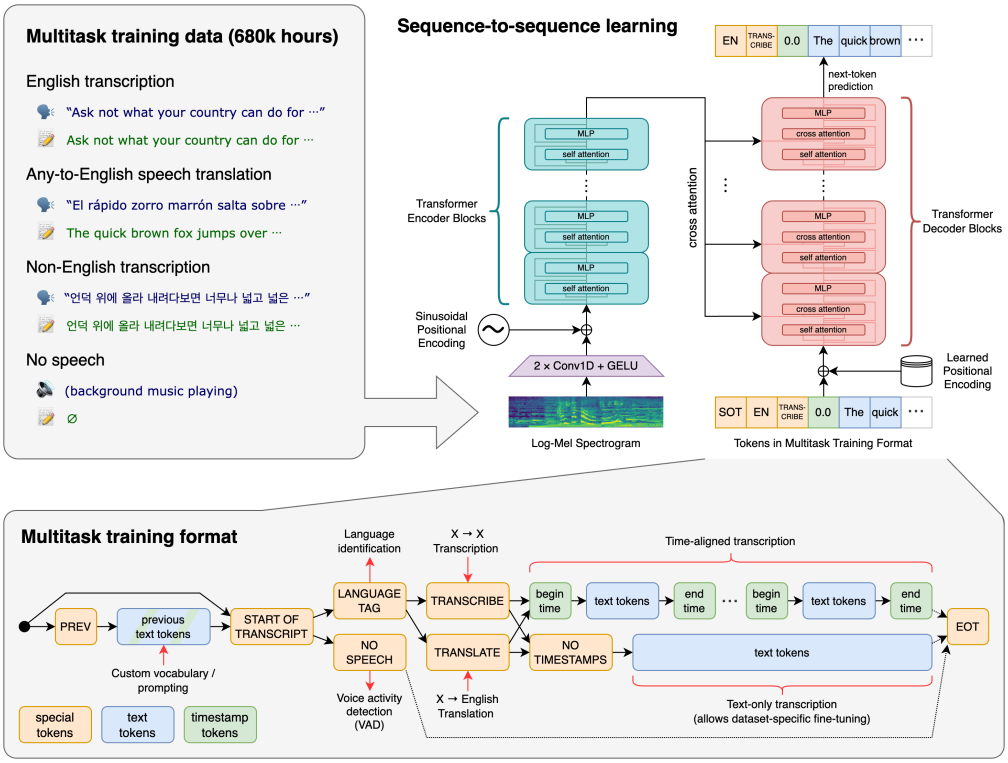


Figure 3.1: *Overview of the architecture and training approach of Whisper. Source: [7]*

3.1.1.2 Multilingual and Multitask Capabilities

Whisper distinguishes itself from other speech recognition models by its focus on multilingualism and multitasking. Its dataset includes over 117,000 hours of non-English audio, covering 96 languages. This broad language scope allows Whisper to excel in zero-shot multilingual transcription tasks. In addition to transcription, Whisper can handle spoken language identification and translation from various languages to English, making it versatile for global applications [23].

3.1.1.3 Zero-Shot Performance and Robustness

A notable feature of Whisper is its ability to perform reliably in zero-shot settings, transcribing audio from datasets it has never encountered during training. Evaluations demonstrate that Whisper approaches human-level accuracy in terms of speech recognition and shows remarkable robustness across varied audio domains, including noisy environments and long-form transcription tasks. The model's generalization capabilities make it well-suited for real-world applications where fine-tuning may not always be feasible [23].

3.1.2 Audio Spectrogram Transformer (AST): A Pure Attention-Based Model for Audio Classification

The Audio Spectrogram Transformer (AST) [24] is the first model to perform audio classification using a purely attention-based architecture, without relying on convolutional layers. AST leverages the transformer framework, which was initially developed for natural language processing, to process spectrograms of audio inputs. This approach captures long-range dependencies within the data, offering superior performance over traditional convolutional neural networks (CNNs) and CNN-attention hybrid models.

3.1.2.1 Architecture

AST transforms audio data into 128-dimensional log-mel spectrograms, which are then split into 16x16 patches. These patches are flattened and linearly projected into 1D patch embeddings. The transformer model's encoder processes these embeddings along with learnable positional encodings to maintain spatial information. A [CLS] token is appended at the beginning of the sequence to handle classification tasks. The output of this [CLS] token is passed to a linear layer for final classification. AST employs a 12-layer, 12-head transformer encoder, which is similar to the Vision Transformer (ViT) architecture.

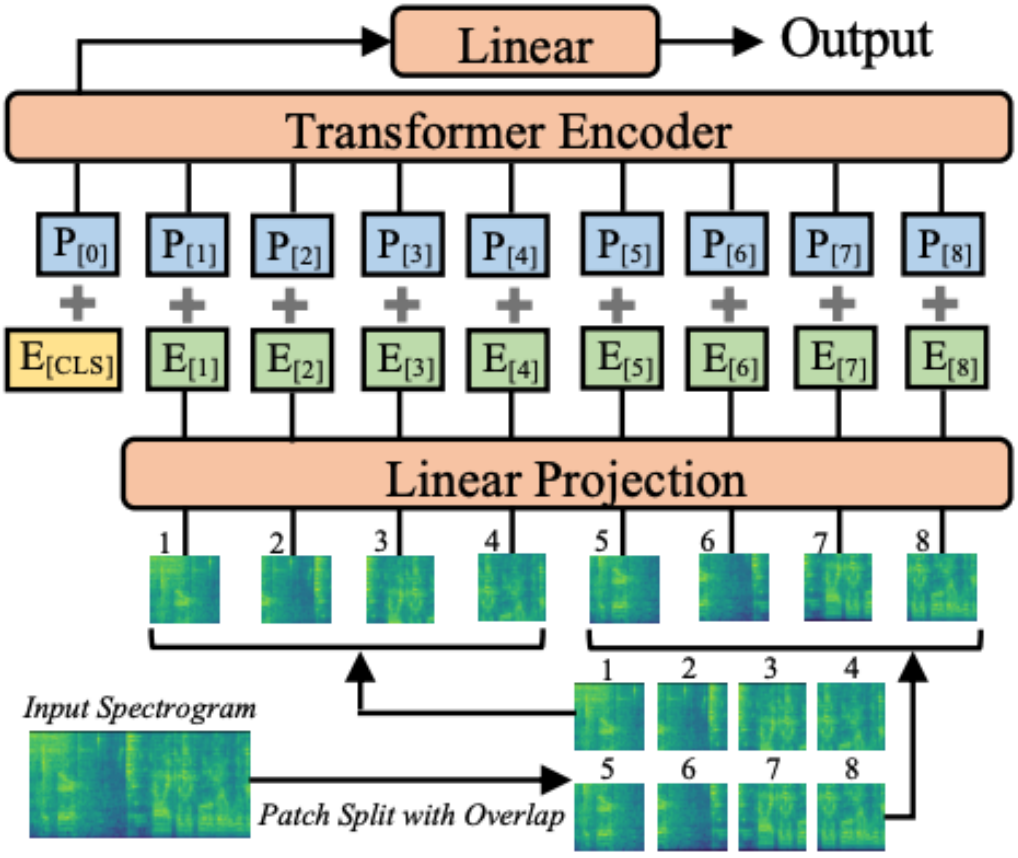


Figure 3.2: *The architecture of AST. Source: [8]*

3.1.2.2 Transfer Learning from Vision Transformers

To overcome the challenges of limited labeled audio data, AST utilizes pre-trained weights from Vision Transformers (ViT) trained on ImageNet. Knowledge transfer is achieved by adapting the positional embeddings and input representations from the visual domain to the audio domain. This cross-modality transfer learning significantly enhances the model's performance across diverse audio tasks.

3.1.2.3 Performance and Applications

AST has been evaluated on several benchmark datasets, including AudioSet, ESC-50, and Speech Commands. It achieves state-of-the-art results across all these datasets, with an mAP of 0.485 on AudioSet, 95.6% accuracy on ESC-50, and 98.1% accuracy on Speech Commands V2. AST's purely attention-based structure allows it to outperform CNN-based and hybrid models, particularly in tasks involving long-range temporal dependencies, without requiring task-specific architecture modifications.

3.1.2.4 Advantages and Limitations

AST's main advantage lies in its simplicity and flexibility. Unlike CNN-based models, which require tuning for different input lengths and tasks, AST can handle variable-length inputs without architectural changes. Furthermore, its attention-based mechanism captures global context better than local convolutional filters. However, AST's reliance on pre-trained models like ViT highlights its need for large-scale datasets for optimal performance. The model also exhibits higher computational costs compared to CNNs, especially when processing longer sequences.

3.1.3 MERT: Acoustic Music Understanding with Large-Scale Self-Supervised Training

MERT (Music undERstanding model with large-scale self-supervised Training) [9] is a large-scale, self-supervised learning (SSL) model designed for acoustic music understanding. Developed to bridge the gap between speech and music processing, MERT is fine-tuned on musical features, enabling it to outperform conventional models in a wide range of Music Information Retrieval (MIR) tasks.

3.1.3.1 Architecture and Training

MERT's core architecture leverages self-supervised learning similar to that of speech-based models, such as HuBERT, by utilizing the masked language modeling (MLM) framework. The model is trained with teacher models to generate pseudo labels for masked audio segments, combining acoustic and musical representations to enhance learning. It integrates two primary teacher models:

- **Acoustic Teacher**: A Residual Vector Quantization-Variational Autoencoder (RVQ-VAE) is used to provide high-resolution acoustic features, aiding the model in understanding musical timbre and structure.

- **Musical Teacher**: The Constant-Q Transform (CQT) focuses on tonal and harmonic structures, which are essential for tasks like pitch detection and chord recognition.

MERT is scaled from 95 million to 330 million parameters, employing a 12-layer transformer-based encoder. The use of convolutional layers alongside transformers allows for efficient and robust feature extraction, making it capable of understanding complex musical patterns.

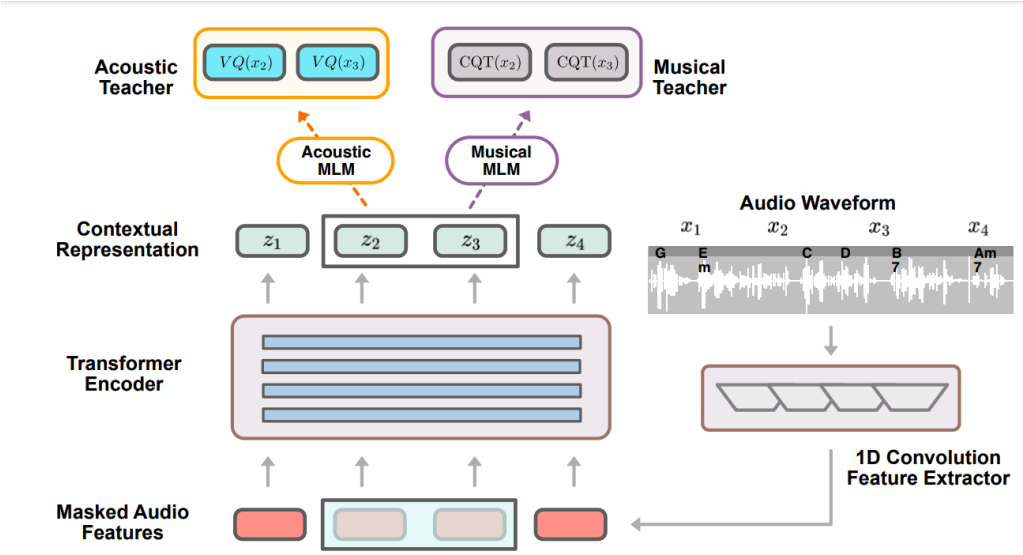


Figure 3.3: *The architecture of MERT. Source: [9]*

3.1.3.2 Applications and Performance

MERT achieves state-of-the-art results across 14 music understanding tasks, including music tagging, beat tracking, pitch detection, source separation, and genre classification. It has demonstrated exceptional performance in capturing both local musical features, such as beats and timbre, and more global features like key detection and emotion recognition.

Its innovative multi-task pre-training paradigm, using both acoustic and musical teacher models, allows MERT to generalize across different MIR tasks without needing task-specific architectures. This general-purpose approach makes it suitable for a wide range of applications in both industry and research.

3.1.3.3 Challenges and Future Directions

Despite its strong performance, MERT faces challenges in processing long musical sequences due to the limitations of 5-second training segments, potentially affecting tasks requiring longer contexts. Addressing this issue in future work could further improve its capabilities in more complex music understanding tasks.

Additionally, the model's pre-training paradigm, particularly with teacher models like RVQ-VAE and CQT, could be refined to offer even more detailed acoustic and musical rep-

resentations. The model's open-source availability is expected to foster further research in music SSL, promoting innovation and broader application in MIR fields.

3.2 Music Understanding in Multimodal Audio Models

Many SOTA multimodal audio models have been trained with a focus on classification tasks regarding speech and non-musical sounds. In this section, we analyze some music understanding multimodal audio models, used in MIR tasks such as music tagging, captioning, and question-answering for musical input.

3.2.1 SALMONN: Towards Generic Hearing Abilities for Large Language Models

SALMONN (Speech Audio Language Music Open Neural Network) [25] is a novel multimodal large language model (LLM) developed to process and understand general auditory inputs. Built by integrating pre-trained language models with speech and audio encoders, SALMONN exhibits the capability to handle three distinct types of sounds: speech, audio events, and music. It achieves competitive performance across a range of auditory tasks, including automatic speech recognition (ASR), translation, audio captioning, emotion recognition, and more.

3.2.1.1 Architecture

SALMONN uses a dual-audio encoder system, combining the Whisper speech encoder and the BEATs audio encoder. The Whisper encoder, sourced from OpenAI, specializes in speech recognition and translation, while BEATs focuses on non-speech audio events through self-supervised learning. Both encoders are synchronized and processed through a window-level Q-Former, which converts variable-length audio input into text-like token sequences. This design enables SALMONN to handle multimodal input efficiently while aligning auditory information with textual language models.

The Q-Former structure interacts with the outputs from both encoders, transforming them into augmented audio tokens, which are then fed into the backbone LLM, Vicuna. To adapt the audio tokens into the Vicuna model's input space, SALMONN employs a LoRA (Low-Rank Adaptation) module to align the audio and text tokens for coherent multimodal reasoning [25].

3.2.1.2 Cross-Modal and Emergent Abilities

One of SALMONN's key innovations is its ability to perform cross-modal tasks, which were unseen during training. This includes tasks like speech translation into untrained languages, audio-based storytelling, and speech audio co-reasoning. These emergent abilities highlight SALMONN's capacity for zero-shot generalization, meaning it can perform well in tasks without prior exposure. SALMONN also supports speech-based question answering, speaker verification, and music captioning, making it a versatile tool for auditory information processing.

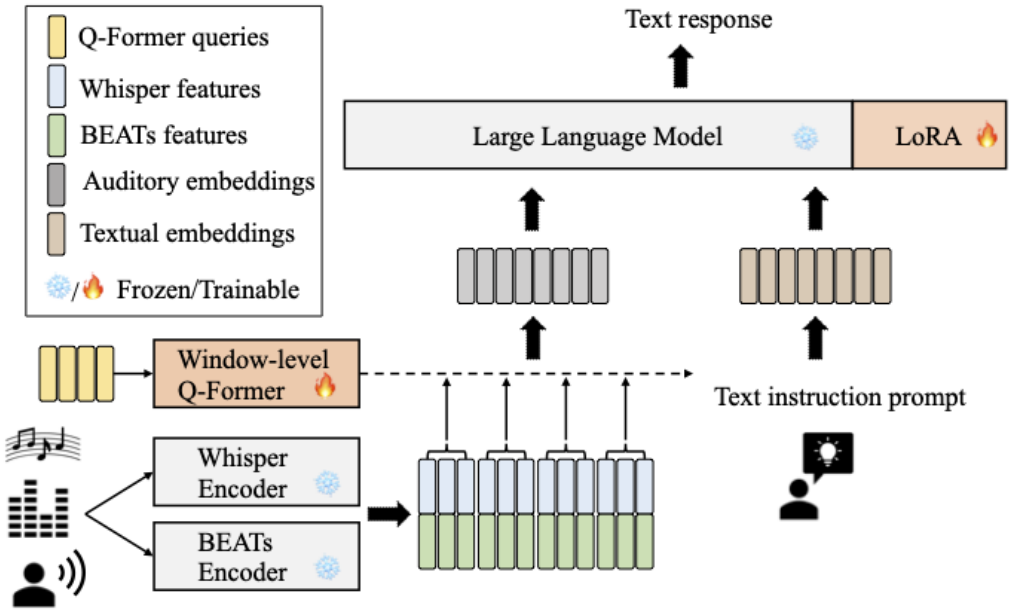


Figure 3.4: *The architecture of SALMONN. Source: [2]*

To activate these emergent cross-modal abilities, SALMONN employs activation tuning. This technique improves the model's performance on untrained tasks by fine-tuning specific components of the model, such as the LoRA and Q-Former modules, without sacrificing performance on pre-trained tasks [25].

3.2.1.3 Applications and Performance

SALMONN excels in a variety of speech and audio tasks, outperforming many existing models in areas such as ASR, translation, and audio captioning. Additionally, its emergent abilities in tasks like storytelling and co-reasoning open up new possibilities for AI applications in areas requiring deep audio understanding [25].

3.2.2 MusiLingo: Bridging Music and Text with Pretrained Language Models

MusiLingo is a music-language model developed to bridge the gap between music and natural language by generating accurate captions and answering music-related queries. The model integrates a MERT (Music Understanding Model) encoder, which extracts acoustic and musical features, with a pre-trained Vicuna LLM. The core design uses a simple linear projection layer that maps music representations into text embedding space, followed by a temporal compression layer to handle music-text alignment efficiently.

3.2.2.1 Architecture and Training

The architecture comprises a MERT encoder for extracting meaningful audio representations from music clips. These representations are passed through a linear adapter layer before being projected into the Vicuna model's text embedding space. The MusiLingo model undergoes two key training phases: a pre-training phase, where it is trained on

music caption datasets, and an instruction-tuning phase, where it is fine-tuned on a Q&A dataset (MusicInstruct). The integration of these two phases allows MusiLingo to generate rich and accurate music-related captions and answer open-ended music queries.

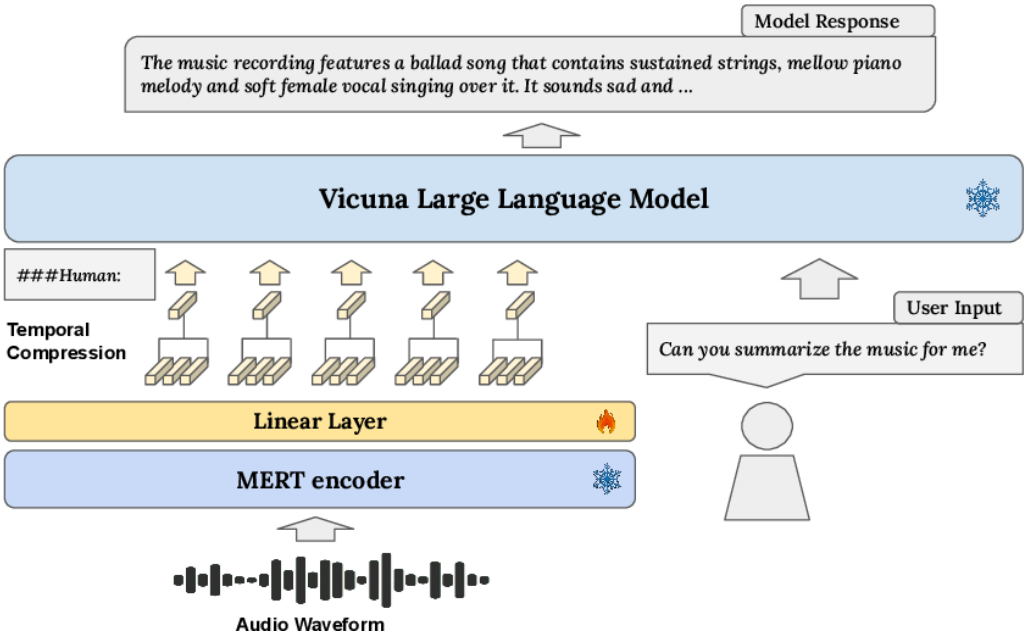


Figure 3.5: *The architecture of MusiLingo. Source: [10]*

3.2.2.2 Music Captioning and Instruction Following

MusiLingo achieves competitive results on tasks like music captioning and answering detailed music questions. The model is fine-tuned on MusicInstruct, a dataset specifically created for open-ended music queries, which enhances its ability to follow instructions and generate human-like responses. MusiLingo outperforms other models like MU-LLaMA in several evaluation metrics, particularly in music Q&A tasks, as evidenced by its strong performance on both short and long-format question-answer datasets.

3.2.2.3 Performance and Applications

MusiLingo demonstrates state-of-the-art (SOTA) performance on several music information retrieval tasks, including music tagging, captioning, and Q&A. It provides detailed answers about musical genres, instruments, moods, and even specific user inquiries, offering a robust system for both music-related research and practical applications, such as recommendation systems or music cataloging. MusiLingo's versatility in handling both objective and subjective questions makes it an important tool for advancing music-language model research.

3.2.3 MU-LLaMA: a Model for Music Question Answering and Music Caption Generation

The MU-LLaMA model [26] represents an advanced solution in the realm of music

understanding, particularly addressing tasks like music question answering and caption generation. The model is built on top of Meta's LLaMA language model and utilizes the MERT (Music Embedding Representation Transformer) model as a music encoder. The goal is to overcome the limitations posed by the scarcity of large-scale public datasets for text-to-music generation (T2M-Gen) by enhancing the model's capability to understand music and generate meaningful text descriptions.

3.2.3.1 Architecture

The architecture of MU-LLaMA is divided into three key components: a pretrained MERT encoder, a Music Understanding Adapter, and the LLaMA language model itself. The MERT encoder processes raw audio inputs into feature embeddings, which are then aggregated and passed through a dense neural network in the adapter. This enables the model to capture complex musical features such as mood, instrumentation, and genre. These features are fed into the top layers of the LLaMA model, allowing it to handle both music question answering and text generation tasks effectively.

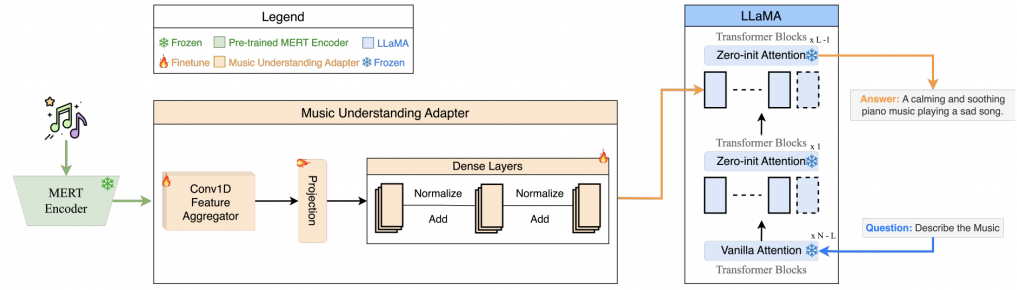


Figure 3.6: *The architecture of MU-LLaMA. Source: [11]*

3.2.3.2 Performance

MU-LLaMA's performance, validated on datasets like MagnaTagATune and MusicCaps, outperforms existing models such as LTU and LLaMA-Adapter across multiple metrics (BLEU, METEOR, ROUGE-L). The paper further highlights the adaptability of MU-LLaMA to various music understanding tasks, marking it as a significant advancement in generating high-quality text-music pairs and enhancing music comprehension for T2M-Gen models.

3.2.4 M²UGen: Multi-modal Music Understanding and Generation with the Power of LLMs

The M²UGen model [27] provides a novel framework for both music understanding and multi-modal music generation using Large Language Models (LLMs). It addresses a gap in existing research, which often focuses on either understanding or generation but rarely both. The M²UGen framework integrates music, image, and video modalities through various pre-trained encoders, including the MERT model for music, Vision Transformer (ViT) for images, and ViViT for videos. These encoders extract feature embeddings that

are aligned and processed by the LLaMA 2 model to perform downstream tasks such as music question answering and text/image/video-to-music generation.

3.2.4.1 Architecture

At the core of M^{2}UGen's music understanding capabilities lies the LLaMA 2 model, which fuses multi-modal context information from its encoders through specialized adapters. These adapters transform the multi-modal features into a format that the LLaMA model can process for both music understanding and generation tasks. For music generation, the system integrates either AudioLDM 2 or MusicGen models as decoders, allowing M^{2}UGen to translate input prompts into coherent music outputs. Through systematic evaluations, M^{2}UGen is shown to outperform state-of-the-art models in music question answering, music editing, and multi-modal music generation, highlighting its significant contribution to the field of AI-driven artistic music creation.

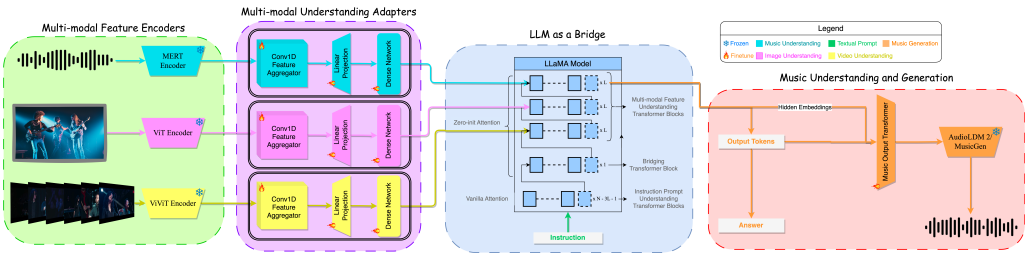


Figure 3.7: *The architecture of M^{2}UGen. Source: [12]*

3.3 Vision Multimodal Models: An Overview

Vision multimodal models are perhaps the most studied within the multimodal domain, primarily due to the integration of visual data with text, enabling tasks like image captioning, visual question answering (VQA), and more recently, multimodal reasoning and dialogue systems. These models integrate visual information with language models to provide rich interpretations of images, videos, and even 3D environments [28].

3.3.1 Overview and Architecture

A vision multimodal model typically consists of an image or video encoder, a pre-trained LLM, and a modality interface that aligns the visual features with textual information. The image encoders, such as CLIP, process visual data into compact feature vectors that are semantically aligned with textual representations [18].

The modality interface plays a pivotal role in bridging the visual and textual modalities. It typically performs token-level or feature-level fusion, allowing the LLM to generate responses grounded in both visual and textual contexts. Recent advancements, such as Flamingo's use of cross-attention layers, have improved how vision and language interact, leading to more sophisticated understanding and reasoning [18].

3.3.2 Applications

Vision multimodal models are widely used in numerous applications. In visual question answering (VQA), these models can answer questions based on an image, demonstrating the ability to reason across modalities. They also play a crucial role in generating image captions, creating stories from visual content, and even offering real-time assistance in understanding complex visual scenes in domains like medicine and robotics.

Another notable application is the embodied agents field, where vision multimodal models guide robots or virtual agents in interacting with the physical world. These agents use visual data to navigate and understand their surroundings, bridging the gap between artificial intelligence and real-world environments [18].

3.3.3 Challenges

One of the prominent challenges in vision multimodal models is multimodal hallucination, where the model incorrectly interprets or adds details not present in the visual input. This can be particularly problematic in domains requiring high precision, such as medical image interpretation. Another challenge lies in the fusion of modalities; while vision models excel at understanding static images, integrating data that require long context, an example of which are temporal data like videos, introduces complexity in handling this information [18].

Additionally, scaling vision multimodal models to handle high-resolution images or detailed scene information without compromising speed or computational efficiency remains a critical research area. Ensuring these models can generalize well across diverse visual domains, from everyday scenes to specialized fields like medical imaging, is essential for their future development.

3.4 LLaVA: Large Language and Vision Assistant

LLaVA (Large Language and Vision Assistant) [29] is an advanced multimodal model designed to integrate vision and language understanding. The model bridges the gap between large language models (LLMs) and vision encoders, enabling the model to follow human instructions and reason about images in real-world applications. LLaVA was developed through visual instruction tuning, which extends instruction-following data to multimodal spaces, allowing the model to process complex vision-language tasks.

3.4.1 Architecture and Data

LLaVA integrates the CLIP (Contrastive Language-Image Pretraining) visual encoder with the Vicuna LLM. The architecture involves using the CLIP-ViT-L-14 model to generate image embeddings, which are then linearly projected into the Vicuna language model's embedding space. This alignment between vision and text tokens allows the model to perform diverse multimodal tasks, such as image captioning, detailed visual description, and visual reasoning.

To train LLaVA, multimodal instruction-following data is generated using GPT-4 and ChatGPT. This data consists of questions, detailed descriptions, and complex reasoning tasks about images, allowing LLaVA to follow detailed instructions and generate accurate responses. The dataset includes approximately 158,000 samples of language-image pairs, which enable the model to learn robust visual reasoning.

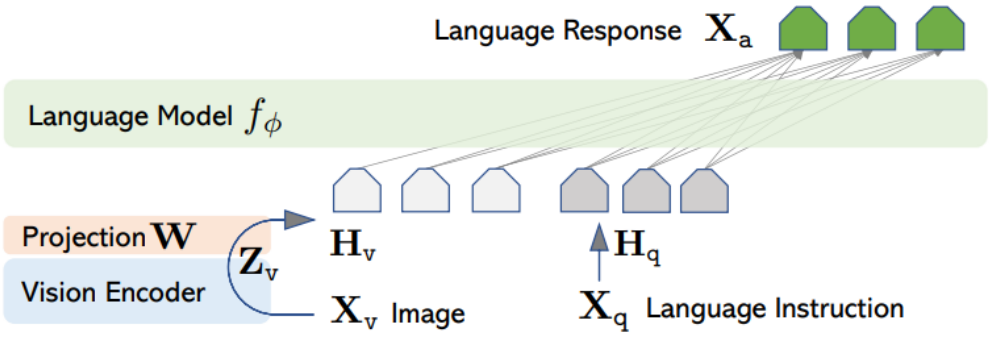


Figure 3.8: *The network architecture of LLaVA. Source: [13]*

3.4.2 Performance and Applications

LLaVA achieves strong performance on tasks such as visual chat, image-based reasoning, and multimodal question-answering. Notably, LLaVA outperforms other multimodal models, including BLIP-2 and Qwen-VL-Chat, particularly in visual instruction-following tasks. The model also demonstrates competitive accuracy on the ScienceQA benchmark, a dataset designed to evaluate multimodal reasoning in scientific contexts, achieving state-of-the-art results when ensembled with GPT-4 [29].

3.4.3 Visual Instruction Tuning

The concept of visual instruction tuning is central to LLaVA's success. This technique involves training the model on machine-generated multimodal data that mimics human instruction-following behavior. By leveraging GPT-4 for data generation, LLaVA is able to handle complex reasoning tasks that require both image and language comprehension. The visual instruction tuning pipeline ensures that the model can generalize across different visual domains, making it suitable for tasks such as object detection, image captioning, and complex image reasoning [29].

3.5 Large Language Models

3.5.1 GPT-2: The Evolution of OpenAI's LLMs

GPT-2 [30] represents the second iteration of the Generative Pre-trained Transformer (GPT) family developed by OpenAI. GPT-2 was a significant step forward in NLP due to its large model size (1.5 billion parameters), robust language generation abilities, and lack of need for task-specific fine-tuning.

3.5.1.1 Architecture and Training

GPT-2 is a decoder-only transformer model, employing a unidirectional approach to text generation. It generates text by predicting the next token in a sequence, considering all previous tokens in the context. GPT-2's autoregressive nature allows it to excel at generative tasks, such as text completion, summarization, and creative writing.

The model was pre-trained on the WebText dataset, which contains millions of web pages. During training, the model learns to predict the next word based on the surrounding context, allowing it to generate coherent and contextually relevant text over long sequences.

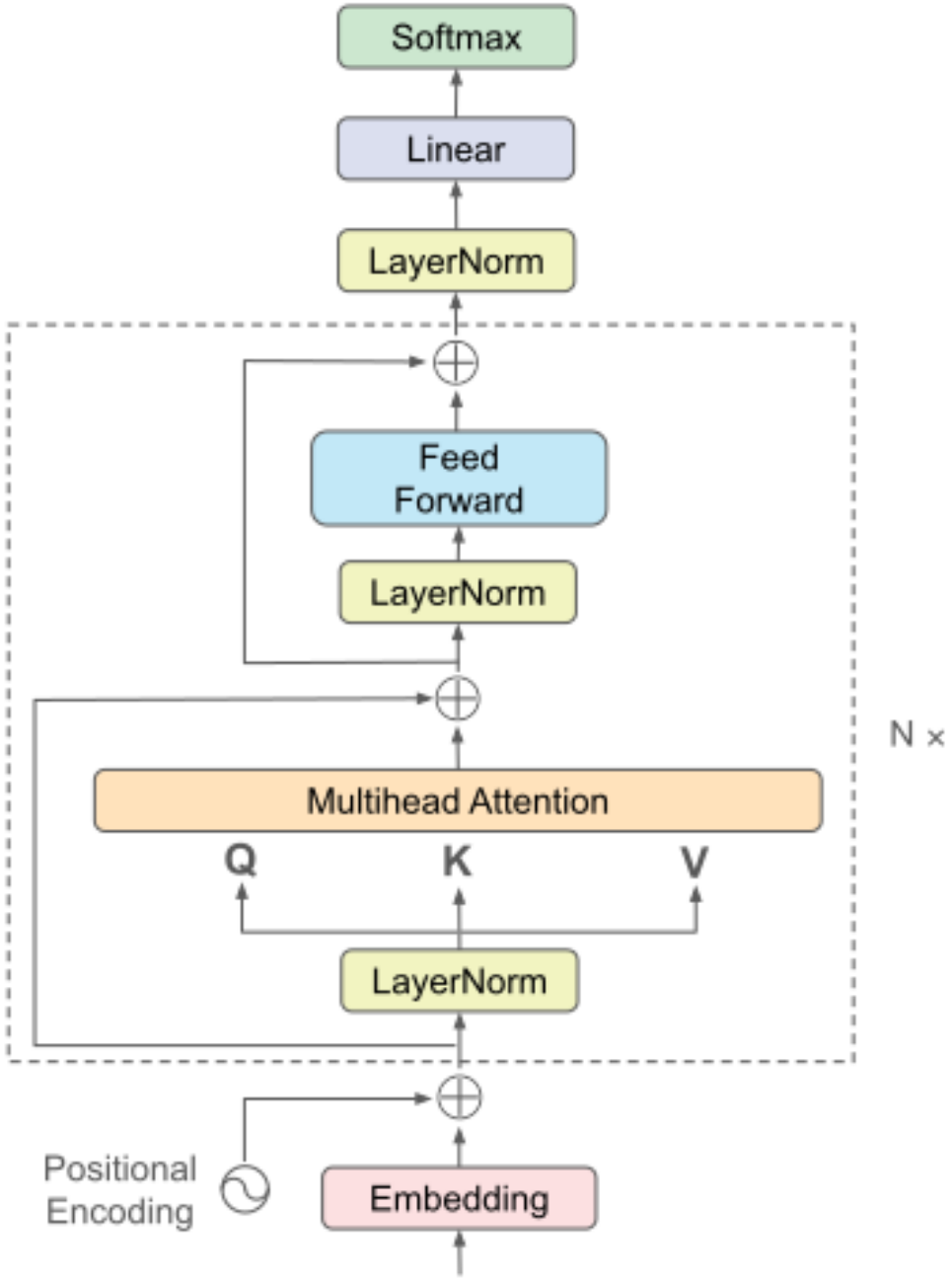


Figure 3.9: *The GPT-2 architecture. Adapted from [14]*

3.5.1.2 Contributions and Limitations

GPT-2's release demonstrated the power of large-scale language models in generating human-like text across various domains. Its ability to perform tasks without explicit supervision, such as answering questions or translating text, showcased the potential of few-shot and zero-shot learning.

However, GPT-2 also has limitations. As a purely autoregressive model, it struggles with long-range dependencies and may produce inconsistent outputs when generating long text sequences. Additionally, its training on open web data can lead to the generation of biased or harmful content, an issue that subsequent models have addressed with reinforcement learning techniques.

3.5.2 LLaMA: A Foundation Model by Meta

LLaMA (Large Language Model Meta AI) [31] is a family of open-source LLMs developed by Meta. LLaMA's primary contribution lies in offering competitive performance with significantly smaller model sizes compared to proprietary models like GPT-3. Released with parameters ranging from 7B to 65B, LLaMA demonstrated that smaller models, when trained on carefully curated data, could outperform larger closed-source models.

3.5.2.1 Architecture and Training

LLaMA utilizes a transformer architecture similar to GPT models, with key optimizations to improve efficiency. For instance, it incorporates SwiGLU activations instead of standard ReLU and rotary positional embeddings instead of absolute positional embeddings. These architectural tweaks enhance the model's performance, particularly in terms of handling long sequences of text.

LLaMA models are trained on a carefully curated collection of publicly available datasets, with over a trillion tokens drawn from diverse sources such as Common Crawl. These models leverage high-quality, diverse datasets to improve their language understanding and generalization capabilities. The emphasis on data quality and efficient architecture allows LLaMA to achieve high performance while requiring relatively fewer parameters compared to some other large language models of similar capability.

3.5.2.2 Applications and Impact

LLaMA has had a profound impact on the open-source community, serving as the backbone for several derivative models, including Vicuna and Alpaca. Its open nature has encouraged widespread experimentation, contributing to the development of more efficient, task-specific LLMs. However, like other LLMs, it still faces challenges related to bias, toxicity, and hallucination in its outputs [20].

3.5.3 Vicuna: A Finetuned LLaMA

Vicuna [32] is a 13B-parameter chat model, derived from the LLaMA-1 model and fine-tuned using instruction-following datasets. Developed by the Vicuna team, it was

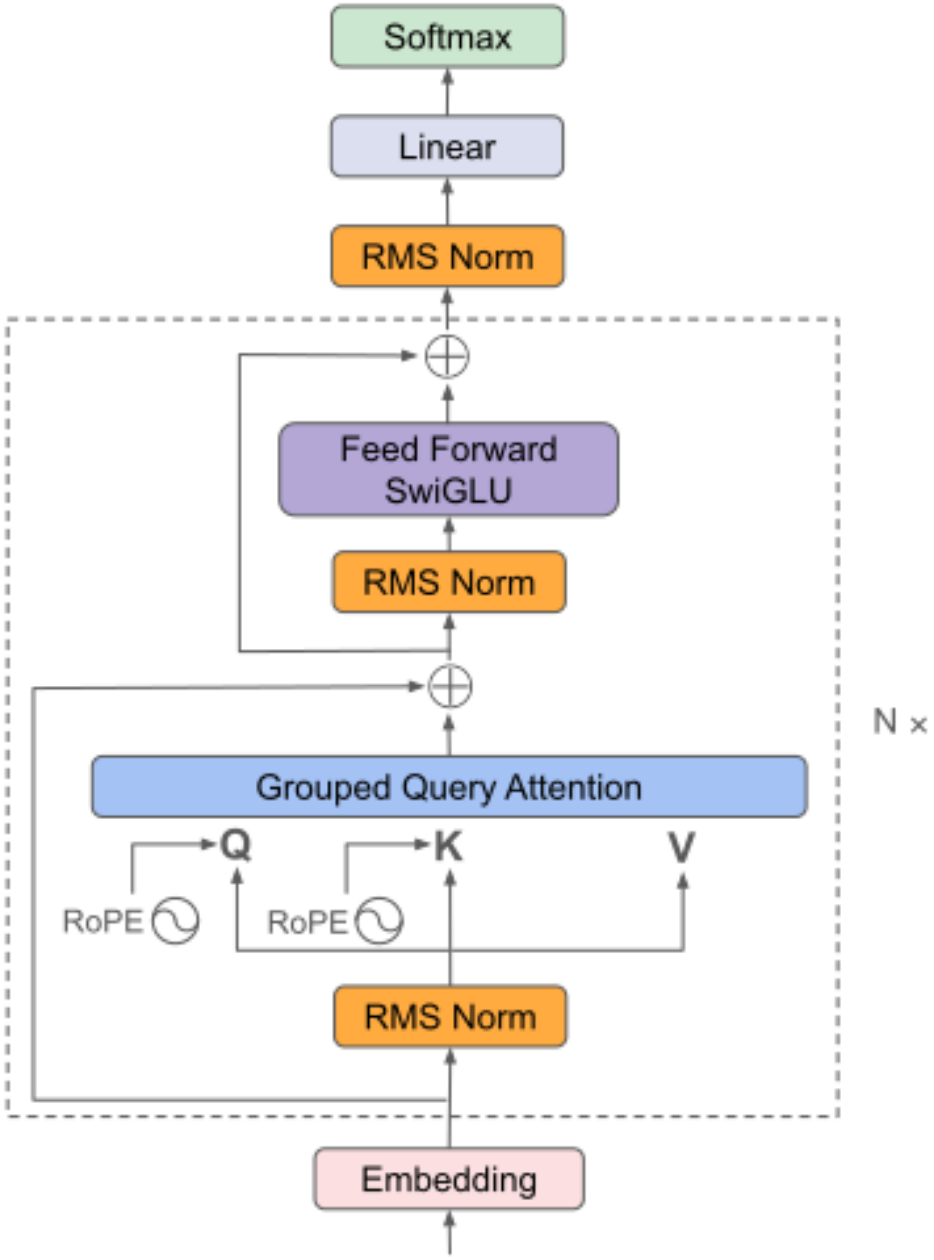


Figure 3.10: *The Llama architecture. Adapted from [14]*

designed to enhance the conversational abilities of LLaMA models by training on real-world dialogue data collected from platforms like ShareGPT.

3.5.3.1 Fine-tuning and Performance

Vicuna's training process involved fine-tuning LLaMA on user-shared conversations, enabling the model to generate more contextually relevant and human-like responses in interactive settings. By employing GPT-4 as an evaluator, the Vicuna team reported that the 13B-parameter model achieved 90% of the performance quality of proprietary models like OpenAI's ChatGPT.

Despite its relatively small size, Vicuna has demonstrated competitive performance in

chat-based applications, including question-answering, summarization, and multi-turn dialogue. Its lightweight nature makes it particularly appealing for deployment in real-time, user-facing systems.

3.5.3.2 Challenges

Although Vicuna has made significant strides in conversational AI, it still grapples with limitations inherent to LLaMA-based models, such as bias and difficulty in maintaining coherence over long conversations. Additionally, while it provides a strong open-source alternative, Vicuna's performance still falls short of proprietary models in certain complex reasoning tasks.

3.5.4 Mistral: A High-Performance LLaMA Variant

Mistral [15] is a recent addition to the LLM landscape, offering a 7B-parameter model that outperforms both LLaMA-2 (13B) and LLaMA-1 (34B) across various benchmarks. This model is built to maximize performance while minimizing computational costs, making it a highly efficient option for a range of NLP tasks.

3.5.4.1 Architecture and Optimization

Mistral uses several architectural enhancements to achieve superior performance. One notable feature is the implementation of grouped-query attention, which reduces inference costs by enabling the model to handle longer sequences more efficiently. Mistral also employs sliding window attention, allowing it to process arbitrary-length sequences without compromising on accuracy or speed.

The model's training regime is designed to optimize for tasks such as code generation, reasoning, and mathematical problem-solving, outperforming larger models like LLaMA-1 34B on several evaluation metrics [15].

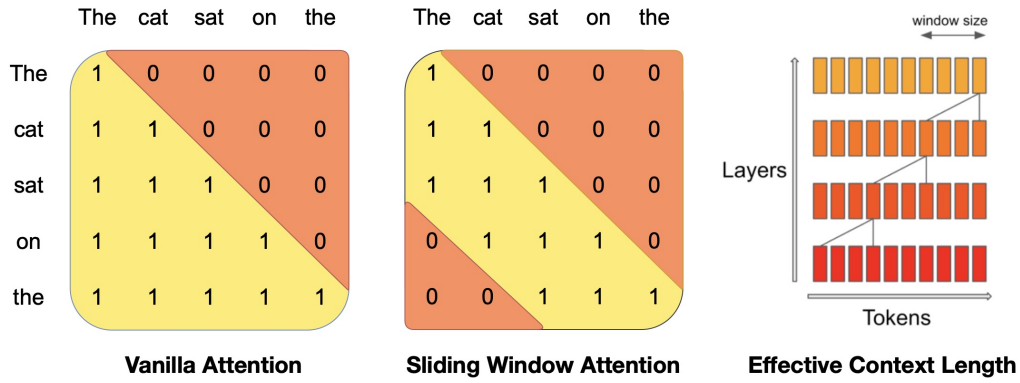


Figure 3.11: *The Mistral sliding window attention. Source: [15]*

3.5.4.2 Advantages and Limitations

Mistral's strength lies in its ability to achieve high performance with fewer parameters, making it a cost-effective choice for researchers and developers seeking efficient LLMs. Its

improvements in attention mechanisms also make it well-suited for tasks requiring long-term dependency modeling.

However, as with other LLaMA-based models, Mistral faces ongoing challenges in handling ambiguous or biased content. Additionally, while it excels in specific domains like reasoning and code generation, its general-purpose language understanding capabilities may lag behind those of larger models.

3.5.5 Mistral-7B-OpenOrca: Fine-tuning on the OpenOrca Dataset

The Mistral-7B-OpenOrca model [33] is a fine-tuned variant of the Mistral-7B model. It was trained on a carefully curated subset of GPT-4-augmented data from the OpenOrca dataset, designed to replicate the dataset used in Microsoft's Orca research. Mistral-7B-OpenOrca outperforms other models in its size category, ranking number 1 on the Hugging Face Leaderboard for models smaller than 30B parameters at release. The model achieved significant performance boosts on several benchmarks, including MMLU, ARC, and HellaSwag, with superior results in reasoning, mathematics, and code generation tasks. The model's fine-tuning process involved 4 epochs of training on 8 A6000 GPUs, achieving remarkable results at a cost-effective level.

This model exemplifies how meticulous data filtering and fine-tuning strategies can enhance the performance of smaller models, enabling them to rival much larger models in various tasks. It is also optimized for deployment on consumer GPUs, providing high accessibility for developers and researchers.

3.5.6 Claude: Exceptionally Good in Creative Writing

3.5.6.1 Claude 3 Overview

Claude 3, developed by Anthropic and introduced in March 2024, represents a significant advancement in large language models (LLMs). The Claude 3 family consists of three key variants: Claude 3 Opus, the most capable model; Claude 3 Sonnet, which balances speed and skill; and Claude 3 Haiku, the fastest and least expensive version. All three models incorporate multimodal capabilities, enabling them to interpret and analyze visual inputs, such as images and charts, alongside text, thus expanding their potential applications [34].

3.5.6.2 Capabilities of Claude in Creative Writing

The Claude 3 models, particularly Opus and Sonnet, demonstrate increased competence in tasks involving creative writing, detailed analysis, and structured outputs. As reported by Anthropic, Claude 3 is better at creative writing tasks compared to Claude 2.1. Also, Claude 3 shows a 63% win rate against a baseline Claude Instant model in the task of creative writing. Given that the authors of the paper "A Confederacy of Models: a Comprehensive Evaluation of LLMs on Creative Writing" [35] report that Claude Instant

1.2 ranked third or higher in all of their rubric items, specifically ranking second in cohesion, and second overall, after GPT-4. This shows that Claude 3, given its reported much better performance in creative writing, can be considered very promising for creating creative and coherent lyrics. In addition, as reported by Anthropic, the newest model Claude 3.5 Sonnet shows even better performance compared to Claude 3 to all of the evaluated tasks [36].

3.6 Evaluation Methods for Generation Tasks

Evaluation methods are much more straightforward in classification tasks, where the performance of a model can be measured easily by comparing the output with the gold labels of the dataset. However, this method cannot always be used in generative tasks like lyric generation. In this chapter, we go over some of the methods used in literature that are also suitable for the evaluation of the models used for the lyric generation task.

3.6.1 Objective Metrics

In the MusicJam paper [1], the metrics that were used are:

- BLEU [37]: the BLEU metric is usually used in machine translation tasks, and measures the overlap of n-grams between gold and generated text. It is represented by the formula:

$$BLEU_N = BP \cdot \sum_{n=1}^{N} exp(w_n log p_n),$$

  where BP is the brevity penalty, which penalizes shorter outputs in the generated text according to the formula $min(1, \frac{reference\_length}{translated\_length})$, p_n reflects gram precision by quantifying the precision of n-grams by evaluating the ratio of shared n-grams between the machine-generated translation and the reference text, compared with the overall count of n-grams in the machine-generated translation, and w_n are the corresponding weights for each n-gram. For example, in $BLEU_2$ and $BLEU_3$, the weights are $(0.5, 0.5)$ and $(0.333, 0.333, 0.334)$ respectively.

  This metric provides insight into how closely generated lyrics align with ground truth data, however, it favors verbatim matches and may fail to capture creative deviations that are still musically appropriate.

- Distinct/Diversity [38]: this score is calculated according to the formula:

$$Distinct_N = \frac{\mid unique(Ngrams) \mid}{\mid Ngrams \mid}.$$

  It evaluates diversity by analyzing the number of unique n-grams in the generated lyrics. While it promotes lexical variety, high diversity can sometimes come at the cost of coherence, especially if the lyrics become too disjointed or unnatural.

- Novelty [39]: this metric calculates the ratio of infrequent n-grams, to total number of n-grams. The paper deems the n-grams that are not among the 2000 most

frequent phrases as infrequent. The corresponding formula is:

$$Novelty_N = \frac{\mid infrequent(Ngrams) \mid}{\mid Ngrams \mid}.$$

Novelty quantifies the originality of the lyrics by comparing them to commonly used phrases. This metric is essential for evaluating creative output but can mislead if rare or obscure phrasing sacrifices clarity or relevance to the theme of the music.

- Coherence [40]: coherence is calculated by counting the number of the repeated words in the lyrics generated for one song, and then taking the average of the results. It is expressed by the formula:

$$Coherence = \frac{1}{M} \cdot \sum_{k=1}^{M} \sum_{i=1}^{n_k} \mathbb{1}(count(w_i) > 1),$$

where M represents the number of the songs, n_k the number of the words generated for the song k and w_i the i-th word of the lyrics.

While it is very important for a song to be coherent, this way of measuring coherence does not capture meaningfully the concept of coherence. For example, if the model repeats the same words, the score will be high but the lyrics won't necessarily be coherent, meaningful or have a natural flow.

Overall, while these metrics provide valuable insights, they each have inherent limitations. BLEU and Coherence emphasize linguistic accuracy over creativity, potentially undervaluing imaginative lyrics. Distinct and Novelty, on the other hand, may prioritize uniqueness at the expense of clarity. A more holistic evaluation could involve human assessments that account for thematic consistency, emotional resonance, and overall musicality.

3.6.2 Similarity Score with Cross Encoder

Cross-encoders are transformer-based models designed to capture the relationship between input pairs. Cross encoders take two inputs and encode them together into a shared representation, which is different from bi-encoders where the two inputs are passed independently into BERT. This shows that, while more computationally expensive, using a cross-encoder can capture more accurately the similarity between two texts [41].

This similarity scoring method helps avoid the weaknesses of scores likes BLEU, which matches only n-grams overlaps and therefore misses when two texts have the same meaning but use synonyms or differently phrased sentences.

3.6.3 JudgeLM: Scalable LLM for Evaluation

JudgeLM is a fine-tuned large language model designed to evaluate the performance of other LLMs in open-ended tasks [17]. Traditional evaluation metrics often fall short in assessing LLMs due to the complexity and variability of their outputs. JudgeLM addresses

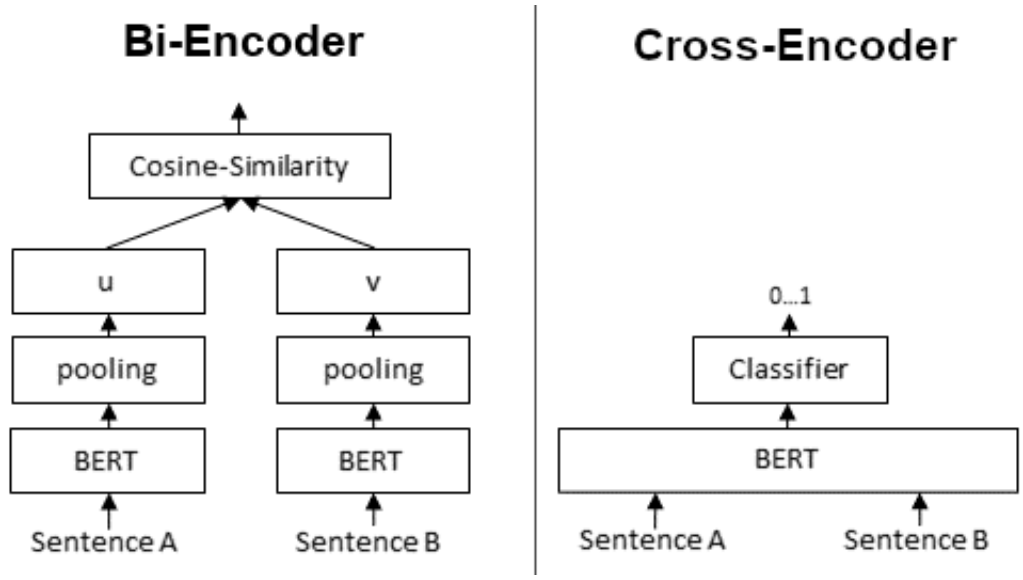


Figure 3.12: *Bi-encoder vs Cross-encoder architecture. Source: [16]*

this by fine-tuning open-source models (such as Vicuna) with a large-scale dataset of 105K seed tasks and GPT-4-generated judgments. It is trained to act as a scalable and efficient evaluator, surpassing human-to-human agreement levels, with a focus on grading, judging, and reasoning.

3.6.3.1 Architecture and Techniques

JudgeLM utilizes a scalable architecture, available in sizes ranging from 7B to 33B parameters. It is fine-tuned using innovative techniques like swap augmentation, reference support, and reference drop to mitigate inherent biases (e.g., position bias, knowledge bias, and format bias). The swap augmentation technique ensures that the model judges content rather than position by training it on data where the positions of the answers are swapped. Reference support allows JudgeLM to leverage reference answers to improve accuracy in fact-based tasks, while reference drop helps the model handle both referenced and non-referenced formats, increasing flexibility [17].

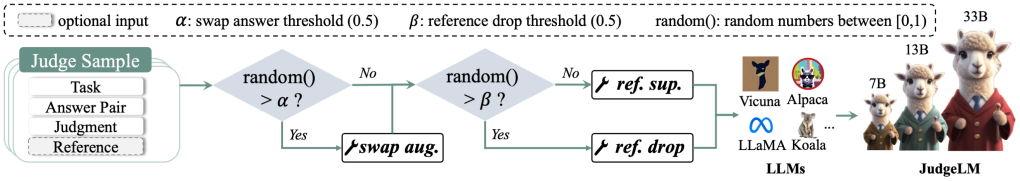


Figure 3.13: *An illustration of the JudgeLM's fine-tuning and the used methods to mitigate bias. Source: [17]*

3.6.3.2 Using It to Evaluate Other Models

JudgeLM is capable of evaluating a range of models across different tasks, including single-answer and multi-answer evaluation, multimodal model judgments, and multi-turn dialogues. In practice, JudgeLM compares model outputs to reference answers or

peer models, providing detailed scores and reasoning. It achieves over 90% agreement with GPT-4 on several benchmarks, such as PandaLM, surpassing even GPT-3.5. Additionally, JudgeLM can efficiently evaluate 5,000 sample pairs in just three minutes using 8 A100 GPUs, making it both cost-effective and scalable compared to traditional human or GPT-4-based evaluations [17].

3.6.4 LLM-based Evaluation

Recent studies show that evaluations made by LLMs can be better than objective metrics, especially for tasks that require creativity and diversity [42]. Given specific criteria to evaluate, strong LLMs can evaluate the generated lyrics on criteria like creativity, coherence, naturality and singability.

A problem often raised with this method is the bias that the LLMs may incorporate when making these judgements: they have been observed to exhibit position bias, which is the propensity to favor certain positions over others, verbosity bias, which is when an LLM favors longer, verbose responses even when they are of lower quality, and self-enhancement bias, which is when LLMs favor the answers generated by themselves [43].

Strong LLMs, and specifically GPT-4 has been shown in the paper "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena" [43] to have 80% of agreement with human annotators, which is equal to the agreement between human annotators.

However, in the paper "Reference-Guided Verdict: LLMs-as-Judges in Automatic Evaluation of Free-Form Text" [44], it was shown that using multiple and diverse LLMs as judges mitigates individual model biases and significantly improves alignment with human judgments, especially in challenging tasks regarding free-form text, where traditional metrics and single-model evaluations fall short. The specific method that they followed was to employ two open-source models and one closed-source model: Mistral-Instruct-7B-v0.3, Llama-3.1-70B and GPT-3.5-turbo.

Κεφάλαιο 4

Methodology

In this section, we present the method that we followed, regarding the preprocessing of the dataset that we used, the combination of modalities and models that we tested for the lyric generation process.

4.1 Our Dataset

In this section, we talk about the dataset that we used for training our models and for inferring our suggested pipelines. We analyze the preprocessing that we performed in order to get the suitable form for our task, and to increase the performance of our models.

4.1.1 About the DALI dataset

The dataset that we used is the DALI dataset, a dataset of synchronized audio, lyrics and vocal notes [45]. The dataset, that is comprised of 7756 songs, contains the MP3 files of a variety of songs, annotated with lyrics, in different levels of granularity: paragraphs, lines, words and notes. Each song is also annotated with metadata, such as genre, language, artist, song title and album.

4.1.2 Dataset Preprocessing

4.1.2.1 Source Separation

The MP3 files of the dataset are not source-separated, meaning they do not have separate MP3 files that correspond to voice and accompaniment. Therefore, in order to acquire the music accompaniment, we used a source-separation library called Spleeter [46], which has the option of two-stem separation, namely for vocals and accompaniment. This separation is crucial in order to ensure that the models don't learn to give outputs based on the vocals.

4.1.2.2 Acquiring the Useful Subset of the Dataset

When we trained our models without doing a more thorough preprocessing, the quality of the lyrics was severely negatively affected. This was due to the fact that the dataset has many entries where the lyrics contain the following characteristics:

- Incorrect spelling of words in the lyrics.
- Separation of one word into smaller parts: in many of the entries of the dataset, because the singer pronounces the words by separating them in smaller chunks, the annotation split the words accordingly (e.g. the word 'tomorrow' was written as 'tomo rrow').
- Unconventional contractions: in a plethora of entries in the dataset, there were words that used apostrophes in an unusual way (e.g. 'ev'rything' instead of 'everything').
- Some entries were wrongly tagged as English songs, when they were in another language.

These peculiarities led us to conduct the following preprocessing:

- Using the granularity of 'lines', we removed the entries of the dataset which had lines with words that were consecutively repeated for more than three times. This was done to avoid making the model repeat the same words during the inference.
- We removed the entries that were wrongly tagged as English in the dataset, firstly by using a language detection model based on the XLM-RoBERTa model [47]. Then, we also removed manually the entries that, even if they were mostly in English, contained some phrases in foreign languages.
- There were entries that had hyphens and other special characters like numbers and parentheses. We removed the entries that have hyphens, and we removed the special characters from the entries, either by replacing them with spaces, or in the case of numbers, by spelling them out (e.g. '2' was replaced by 'two').
- Because there were many entries with contractions, we could not pick them out and remove them, so we replaced them with their full form. This was done for both the usual contractions and the unconventional ones, because we observed that the VAE model generated many contractions, which were unusual and didn't make sense, before this preprocessing.
- Because of the high frequency of incorrect spellings and separated words, we fixed these entries in the following way: we used a Python library called Spellchecker [48] to detect the words that were deemed as 'unknown' by the library, and then we manually fixed them.

This preprocessing resulted in 3111 songs, which were split in training and test set in percentage of 80-20.

4.2 The Tested Models for each Combination of Modalities

In this chapter, we analyze each of the combinations of modalities that we tested for the lyric generation process. Specifically, we dive into the models that correspond to each

of these combinations of modalities, by analyzing their architectures and the prompts that we used.

4.2.1 Text to Text

The text to text generation process is comprised of an LLM that takes into the input the beginning of a song, and is asked to generate the rest of it.

4.2.1.1 Out-of-the-box: Claude 3.5 Sonnet, GPT-2, OpenOrca, Vicuna

Firstly, in the text to text generation, we tested the LLMs that are also later used with the addition of other modalities, making these experiments also serve as ablation studies for the proposed following pipelines. The LLMs that were tested are Claude 3.5 Sonnet, GPT-2, Mistral OpenOrca and Vicuna. This testing is initially done out-of-the box, namely without any finetuning on our selected dataset. Because GPT-2 is a completion, and not an instruction model, we just provide the first line of the song without any instruction. For the other three models, the prompt that was used can be seen below:

Prompt used to generate lyrics with Claude OpenOrca and Vicuna in Text-to-Text

```
You are a helpful assistant that creates song lyrics given the first
line of the song. The lyrics should be coherent and creative.
First line: {first_line}
Lyrics:
```

4.2.1.2 Finetuned LLMs: GPT-2, OpenOrca

We also tested whether finetuning GPT-2 and OpenOrca on the DALI dataset would improve the quality of the output. Regarding the form of the dataset for this specific training setup, each training sample is a song, and specifically, the input given to the model during training is of the form:

```
### These are some song lyrics: {lyrics}
```

We employed the LoRA (Low-Rank Adaptation) technique for fine-tuning, which allows for efficient adaptation of large language models.

4.2.2 Text & Audio to Text

The text & audio to text generation process consists of giving the music accompaniment and previous lyrics as input. The two models that were tested and belong in this category are the VAE AST-GPT2 model, and the Whisper-OpenOrca model.

4.2.2.1 VAE AST-GPT2 model: a Reproducibility Study

Aiming to reproduce the work of Chuer Chen et al. [1], we follow a similar approach to the one described in their paper. Given that the code, their preprocessing method, and

therefore the subset of the dataset that they used, aren't available, we tried to follow as closely their work with the information that is given in the paper.

In the lyric generation model of their work, they introduce a variational autoencoder based on GPT-2. The architecture of the model resembles a Transformer, where the encoder is the audio encoder AST, and the decoder is GPT-2. The model works as follows: the mel-spectrogram of a 5-second music accompaniment is fed into the AST encoder. From the output of the encoder, the hidden representation H_{music} is retained. The hidden representation is then transformed into two vectors, which are meant to capture the distribution of the music, which are computed according to the following formulas:

$$\mu = W_{\mu} H_{music}$$

$$\sigma = exp(\frac{W_{\sigma} H_{music}}{2})$$

Then, reparameterization is performed, which is a standard process found in VAE models, which is a random sampling from the normal distribution, and which ensures that the latent vector is non-deterministic. We retain this vector in this way:

$$z = \mu + \sigma \odot \epsilon, \epsilon \sim \mathcal{N}(0, I)$$

After, this latent vector is fed into the multimodal decoder through a cross-attention layer. The decoder takes the previous line of lyrics in its input, which is used as context for the generation of the next lyric line. The architecture of the model can be seen in Figure 4.1.

The paper reports that they used the DALI dataset, from which they also retained only the music accompaniment. They report that they used 5-second lines, and that they labeled each lyric line with its previous one for the context that is fed into the decoder, using the "<START>" special token for the first line of each song. They report that they retained 2590 songs from the DALI dataset, and splitting them into 2072 and 518 for the training and test set respectively.

We followed the same training method that they used: the encoder and multimodal decoder are comprised of 12 layers each, with 12 attention heads. We also used a loss that combines reconstruction loss and KL divergence as the training objective, described by the equations below:

$$L_{\partial}(x, y, z, \hat{y}) = L_{reconstr}(y, \hat{y}) + \beta KL(q_{\phi}(z \mid x) \parallel p(z))$$

where $L_{reconstr}$ calculates the dissimilarity between the generated lyrics y and the gold lyric $\hat{y}$, $KL(q_{\phi}(z \mid x) \parallel p(z))$ assesses the difference between the $p(z)$ and the distribution produced by the encoder, and β is hyper-parameter controlling the loss contribution from the KL divergence.

4.2.2.2 Whisper-OpenOrca model: Harnessing the Power of Advanced Models

By reviewing the results of the previous model, we noticed that their quality was not sufficient. Because of the limitation of the context with the previous method, the

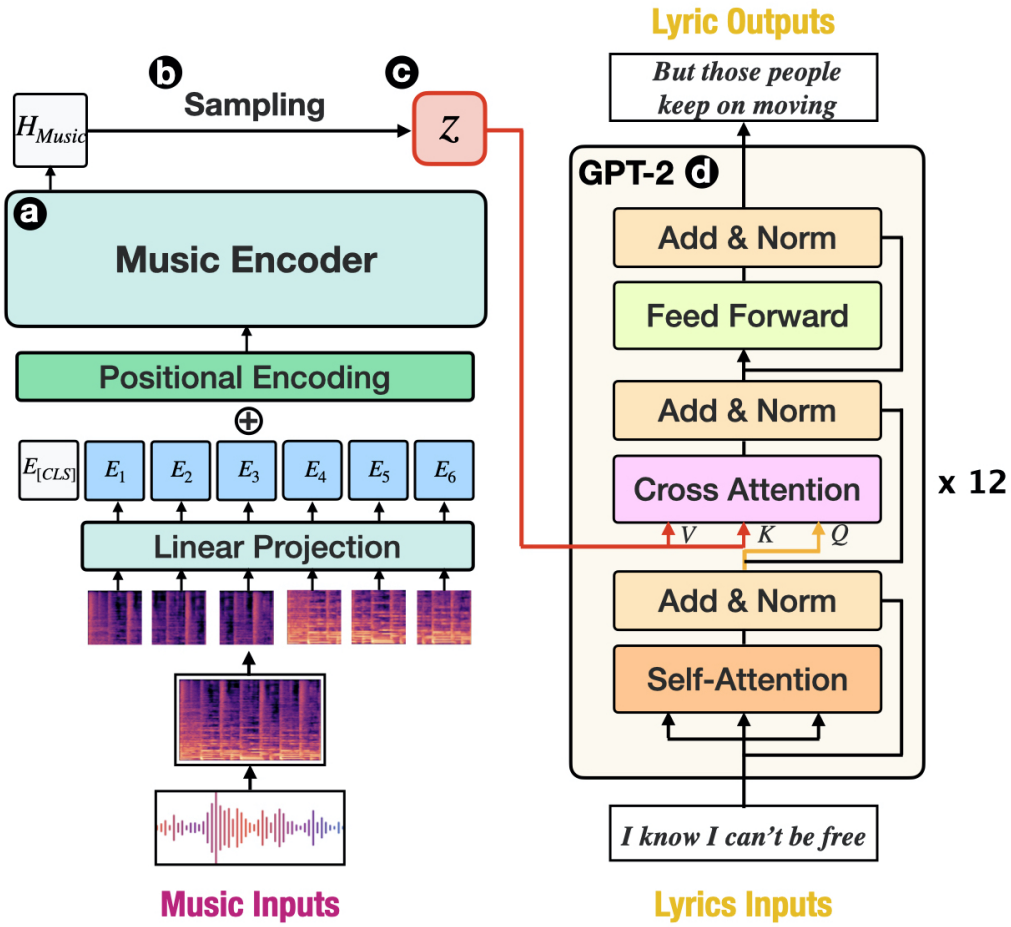


Figure 4.1: *The VAE AST-GPT2 model. Source: [1]*

generated lyrics weren't meaningful and coherent to a sufficient degree (the results are further explored in the section 5.3). This led us to explore other models with newer architectures and further pretraining.

In this model, we utilize the Whisper model as the audio encoder, and the Mistral OpenOrca as the LLM. These models are loaded with their pretrained weights, and the alignment between audio representations and lyrics is achieved by a projection layer between these two models. This projection layer is the only trainable module of this model. The employed loss function is the cross-entropy loss between the gold lyrics ($\hat{y}$) and the generated lyrics (y), the formula of which can be seen below:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

In order to fix the problem of limited context, which reduces significantly the coherence of the output lyrics, we trained the model by giving the whole previous gold lyrics context into the prompt of the LLM. In a similar manner, during inference, the trained model is prompted to either start creating or continue the given music lyrics, given the music accompaniment. Because this LLM is an instruct model, in contrary to GPT-2 which is a completion model, instead of only giving the previous lyrics in the prompt, we also give

an instruction. The specific prompt that was used can be seen below:

Prompt used to provide the OpenOrca LLM with the lyric context in Whisper-OpenOrca model

```
<|im_start|>system
You are an assistant that helps create the lyrics of a song in a way that
matches the given music accompaniment. You are given the previous lyrics of
the song and you create the next lyrics. When the token '<START>' is given as
previous lyrics, you write the beginning of the song.
<|im_end|>
<|im_start|>user
Previous lyrics: {previous_lyrics}
```

The architecture of the model can be viewed in Figure 4.2.

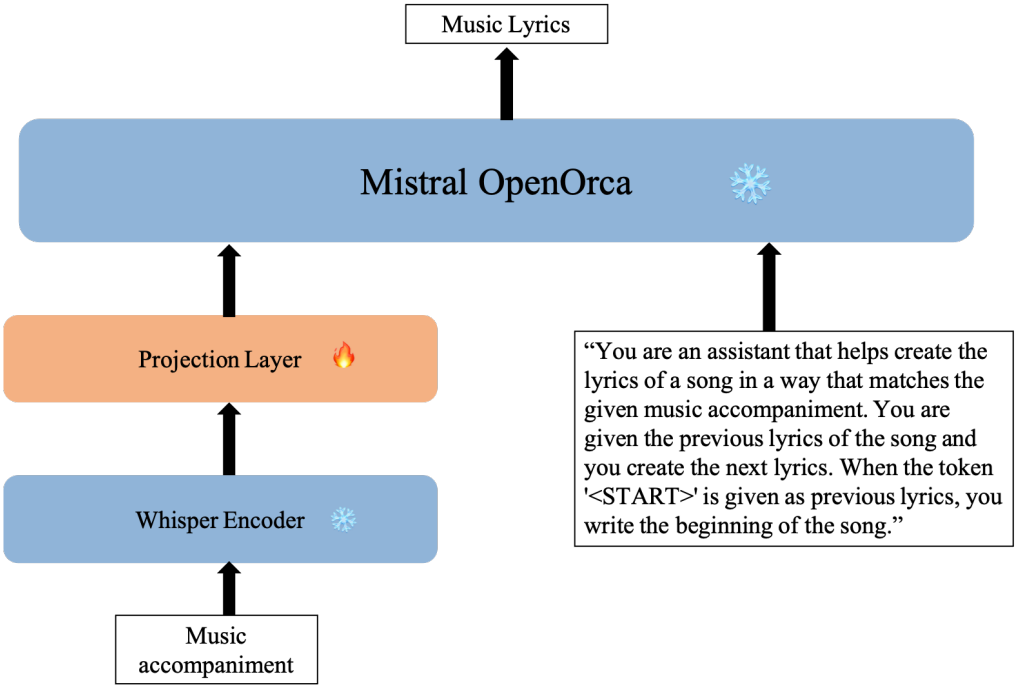


Figure 4.2: *The Whisper-OpenOrca model.*

4.2.3 Text & Audio to Text to Text

The audio to text to text generation process consists of an extra generation step compared to the previous method: music tags are intermediately produced, which are then used to create the final song lyrics. The model used in this setup is SALMONN-Claude. With this and the next model, we wanted to examine how the lyric generation would be affected when we pass from different modalities before we end up to the lyric generation process.

4.2.3.1 SALMONN-Claude: Guiding Lyric Generation with Autogenerated Music Tags

With this model, we harnessed the creativity and extraordinary performance of Claude regarding creative writing tasks. Specifically, we used the pretrained model SALMONN to extract tags from the music. We specifically used prompts that were found in the appendix of the corresponding paper [25], because they were showcased to have good results. We asked the model to give a detailed description of the given music accompaniment, and to extract the emotion of the music, and the musical instruments that it can hear. These tags were then used to prompt Claude to generate lyrics based on these given music tags.

We performed prompt tuning in order to improve the outputs of the music understanding model, and we ended up using these:

Prompt that extracts detailed description of the song in SALMONN-Claude pipeline

```
Please describe the music in detail.
```

Prompt that extracts the emotion of the song in SALMONN-Claude pipeline

```
What is the emotion of the music? Explain the reason in detail.
```

Prompt that extracts the musical instruments in the song in SALMONN-Claude pipeline

```
Which musical instruments do you hear?
```

We also tested if few-shot learning could improve even further the quality of the lyrics, by formulating two different prompts for the Claude model: a zero-shot one, and a few-shot one. The form of the prompts given to Claude can be seen below (we give a 2-shot example for the few shot one, for simplicity):

Zero shot prompt given to the Claude model in SALMONN-Claude pipeline

```
You are a helpful assistant that writes song lyrics based on provided tags
like instruments that are used, the sentiment of the song, and an overall
description of the background music. The lyrics should be original, coherent,
and singable. If an existing song is referenced, the lyrics should be a new
creation inspired by the music accompaniment of the song.
Instruments: {instruments}
Sentiment: {sentiment}
Description: {description}
```

Few shot prompt given to the Claude model in SALMONN-Claude pipeline

```
You are a helpful assistant that writes song lyrics based on provided tags
like instruments that are used, the sentiment of the song, and an overall
description of the background music.  The lyrics should be original, coherent,
and singable.  If an existing song is referenced, the lyrics should be a new
creation inspired by the music accompaniment of the song.
<examples>
    <example>
        Instruments:  {instruments_example1}
        Sentiment:  {sentiment_example1}
        Description:  {description_example1}
        Lyrics:  {lyrics_example1}
    </example>
    <example>
        Instruments:  {instruments_example2}
        Sentiment:  {sentiment_example2}
        Description:  {description_example2}
        Lyrics:  {lyrics_example2}
    </example>
</examples>
```

For the few-shot prompting, we gave to the model six examples, two from each of the most predominant genres present in the DALI dataset: pop, rock and alternative.

The described pipeline can be seen in Figure 4.3.

4.2.4 Text & Audio to Text to Image to Text

The text & audio to text to image to text generation process consists of an extra step compared to the previous setup, where there is an image generation based on a description generated from the music understanding model. The lyrics are then produced from the image.

4.2.4.1 SALMONN-Stable Diffusion-LLaVA: adding other modalities in the lyrics generation process

With this model, we wanted to test if the creativity of the lyric generation process would increase by adding the vision modality. The pipeline of this model consists of the following modules: first, we feed the music accompaniment to SALMONN, and we prompt it to describe a paused movie scene that could be accompanied by the given music accompaniment. Given that this model has been trained to respond sufficiently to creative tasks, the output of the model is satisfactory. The output of the model is then fed to Claude, which is prompted to modify the given description as necessary, so it can be used as a Stable Diffusion prompt. Afterwards, Claude's response is fed into Stable

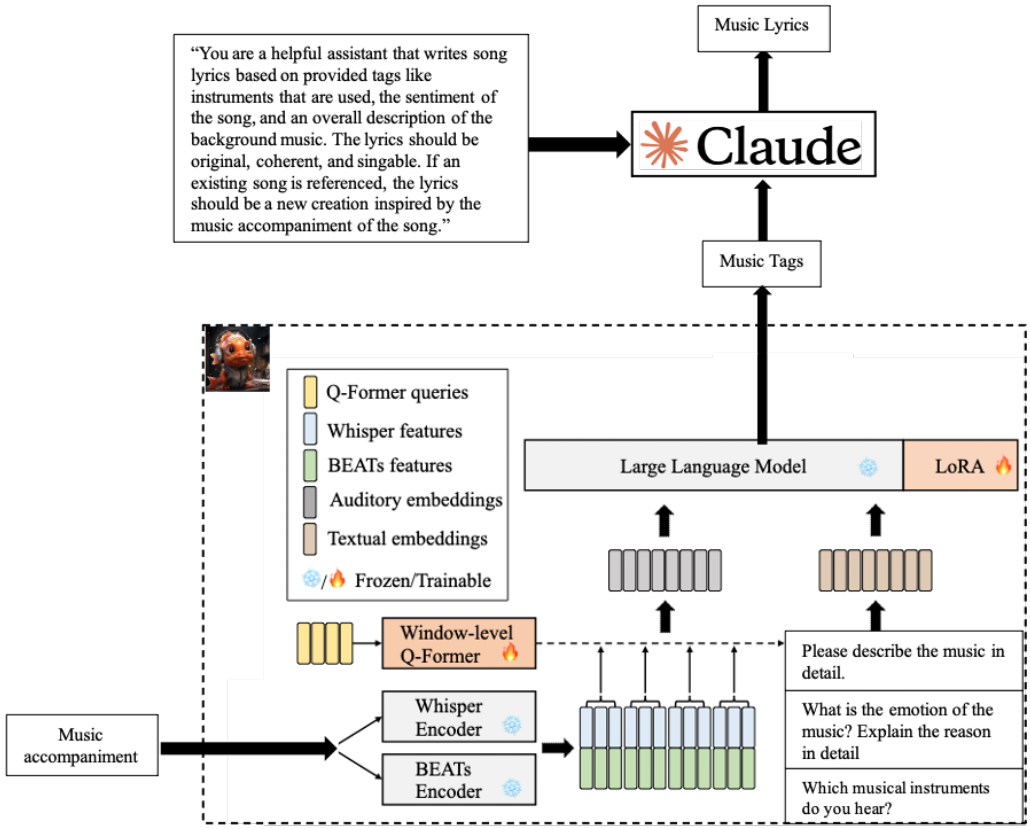


Figure 4.3: *The Text & Audio to Text to Text pipeline. The dashed line highlights the SALMONN model. Adapted from [2], [3] and [4]*

Diffusion, which generates an image, and this image is finally fed into the LLaVA model, which is prompted to generate lyrics based on the given image.

We additionally test the performance of this pipeline by doing few-shot learning in the LLaVA model. The prompts can be viewed below:

Prompt for the SALMONN model in SALMONN-Stable Diffusion-LLaVA pipeline

```
Describe a paused movie scene that would be accompanied by this music. Analyze
shortly the setting, characters, and plot. The description should be inspired
by the mood and atmosphere of the music.
```

Prompt for the Claude model in SALMONN-Stable Diffusion-LLaVA pipeline

```
You are a helpful assistant that turns descriptions of movie scenes into Stable
Diffusion prompts, removing any references to music.
Description: {description}
Stable Diffusion prompt:
```

Zero-shot prompt for the LLaVA model in SALMONN-Stable Diffusion-LLaVA pipeline

```
Create song lyrics that match the atmosphere and overall sentiment depicted in
this image.
```

Few-shot prompt for the LLaVA model in SALMONN-Stable Diffusion-LLaVA pipeline

```
Create song lyrics that match the atmosphere and overall sentiment depicted in
this image.  Some examples of lyrics are:
Example 1:  {lyrics1}
Example 2:  {lyrics2}
Example 3:  {lyrics3}
Example 4:  {lyrics4}
Example 5:  {lyrics5}
Example 6:  {lyrics6}
```

The described pipeline can be seen in Figure 4.4.

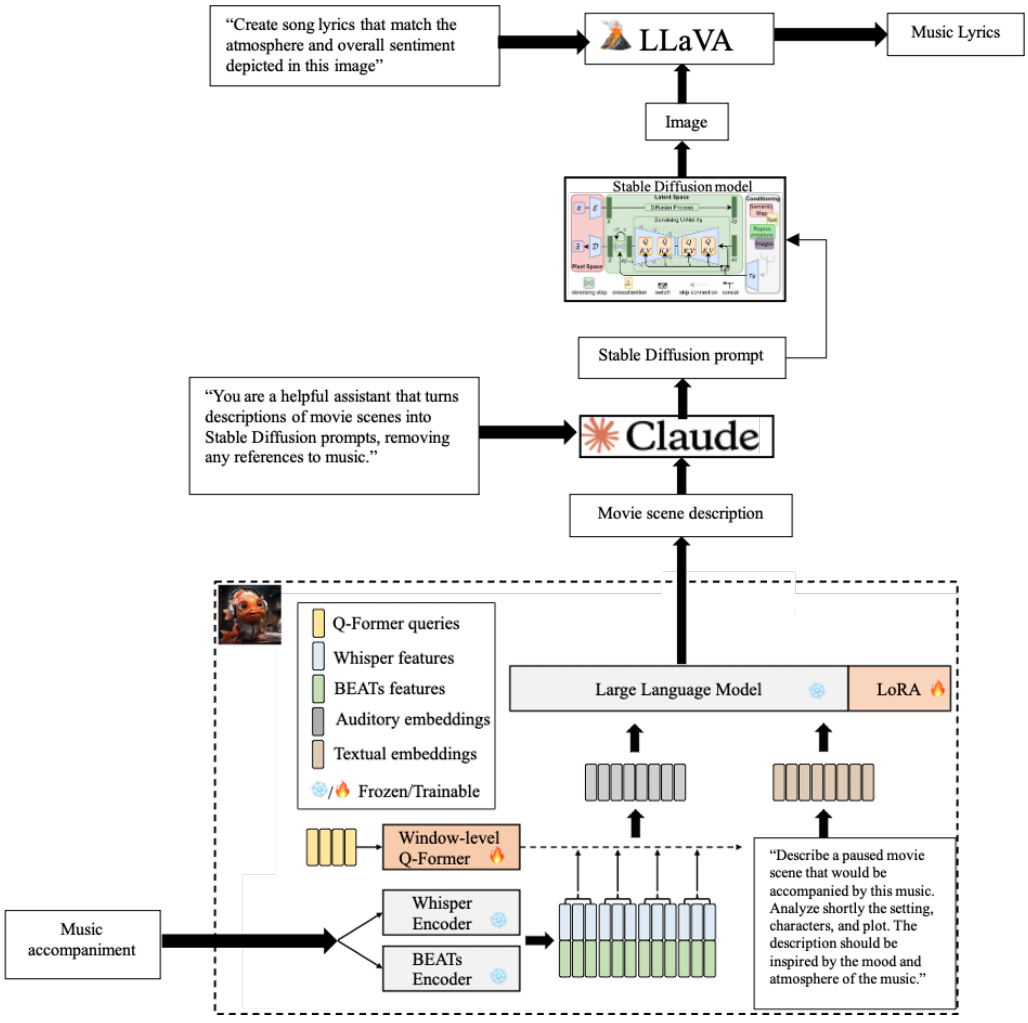


Figure 4.4: *The Text & Audio to Image to Text pipeline. The dashed line highlights the SALMONN model. Adapted from [2], [3], [4] and [5]*

Furthermore, we tested a configuration as an ablation study for this pipeline. Given that the LLaVA model is based on the Vicuna-7b-v1 model, we omitted the vision modality, and we got the SALMONN-Vicuna pipeline. In this pipeline, the SALMONN model gives the movie scene description, and then the Vicuna model is prompted to generate the lyrics

given the description from the SALMONN model. The prompt used for the Vicuna model, in order to get the lyrics from the movie scene description, is shown below:

Prompt for the Vicuna model in the SALMONN-Vicuna pipeline

```
USER: Create music lyrics based on the following description.  Avoid
referencing the style of music in your lyrics:  {movie_scene_description}
ASSISTANT:
```

Κεφάλαιο 5

Experiments and Results

Building upon the methodological framework presented in chapter 4, this chapter details our experimental implementation and presents our findings. We first describe the specific configurations and training processes for each approach, followed by a comprehensive analysis of our results.

5.1 Implementation Details

Following our methodological framework, we implemented and evaluated four distinct approaches to lyric generation. For each approach, we detail the specific configurations and resources required for reproducibility.

5.1.1 Text to Text Implementation

Our initial experimental approach focused on text-to-text architectures, where we systematically evaluated four language models chosen to represent different architectural paradigms and parameter scales. The selected models were Claude 3.5 Sonnet, GPT-2, Mistral OpenOrca (7B parameters), and Vicuna 1.5 (7B parameters). This diverse selection enabled us to examine how different model architectures and scales influence the quality of lyric generation, while also providing insights into the trade-offs between model complexity and generation capability.

For the initial inference phase, we established consistent generation parameters across all models, setting the temperature to 0.6 and maximum token length to 512. These values were determined through preliminary experimentation, where we found that a temperature of 0.6 provided an optimal balance between creativity and coherence in lyric generation, while 512 tokens sufficiently accommodated the typical length of song continuations.

Following the initial evaluation, we proceeded with fine-tuning two of the models—GPT-2 and Mistral OpenOrca—using the Low-Rank Adaptation (LoRA) technique. LoRA was selected for its efficiency in adapting large language models while maintaining reasonable computational requirements, a crucial consideration for our research scope. The fine-tuning configurations were tailored to each model's architecture:

For GPT-2, we implemented LoRA with a rank of 16 and alpha of 32, targeting the

primary linear layers (c_attn, c_proj, c_fc, and lm_head). This configuration was chosen based on empirical testing and represented an effective balance between adaptation capacity and computational efficiency. Similarly, for Mistral OpenOrca, we employed a higher rank of 32 and alpha of 64, targeting an expanded set of linear layers due to the model's more complex architecture. Both configurations utilized a dropout rate of 0.05 to prevent overfitting while maintaining effective learning.

The training process for both models employed consistent hyperparameters: a batch size of 8, learning rate of 2.5e-5, and the 8-bit AdamW optimizer with paged optimization. These parameters were selected to ensure stable training while managing memory constraints. The inclusion of two warm-up steps proved sufficient to stabilize the initial training phase without significantly impacting the overall training duration.

This systematic approach to model selection and parameter tuning allowed us to comprehensively evaluate the effectiveness of different text-to-text architectures in the context of lyric generation, while maintaining experimental rigor and computational feasibility.

5.1.2 Text & Audio to Text Implementation

Building upon our text-only approach, we explored multimodal architectures that could leverage both textual and audio information for lyric generation. This investigation centered on two distinct approaches: a VAE-based architecture combining AST and GPT-2, and an integrated system utilizing Whisper and OpenOrca. These architectures were selected to explore different paradigms of multimodal integration in the context of lyric generation.

For the VAE AST-GPT2 implementation, we developed a training regime spanning 40 epochs with a batch size of 32. The choice of the Adam optimizer with a learning rate of 5e-5 was informed by the model's complexity and the need for stable convergence in multimodal learning. A key aspect of our implementation was the management of the KL weight (β), which we carefully scheduled throughout the training process. Initially set to 1e-5 for the first half of training, β was linearly increased to 1 during the second half. This progressive scheduling strategy was crucial for maintaining the delicate balance between the variational bottleneck and reconstruction quality, while preventing posterior collapse—a common challenge in VAE architectures.

Our implementation strategy for this architecture focused on efficient parameter updating. Through careful analysis of the original architecture, we identified that the cross-attention layer and reparameterization weight matrices were the critical components requiring adaptation. Consequently, we maintained the pretrained weights of both the AST and GPT-2 components, focusing our training exclusively on these interaction layers. This targeted approach allowed us to preserve the robust feature extraction capabilities of the pretrained models while optimizing their integration for lyric generation.

The Whisper-OpenOrca implementation required a different approach due to the distinct characteristics of the Whisper model. During training, we employed the Adam optimizer with a higher learning rate of 1.5e-3, conducted over 20 epochs. This configuration was determined through empirical testing to best accommodate the model's architec-

ture and our training objectives. A significant consideration in this implementation was Whisper's fixed 30-second input window. To address this constraint, we developed a preprocessing pipeline that intelligently segmented and aligned the DALI dataset entries. Our solution involved concatenating lyric lines to optimally fit the 30-second windows, with appropriate padding for shorter segments. This resulted in training pairs consisting of 30-second current lyric segments paired with their complete preceding lyrical context.

This structured approach to multimodal integration allowed us to effectively combine audio and textual information while managing the inherent complexities of each architecture. The distinct implementations provided valuable insights into different strategies for multimodal lyric generation, while maintaining computational feasibility and training stability.

5.1.3 Text & Audio to Text to Text Implementation

Our investigation into more complex multimodal chains led us to explore a two-stage processing pipeline combining SALMONN-7B [25] and Claude 3.5 Sonnet [36]. This architecture was designed to leverage both audio and textual information in a sequential manner, allowing for progressive refinement of the generated lyrics.

In implementing this pipeline, we carefully calibrated the temperature parameters for each model: 1.0 for SALMONN and 0.7 for Claude. The selection of these values emerged from extensive preliminary testing, where we found that a higher temperature for SALMONN promoted more diverse initial representations of the audio-textual features, while the moderately lower temperature for Claude helped maintain coherence in the final lyric generation phase. This configuration created an effective balance between creative exploration and semantic consistency, addressing one of the key challenges in multimodal lyric generation.

5.1.4 Text & Audio to Text to Image to Text Implementation

Building upon our previous pipeline architecture, we developed an extended multimodal chain incorporating visual modality through the integration of Stable Diffusion 2 [62] and LLaVA-v1.5-7b [29]. This novel approach combined the strengths of SALMONN-7B's audio-textual processing with the visual-semantic capabilities of Stable Diffusion and LLaVA, creating a comprehensive multimodal system for lyric generation.

The implementation maintained the previously established temperature parameters for SALMONN (1.0) and Claude (0.7), while introducing a carefully selected temperature of 0.6 for the LLaVA model. This configuration was chosen after extensive experimentation, finding that a lower temperature for LLaVA provided more consistent and contextually relevant interpretations of the generated images, which proved crucial for maintaining semantic coherence throughout the extended pipeline.

The integration of visual processing into the pipeline presented unique challenges in maintaining semantic consistency across modalities. Our implementation focused on creating a balanced flow of information, where each stage of the pipeline contributed meaningfully to the final lyric generation process. The careful selection of model param-

eters and processing sequence helped ensure that the visual modality enhanced rather than disrupted the lyric generation process, providing additional semantic context while maintaining musical and linguistic coherence.

5.2 Computational Requirements and Resource Analysis

An important aspect of our research is understanding the computational demands and efficiency of different approaches to lyric generation. This analysis not only provides practical insights for future implementations but also helps contextualize the scalability and accessibility of various architectures. In this section, we present a detailed analysis of the computational requirements across our implemented models, examining both training and inference characteristics.

5.2.1 Parameter Efficiency Analysis

Table 5.1 presents the parameter distribution across our trained models, revealing interesting patterns in architectural efficiency. The parameter utilization varies significantly across architectures, with the VAE-AST-GPT2 showing the highest proportion of trainable parameters (10.46%). This higher percentage reflects the more extensive adaptation required for effective multimodal integration in this architecture. In contrast, the Whisper-OpenOrca implementation achieves its objectives with remarkably few trainable parameters (0.12%), demonstrating the efficiency of its transfer learning approach.

The GPT-2 and OpenOrca implementations, utilizing LoRA fine-tuning, maintain similar proportions of trainable parameters (2.49% and 2.22% respectively) despite their vastly different scales. This consistency validates LoRA's effectiveness in maintaining parameter efficiency across model scales, while still allowing sufficient adaptation for our specific task.

5.2.2 Training Resource Requirements

The training requirements, detailed in Table 5.2, reveal the varying computational demands of different architectures. The contrast between training durations is particularly noteworthy, ranging from 1.5 hours for GPT-2 to 90 hours for Whisper-OpenOrca. These differences reflect not just the computational complexity of each model, but also the challenges inherent in different approaches to multimodal integration.

The precision requirements also provide interesting insights. While some models operated efficiently with float32 precision, others required more sophisticated approaches like nf4 quantization with bfloat16 precision. These variations in precision requirements highlight the balance between computational efficiency and model performance, particularly in larger architectures like OpenOrca and Whisper-OpenOrca.

5.2.3 Inference Performance Analysis

The inference characteristics, presented in Table 5.3, demonstrate notable variations in processing efficiency across models. The text-to-text models (Claude and GPT-2)

achieved the fastest inference times at 16 minutes for 100 songs, while more complex pipelines like SALMONN-Stable Diffusion-LLaVA required up to 150 minutes for the same task. These differences in inference time reflect the computational overhead of processing multiple modalities and the complexity of inter-model communication in pipeline architectures.

Notably, the GPU memory requirements remain relatively consistent across most models at 16GB, with some variations for simpler architectures. The more complex multimodal pipelines, such as SALMONN-Stable Diffusion-LLaVA, also maintain the same 16GB memory requirement as simpler models, which is achieved through sequential processing of each modality. Despite handling multiple modalities and larger model architectures, these pipelines avoid increased memory requirements by efficiently loading and unloading components as needed. However, this memory efficiency trades off against processing time, as evidenced by the longer inference times for these models.

These resource requirements and performance characteristics provide valuable insights for future research and practical implementations in the field of automated lyric generation. They highlight the trade-offs between model complexity, computational efficiency, and generation quality, offering important considerations for both research and practical applications in this domain.

Model	No. of total parameters	No. of trainable parameters	Trainable parameters (%)
GPT-2 finetuned	127,615,504	3,175,696	2.49%
OpenOrca finetuned	3,837,128,768	85,041,216	2.22%
VAE-AST-GPT2	236,252,930	24,701,440	10.46%
Whisper-OpenOrca	4,392,381,952	5,245,440	0.12%

Table 5.1: *Trainable parameters for each trained model*

Model	GPU requirements	LLM Precision	Training time	Epochs
GPT-2 finetuned	16GB	float32	1.5 hours	6
OpenOrca finetuned	24GB	nf4 quantization - bfloat16 precision	7 hours	6
VAE-AST-GPT2	16GB	float32	24 hours	40
Whisper-OpenOrca	24GB	nf4 quantization - bfloat16 precision	90 hours	20

Table 5.2: *GPU requirements, precision, training time and number of epochs for the trainable models*

Model	GPU requirements	Inference time (for 100 songs)
Claude out-of-the-box	-	16 minutes
GPT-2 out-of-the-box	< 8GB	16 minutes
OpenOrca out-of-the-box	16GB	50 minutes
Vicuna out-of-the-box	16GB	50 minutes
SALMONN-Claude zero-shot	16GB	80 minutes
SALMONN-Claude few-shot	16GB	80 minutes
SALMONN-Stable Diffusion-LLaVA zero-shot	16GB	150 minutes
SALMONN-Stable Diffusion-LLaVA few-shot	16GB	150 minutes
SALMONN-Vicuna	16GB	95 minutes

Table 5.3: *GPU requirements and inference time for the non-trained models*

5.3 Our Evaluation Methods

During the inference, we evaluated the performance of the models on the test set. Because of the cost of some of the submodules of the last two pipelines, the evaluation was done on 100 songs. The evaluation consists, as analyzed in more detail in section 3.6, of the calculation of similarity score by employing a cross-encoder, and LLM methods like JudgeLM and prompt-based with a combination of strong and diverse LLMs. We give more details for each metric, and the models used for the LLM evaluations. We then present the results of these evaluations. We also employ a user study, in order to get the opinion of users regarding the models that scored best overall with the previous evaluation methods.

5.3.1 Semantic textual similarity

Recognizing the weakness of the BLEU score, we decided to use a method that calculates the similarity between the gold and the generated lyrics, but without the strict check of the n-gram overlaps. For this task, we utilized a cross-encoder model. Bi-encoders produce fixed-dimensional sentence representations and are computationally efficient, however, they usually underperform cross-encoders, which can leverage their attention heads to exploit inter-sentence interactions for better performance [49]. The specific model that was used for the calculation of the similarity scores is ms-marco-MiniLM-L-12-v2 [50].

The exact method of calculating the similarity score is taking the average for all the individually calculated scores for each generated-gold lyric pair. The similarity score can range from 0 to 1. The results of this method can be seen in the Table 5.4.

Model	Similarity Score
SALMONN-Stable Diffusion-LLaVA few-shot	0.89
Vicuna out-of-the-box	0.85
OpenOrca out-of-the-box	0.81
GPT-2 out-of-the-box	0.71
Claude out-of-the-box	0.69
OpenOrca finetuned	0.58
SALMONN-Claude zero-shot	0.51
GPT-2 finetuned	0.50
SALMONN-Stable Diffusion-LLaVA zero-shot	0.48
SALMONN-Vicuna	0.36
Whisper-OpenOrca	0.26
SALMONN-Claude few-shot	0.25
VAE-AST-GPT2	0.07

Table 5.4: *The similarity scores calculated with the cross-encoder model*

5.3.2 Using LLM-as-a-Judge to evaluate lyrics quality

We also use the JudgeLM model, which has been analyzed in subsection 3.6.3. We give to the model that the task that the LLMs had to complete was to generate coherent and creative lyrics. Due to the numerous models and configurations tested, it was not feasible

to prompt JudgeLM to grade all outputs simultaneously, given its context length limitation (2048). To address this and incorporate a method that compares models before grading them—as single answer grading may not always discern subtle differences between specific pairs [43]—we employed JudgeLM for pairwise grading, resulting in 78 comparisons. The model, considering the gold lyrics as well, assigns a grade from 0 to 10 to each model in the pair based on lyrics quality. Using this method, we initially calculate the average rating for each pair. Then, we present a ranking order by aggregating all the scores that each model got in each pairwise comparison. The results of this evaluation can be seen in Table 5.5.

Model	**Aggregated Score**
SALMONN-Claude zero-shot	96.87
SALMONN-Claude few-shot	96.35
Claude out-of-the-box	95.91
OpenOrca out-of-the-box	92.80
SALMONN-Stable Diffusion-LLaVA zero-shot	92.53
Vicuna out-of-the-box	90.02
SALMONN-Vicuna	89.05
OpenOrca finetuned	49.76
GPT-2 finetuned	38.59
Whisper-OpenOrca	24.16
SALMONN-Stable Diffusion-LLaVA few-shot	16.26
GPT-2 out-of-the-box	15.20
VAE-AST-GPT2	13.36

Table 5.5: *The aggregated scores determined by JudgeLM, in descending order*

5.3.3 Prompt-based LLM evaluation

Following a similar method to the one of the paper "Reference-Guided Verdict: LLMs-as-Judges in Automatic Evaluation of Free-Form Text" [44] that was also explained in subsection 3.6.4, we used three (one closed-source, two open-source) LLMs: Claude 3.5 Sonnet, Mistral-Instruct-7B-v0.3 [51], and Llama-3.1-Instruct-70B [52]. We prompted them to evaluate the generated lyrics, by giving a grade of 0 to 10, for each of the following criteria: coherence, creativity, singability, naturality. Afterwards, the given grades were averaged programmatically to extract the score for each song, and then averaged again to get the overall scores given by each LLM-as-a-judge to each model. The specific prompt that was used for this task is the following:

System prompt for prompt-based evaluation

```
You are a helpful assistant that judges the quality of generated lyrics. You
have to give a score from 0 to 10 for each of the following criteria:
Coherence: How well the lyrics make sense and are logically connected.
Creativity: How original and imaginative the lyrics are.
Singability: How easy it is to sing the lyrics.
Naturality: How natural and fluent the lyrics sound.
The form of your answer should be of the form
"Coherence:...\nCreativity:...\nSingability:...\nNaturality:...".
If the given text does not resemble lyrics, provide low scores accordingly.
```

The LLMs were set to low temperatures (0.2-0.4) for this evaluation task to avoid very unpredictable answers. The final results of this evaluation method can be seen in Table 5.7.

Model	Claude	LLaMA	Mistral	Average
SALMONN-Stable Diffusion-LLaVA zero-shot	8.19	8.31	8.50	8.33
SALMONN-Claude zero-shot	8.43	8.15	8.20	8.26
Claude out-of-the-box	8.31	8.08	8.30	8.23
OpenOrca out-of-the-box	7.87	7.89	8.61	8.12
SALMONN-Claude few-shot	8.26	7.80	8.21	8.09
Vicuna out-of-the-box	7.68	7.63	8.54	7.95
SALMONN-Vicuna	7.54	7.34	8.29	7.72
OpenOrca finetuned	5.89	6.74	7.31	6.64
Whisper-OpenOrca	4.40	5.27	6.38	5.35
GPT-2 finetuned	3.66	5.56	6.65	5.29
SALMONN-Stable Diffusion-LLaVA few-shot	2.96	2.91	6.34	4.07
GPT-2 out-of-the-box	1.83	1.36	5.55	2.91
VAE-AST-GPT2	0.35	0.87	1.53	0.92

Table 5.7: *The LLM evaluations, ordered in descending order of average grades*

5.4 User Study: Ranking our Methods with Human Annotations

Given the inherently creative nature of lyric generation, we believe it is crucial to incorporate human judgment alongside automated evaluation methods. While our user study is conducted on a smaller scale compared to other evaluation techniques, it provides valuable insights into the perceived quality and creativity of the generated lyrics from an end-user perspective.

We structured our user study as follows: Participants were first asked to listen to 1 minute of musical accompaniment for each song. They were then presented with two candidate sets of lyrics and asked to choose, first in terms of the quality of the lyrics (coherence, structure) and then in terms of correlation between lyrics and the given music.

The study encompassed 12 songs in total, and we examined only the models that scored the highest with the other evaluation methods, for each set of modalities, which are Claude out-of-the-box, Whisper-OpenOrca, SALMONN-Claude zero-shot and SALMONN-Stable Diffusion-LLaVA zero-shot. To establish a credible ranking of our models, each music featured all six possible pairs of candidate lyrics.

To prevent participant fatigue and maintain engagement, we implemented a random assignment method for the evaluation pairs. Each participant was tasked with grading 12 pairs in total, meaning that the user study had 6 groups of questions, which was randomly assigned each time the user study form is opened. This randomization ensured that every participant evaluated two instances of each pair of models, and one pair of models for each song, striking a balance between comprehensive coverage and a manageable workload for the annotators.

The method that was used to calculate the ranking of the models from this user study is the Bradley-Terry model [53], which is a probability model used to predict the outcome of pairwise comparisons and has found widespread application in various fields, including the ranking of AI models [54]. In the Bradley-Terry model, each candidate (or item) is assigned a strength parameter π_i, for candidate i. The model assumes that the probability that candidate i beats candidate j in a pairwise comparison is:

$$P(i \, beats \, j) = \frac{\pi_i}{\pi_i + \pi_j}.$$

To rank the candidates, the Bradley-Terry model uses maximum likelihood estimation (MLE) to find the best values for each π_i that maximize the likelihood of observing the pairwise comparison data. The likelihood is a function of the observed wins and losses between all pairs of candidates. Given the pairwise outcomes, the model tries to find the π_i values that best explain the data. This typically requires solving the following equation for each candidate i:

$$log(\pi_i) = \sum_{j \neq i} \frac{W_{ij}}{\pi_i + \pi_j},$$

where W_{ij} is the number of times candidate i beat candidate j. After the estimation of the π_i values, the candidates with the highest π values are ranked highest.

The results, which were obtained by the participation of 28 users, can be seen in Table 5.8 and Table 5.9.

Model	Bradley-Terry Probabilities for Coherence and Structure
SALMONN-Stable Diffusion-LLaVA zero-shot	0.316393
Claude out-of-the-box	0.304484
SALMONN-Claude zero-shot	0.282116
Whisper-OpenOrca	0.097007

Table 5.8: *The Bradley-Terry probabilities for the coherence and structure of the lyrics, ordered in descending order*

We also performed statistical analysis on the results of the user study, by employing the z-test and calculating the p-values for each pair of models. The Z-score measures how far the observed win proportion is from the expected proportion under the null

Model	Bradley-Terry Probabilities for Music Correlation
SALMONN-Claude zero-shot	0.355462
SALMONN-Stable Diffusion-LLaVA zero-shot	0.241730
Claude out-of-the-box	0.228919
Whisper-OpenOrca	0.173889

Table 5.9: *The Bradley-Terry probabilities for the correlation of the lyrics with the music, ordered in descending order*

hypothesis. The null hypothesis assumes that both models are equally likely to win (i.e., a 50-50 split of wins). When the magnitude of the Z-scores is close to zero (around 0 to ±1), they suggest the results are not far from the null hypothesis, and z-scores further away from zero (e.g., greater than ±1.96 for a 95% confidence level and greater than ±1.645 for a 90% confidence level) suggest a statistically significant deviation from the null hypothesis [55]. Regading the p-value, a small p-value (typically $p < 0.05$) suggests that the observed difference in wins is statistically significant, meaning it's unlikely to have occurred by chance, and we might reject the null hypothesis (i.e., the models are not equally preferred). A large p-value ($p > 0.05$) suggests that the observed difference in wins is not statistically significant, meaning we fail to reject the null hypothesis (i.e., the models are equally preferred) [56]. The statistical analysis of the results (z-test and p-values) for each pair of models can be seen in Table 5.10 and Table 5.11.

Model Pair	Z-score (magnitude of value)	p-value
Claude-out-of-the-box & Whisper-OpenOrca	3.21	0.00134
Claude-out-of-the-box & SALMONN-Claude	1.07	0.285
Claude-out-of-the-box & SALMONN-Stable Diffusion-LLaVA	0.27	0.789
Whisper-OpenOrca & SALMONN-Claude	4.81	1.50e-06
Whisper-OpenOrca & SALMONN-Stable Diffusion-LLaVA	3.47	0.00051
SALMONN-Claude & SALMONN-Stable Diffusion-LLaVA	0.80	0.423

Table 5.10: *The z-test and p-values for each pair of models, for the criterion of structure/-coherence*

Model Pair	Z-score (magnitude of value)	p-value
Claude-out-of-the-box & Whisper-OpenOrca	1.60	0.109
Claude-out-of-the-box & SALMONN-Claude	1.87	0.061
Claude-out-of-the-box & SALMONN-Stable Diffusion-LLaVA	0.53	0.593
Whisper-OpenOrca & SALMONN-Claude	2.67	0.0075
Whisper-OpenOrca & SALMONN-Stable Diffusion-LLaVA	0.53	0.593
SALMONN-Claude & SALMONN-Stable Diffusion-LLaVA	1.07	0.285

Table 5.11: *The z-test and p-values for each pair of models, for the criterion of music-lyric correlation*

With this statistical analysis, we can say that, for the criterion of structure/coherence, Whisper-OpenOrca appears to perform worse compared to the other models, particularly when compared to SALMONN-Claude and SALMONN-Stable Diffusion-LLaVA. On the other hand, SALMONN-Claude and SALMONN-Stable Diffusion-LLaVA are performing similarly. For the criterion of music-lyric correlation, SALMONN-Claude appears to significantly outperform Whisper-OpenOrca. The comparison between Claude out-of-the-box

and SALMONN-Claude shows statistical significance with 90% - instead of 95% - confidence, indicating that SALMONN-Claude slightly outperforms Claude out-of-the-box. The other pairs show similar performance.

Κεφάλαιο 6

Conclusion

6.1 Discussion

This thesis explores the task of automatic lyric generation process and how the quality of the produced lyrics is affected with the addition of different modalities. To our knowledge, the incorporation of the modality of vision, along with text and audio, has not been explored.

We explored four sets of different usage of modalities. The first one was text to text, in which we tested how four LLMs would handle the lyric generation task, first by using them out-of-box, then by fine-tuning them in the DALI dataset. This showed that the instruction LLMs did well even without fine-tuning, which is due to their rich pretraining on creative writing texts like songs and poems. In the user study, the selected LLM for the evaluation of the 'text to text' combination was Claude, which showed the best performance in the LLM evaluations. This model performed fairly well in both criteria of the user study, but didn't particularly excel in either.

The second combination of modalities that we explored was by also adding music supervision (text & audio to text). For this, we explored two models: one was in the context of reproducing previous work that used a Variational Autoencoder model, with transformer-like architecture, that incorporated the Audio Spectrogram Transformer as audio encoder, and GPT-2 as the decoder. This model, which we tried to reproduce as faithfully as the paper presented it, was the weakest one of all we studied. We attribute this to the fact that code, the specifics of the preprocessing of the dataset, and the IDs of the subset of the songs of the dataset that they used for the training were not available. Additionally, GPT-2 is an older and weaker model compared to the other LLMs that we used in our other models. Given that in the task of lyric generation, the quality of lyrics is mainly affected by the performance of the LLM, and the other modalities are used to enhance the lyric generation, the use of weaker LLMs leads to lower quality of lyrics.

The other configuration that we tested for the combination of text & audio to text modalities, is the Whisper-OpenOrca model, which uses the Whisper audio encoder, the Mistral OpenOrca LLM, and a trainable projection layer between these two modules, which aligns the music representation with the lyric generation. This model did much better than the other model under this combination of modalities. This can be attributed to the fact that both the audio encoder and the LLM are more recent, and stronger, models,

meaning they exploit their diverse knowledge from their extensive pretraining. In the user study, the Whisper-OpenOrca model underperformed in both criteria in comparison to the other models included in the study, which reflects the LLM evaluations.

The third combination of modalities that we studied added an intermediate production of text compared to the previous one (text & audio to text). This combination was tested with the SALMONN-Claude model, with which we extracted music tags relating to the given audio. Then, these tags were given to the Claude model to generate lyrics. This addition was shown to score slightly better compared to the text to text setup with Claude, which serves as an ablation study of the audio modality in the pipeline. It also scored significantly better than the other 'text to text' and 'text & audio to text' setups, showcasing the ability of this proposed pipeline to enhance the lyric generation process. From the user study, we also observe that this pipeline achieved the best score regarding lyrics and music correlation, which further supports the claim that the addition of the step of music tag extraction enhances the lyric generation process.

The fourth combination of modalities additionally incorporated the vision modality to the third setup (text & audio to text to vision to text). This was explored with the SALMONN-Stable Diffusion-LLaVA model. The SALMONN model generated a description of a movie scene that could accompany the given audio, then this audio was turned, with the help of Claude, to a Stable Diffusion prompt, which was used to visualize that scene. The image was then given to the LLaVA model to generate the final lyrics. The addition of the vision modality proved to increase the quality of the lyrics, and its performance was on par with the previous model, proving that the addition of extra modalities to the lyric generation process can enhance the creativity and therefore the lyric generation process. This is further confirmed by the performance of the the SALMONN-Vicuna model, which serves as an ablation study for the vision modality in the SALMONN-Stable Diffusion-LLaVA model, and the Vicuna out-of-the-box model, which serves as an ablation study for both the audio and image modalities. These two models score significantly less than the proposed pipeline. Additionally, the user study showed that the SALMONN-Stable Diffusion-LLaVA model scored slightly better than the other models in the coherence and structure of the lyrics, while also maintaining a reasonable correlation with the music, striking a balance between these two criteria.

For the two previous setups, in addition to zero shot prompting, we tried few shot prompting. Specifically, in the SALMONN-Claude model, we incorporated some examples of music tags and gold lyrics to the prompt of the Claude model, and in the SALMONN-Stable Diffusion-LLaVA model, the few shot prompting was incorporated in the LLaVA model, by giving examples of gold lyrics. One of the few shot setups showed an increase in the similarity score, however both few shot setups showed a decrease in performance with the LLM-based evaluations. This is due to the fact that, with the few-shot setup, the models were guided to produce similar lyrics to the given examples, but the creativity of the produced lyrics was decreased.

These conclusions were drawn mostly from the LLM evaluations. Given that, even human composers would come up with different lyrics for a given music, or they would continue a song differently given its first line, we think that the similarity scores and

objective metrics are less important than the LLM evaluations and the user study, which give a better view of the quality of the lyrics. Given that we used a method that mitigates the bias of LLMs for the LLM evaluations, we trust these results more. Furthermore, the user study demonstrated a correlation between LLM evaluations and human judgments, and specifically confirming that the Whisper-Mistral model is significantly worse, while the other models of the user study have similar performance. This suggests that LLM evaluations serve as a reliable proxy for assessing the quality of the lyrics.

In conclusion, this thesis contributes to the advancements of generative tasks in the music information retrieval. Given that automatic lyric generation is a field that has not been studied this extensively, and specifically in the direction that we followed, compared to other music information retrieval tasks, we brought new ideas to this field, which can encourage other researchers to study and contribute to this domain.

In summary, our key findings are:

- Instruction-tuned LLMs demonstrate strong baseline performance in lyric generation even without domain-specific training, particularly evident in Claude's performance.
- The incorporation of music supervision through tags (SALMONN-Claude) significantly improves the correlation between generated lyrics and music, compared to pure text-to-text approaches.
- Our novel multimodal pipeline (SALMONN-Stable Diffusion-LLaVA) achieves the best balance between lyrical coherence and musical correlation, suggesting that thoughtfully integrated multiple modalities can enhance creative generation.
- Few-shot prompting shows a trade-off between similarity metrics and creative quality - while it improves similarity scores, it tends to decrease overall creative performance.
- LLM-based evaluation methods show strong correlation with human judgments, suggesting their validity as assessment tools for creative text generation tasks.
- The quality of the base LLM significantly impacts the final output quality, as demonstrated by the performance difference between older models (GPT-2) and newer architectures.
- The addition of intermediate steps (like tag extraction or scene visualization) in the generation pipeline can enhance the final output quality while maintaining creative freedom.

6.2 Limitations and Future Work

Our research was constrained by the computational resources at our disposal. For the majority of our work, we relied on free GPU resources provided by platforms such as Google Colab and Kaggle, as well as the SLP-NTUA lab's server equipped with two 12GB

GPUs (NVIDIA GeForce GTX 1080 Ti and GeForce GTX TITAN X). For more demanding models and stronger LLMs, we selectively utilized on-demand AWS resources.

These limitations impacted the performance of our trained models in several ways. Specifically, training methods were restricted due to computational constraints. We predominantly used lighter models and the scope of experimentation with larger, more resource-intensive models was limited. The use of stronger models would have maybe further increased the quality of the results. Other music understanding models, like MU-LLaMA [26], M^2UGen [27] and MusiLingo [10] could not be experimented with in the third configuration of modalities, for the generation of the music tags, because of their more computationally demanding nature.

Another limitation that doesn't only affect our work, but the lyric generation domain in general, is the limited available datasets. There are many music datasets that are annotated with tags and metadata (e.g. the Million Song Dataset [57]), but not with aligned music lyrics. To our knowledge, the DALI dataset is currently the only available dataset, of this size, with songs as MP3 files, with aligned lyrics. Other available aligned datasets, like the Lakh MIDI dataset [58], have used MIDI files instead of MP3 files. Additionally, the majority of the entries of the DALI dataset are pop songs, making the trained models more likely to limit their output in themes presented in pop music, such as love, which restricts the creativity and diversity in their outputs.

In future research, we believe it would be beneficial to explore the incorporation of the other music understanding models that were mentioned, as well as other strong LLMs in the lyric generation process. Moreover, the fine-tuning of the submodule of the lyric generation process in the 'text & audio to text to vision to text' setup would be an interesting study, or even incorporating a Vision Language model with architecture that supports few shot learning with images, like the Flamingo model [59]. Additionally, another extension of the task that we studied would be to generate lyrics conditioned on artist or genre, and explore the differentiation of the outputs given the same music accompaniment but different artist or genre. Finally, another extension to this thesis would be to train a model to additionally generate melody that would suit the generated lyrics, similarly to "Lyrics and Vocal Melody Generation conditioned on Accompaniment" [60], with the difference that this work was done with symbolic music (MIDI) rather than audio.

Appendices

Appendix A

Examples of Generated Lyrics

In this appendix, we show the results of some of the models that we tested, specifically for Whisper-OpenOrca, OpenOrca out-of-the-box, Vicuna-out-of-the-box, Claude out-of-the-box, SALMONN-Claude, SALMONN-Vicuna and SALMONN-Stable Diffusion-LLaVA. Specifically, we have provided the lyrics that were generated by these models, given two different music accompaniments.

Lyrics given the music accompaniment of "Are We the Waiting" by Green Day:

Whisper-OpenOrca model:

never thought you would haunt me
with each passing day
never thought you would haunt me
with your words

never thought you to haunt me
nothing but tequila
my brother in arms
pine away for love like a drunken sailor
lost at sea over and over again

thirty years waisting away
for nothing but dreams of you
in my night dream
hefty salvage gave me twenty grand
had me drink it down nice and slow

do not turn that plow
hit the ventana and get out there
do not start a war nobody wants to fight
try to stay alive for another night

OpenOrca out-of-the-box model:

Starry nights filled with dreams and desire,
Whispering secrets under the moonlit sky,
As we dance beneath the twinkling stars,
Feeling the magic in the air.

In the glow of a million candles,
We make promises that never fade away,
For this love will last forever,
Burning bright like the stars above.

As we gaze into each other's eyes,
The world around us fades away,
Our hearts entwine in this celestial dance,
And the stars guide us through the darkest nights.

In the warmth of this embrace,
We find solace and serenity,
Together, we shine like the stars above,
And our love will never die.

Vicuna-out-of-the-box model:

Verse 1:
Starry nights, I see your face
Illuminated by the light of the moon
Your eyes shine bright, like a star in the sky
I'm drawn to you, like a moth to a flame

Chorus:
Oh, starry nights
You're the one I need
In the darkness, you're my guiding light
Starry nights, you're my everything

Verse 2:
We lay down on the grass
Watching the stars as they dance in the sky
Our hearts beating as one
Feeling alive, like we could fly

Chorus:
Oh, starry nights
You're the one I need
In the darkness, you're my guiding light

Starry nights, you're my everything

Bridge:
We're like two shooting stars
Flying through the night
Our love shining bright, like a supernova
We'll be together, forever and ever

Chorus:
Oh, starry nights
You're the one I need
In the darkness, you're my guiding light
Starry nights, you're my everything

Claude out-of-the-box model:

Starry nights, a canvas in the sky
Twinkling dreams, painted way up high
Constellations tell their ancient tales
As we wander down moonlit trails

Chorus:
Oh, these starry nights
Fill our hearts with wonder
Celestial lights
Pull us closer, pull us under

Whispered wishes on shooting stars
Hopes and dreams traveling so far
Hand in hand, we count the glowing spheres
Time stands still, and magic appears

(Repeat Chorus)

Milky Way stretches out above
A cosmic dance of light and love
In this moment, we feel so small
Yet connected to it all

Bridge:
Universe, vast and wide
Holds secrets still untried
But here beneath the starry dome
We've found our way, we've found our home

(Repeat Chorus)

Starry nights, they fade with dawn's first light
But memories linger of this perfect night
Till dusk returns and day departs
We'll keep the stars inside our hearts

(Repeat Chorus)

Starry nights, a canvas in the sky
Endless wonder, never saying goodbye

SALMONN-Claude model:

Verse 1:
Strumming chords in the fading light
Memories echo through the night
The guitar weeps a lonesome tune
As I sit here, missing you

Chorus:
Silent strings, they tell our tale
Of love that couldn't prevail
In this instrumental serenade
I hear the promises we made

Verse 2:
Each note falls like a teardrop
On this heart that just won't stop
Beating for what used to be
A melody of you and me

(Repeat Chorus)

Bridge:
The solo cries out in pain
A wordless story of love's remains
In every bend and every slide
I feel the emotions I can't hide

Verse 3:
As the final chord fades away
I'm left with what I cannot say
This song without words speaks true
Of how I'll always remember you

(Repeat Chorus)

Outro:
The guitar gently whispers your name
In this bittersweet refrain

SALMONN-Vicuna model:

Verse 1:
In a dark and ominous forest
Lost and scared, a group of travelers
Struggling to find a way out
Danger lurks in the shadows

Chorus:
Haunting melancholic music
Perfectly captures the mood
Fear and desperation
Elevating the tension

Verse 2:
The music adds to the atmosphere
Making the audience feel alive
In the midst of the treacherous terrain
The travelers fight to survive

Chorus:
Haunting melancholic music
Perfectly captures the mood
Fear and desperation
Elevating the tension

Bridge:
The music perfectly captures
The mood of the forest and the characters
Enhancing the emotional impact
Of this paused movie scene

Chorus:
Haunting melancholic music
Perfectly captures the mood
Fear and desperation
Elevating the tension

Outro:

The music adds to the atmosphere
Making the audience feel alive
In the midst of the treacherous terrain
The travelers fight to survive.

SALMONN-Stable Diffusion-LLaVA model:

Verse 1: In the midst of a misty glade,
Where the trees stand tall and the ground is made,
A path that leads to the unknown,
Where the whispers of the wind are alone.

Chorus:
In the heart of the mist,
Where dreams and reality blend,
A journey through the fog,
Where the secrets of the forest are told.

Verse 2:
Tents of solitude pitch at night,
Where the stars are hidden by the morning light,
A place where the soul can find its place,
In the quiet whispers of the forest's grace.

Chorus:
In the heart of the mist,
Where dreams and reality blend,
A journey through the fog,
Where the secrets of the forest are told.

Bridge:
The leaves rustle like a symphony,
In this place where time seems to freeze,
The air is thick with mystery,
A place where the heart finds its ease.

Chorus:
In the heart of the mist,
Where dreams and reality blend,
A journey through the fog,
Where the secrets of the forest are told.

Outro:
So if you feel the need to escape,

To find a place where you can retrace,
Take the path through the woodland's shade,
And let the forest be your guide.

Chorus:
In the heart of the mist,
Where dreams and reality blend,
A journey through the fog,
Where the secrets of the forest are told.

Lyrics given the music accompaniment of "Yellow Boomerang" by Middle of the Road:

Whisper-OpenOrca model:

Everyone here is one family
Lift your heart up and fly so high with the birds
So trust me and believe
Feel the nature so it is wake you up this very night
Trampling over me
Everyone here is on the make
You want to ask why are you running like a race
Leave it all behind
Take your time it is over now
So your night it is won now
Get up on your feet again

Hello love it is a friendly attack
That it is the point of telling off on the make
Come and dance along with my faithfulness
Hell it was one step away from touching you just a few

So you are the icing
So you are the cherry
So you are the walnut they leave out of every single little tree
Resist the serpent it bites the apple
You friendly serpent, i can see through you and think nothing of it
But just, just be careful I

So you are the icing
So you are the cherry
So you are the walnut they have taken out of every single little tree
Hey do not put another salesman on every single street
The truth in you are spreading
Move it will be seen as a sin on the day that you fall

By only two, by only two tonight
If you have something i am waiting
It it is going to be like a copyright in rewind, snip, snip
I believe you made out of gold
So come and dance along with my faithfulness
Because, indeed, i could not believe

OpenOrca out-of-the-box model:

You're like my yellow boomerang, always coming back to me
With your warm, sunny smile, it's like a summer day
You always land in my heart, never stray too far away

When I throw you into the sky, you never fail to return
You're the brightest shade of yellow, like sunshine in my life
You're my boomerang, always finding your way back to me

You're like my yellow boomerang, always coming back to me
With your carefree spirit, you're a breath of fresh air
In the darkest times, you're the light that guides my way

In the chaos of life, you're the calm that I adore
You're the reason I keep going, my yellow boomerang, my love.

Vicuna-out-of-the-box model:

Verse 1:
You're like my yellow boomerang,
Flying back to me again and again,
With every turn, I feel your grip,
And I'm reminded of the love we've been given.

Chorus:
You're my sunshine in the rain,
My shelter from the storm,
You're the one who makes me feel alive,
And I know that I can always count on you.

Verse 2:
You're my rock, my guiding light,
My everything, my everything tonight,
With every beat, my heart skips a beat,
And I know that I'm exactly where I'm meant to be.

Chorus:
You're my sunshine in the rain,
My shelter from the storm,
You're the one who makes me feel alive,
And I know that I can always count on you.

Bridge:
Like a boomerang, you never leave,
You're always there to lift me up and bring me peace,
With every spin, I'm reminded of our love,
And I know that we'll be together forever, like a boomerang.

Chorus:
You're my sunshine in the rain,
My shelter from the storm,
You're the one who makes me feel alive,
And I know that I can always count on you.

Claude out-of-the-box model:

You're like my yellow boomerang
Always coming back to me
No matter how far I throw you away
You return so faithfully

Chorus:
Round and round, you spin through the air
A curved path that leads you home
My bright spot on the darkest day
With you, I'm never alone

Verse 2:
When life tries to knock me down
You lift me up, help me stand tall
A loyal friend I can count on
To answer every call

(Repeat Chorus)

Bridge:
Through sunny days and stormy nights
You're the constant in my world
My yellow boomerang, my guiding light
Our bond will stay unfurled

(Repeat Chorus)

Outro:
You're like my yellow boomerang
Always finding your way back to me

SALMONN-Claude model:

Verse 1:
Cruising down this neon-lit street
Wah-wah guitar crying at my feet
Memories of you still burning bright
In the rearview mirror of the night

Chorus:
On this Heartbreak Highway
Where love's dreams collide
Wah-wah whispers your name
As I take this wild ride
Can't escape the rhythm
Of this classic rock and roll
On the Heartbreak Highway
Where I'm losing control

Verse 2:
Disco lights flash through my mind
Echoes of a love left behind
The radio plays our favorite tune
Reminding me of nights beneath the moon

(Repeat Chorus)

Bridge:
The strings of my heart
Strumming like this old guitar
We danced to the beat
Now I'm dancing alone, so far

Verse 3:
Years may pass, but this song remains
A timeless anthem of love's sweet pain
Through speakers old and new it'll play
Keeping our story alive today

(Repeat Chorus)

Outro:
On this Heartbreak Highway
Where love never dies
Wah-wah guitar sings the blues
As I say my goodbyes

SALMONN-Vicuna model:

Verse 1:
As the music starts to play,
I'm transported to a cozy café,
In the heart of Paris, so romantic and so sweet.
The main characters, a Frenchman and an American,
Sitting across from each other, deep in conversation.

Chorus:
Laughing and smiling, sharing stories of their travels,
The mood is exciting, the melody so catchy,
It's a grand adventure, a romantic escapade,
The music and scene, perfectly in sync, so enchanting.

Verse 2:
The upbeat tempo, a sense of anticipation,
The jazzy instrumentation, a touch of sophistication,
It all comes together to create a mood so nostalgic,
A romantic adventure, that takes me back in time.

Chorus:
Laughing and smiling, sharing stories of their travels,
The mood is exciting, the melody so catchy,
It's a grand adventure, a romantic escapade,
The music and scene, perfectly in sync, so enchanting.

Bridge:
As the scene unfolds, I'm transported to a different place,
A world of romance, and a time of grace,
The music and scene, a perfect match,
A timeless romance, that never will go out of fashion.

Chorus:
Laughing and smiling, sharing stories of their travels,
The mood is exciting, the melody so catchy,
It's a grand adventure, a romantic escapade,
The music and scene, perfectly in sync, so enchanting.

Outro:
As the music fades, I'm left with a smile,
A romantic and nostalgic feeling, that I can't deny,
The music and scene, a perfect match,
A timeless romance, that never will go out of fashion.

SALMONN-Stable Diffusion-LLaVA model:

Verse 1:
In the glow of city lights, we find our hearts take flight,
Sitting close at a table for two,
Lost in laughter and the joy that we knew.

Chorus:
Cheers to a night of love and cheer,
In this cozy café, we're casting away our fears.
With every sip of wine, our love entwines,
In this moment, time seems to slow down.

Verse 2:
The clinking of glasses, a melody that's free,
The soft lit ambiance, a warmth that's meant to be,
The flicker of a candle, casting dancing shadows,
In this intimate setting, our love seems to grow.

Chorus:
Cheers to a night of love and cheer,
In this cozy café, we're casting away our fears.
With every sip of wine, our love entwines,
In this moment, time seems to slow down.

Bridge:
In the laughter and the light, we find our dreams take flight,
In the gentle touch of your hand, I feel you near,
With every sip of coffee, we share our stories,
In this snapshot of time, our love is clear.

Chorus:
Cheers to a night of love and cheer,
In this cozy café, we're casting away our fears.
With every sip of wine, our love entwines,
In this moment, time seems to slow down.

Outro:

So here's to us, in this warm embrace,
In this intimate setting, we find our place,
With every sip of our drink,
We're creating a memory that's meant to last.

Bibliography

[1] Chuer Chen, Nan Cao, Jiani Hou, Yi Guo, Yulei Zhang and Yang Shi. *MusicJam: Visualizing Music Insights via Generated Narrative Illustrations*, 2023. arXiv:2308.11329, [cs.HC]. Available at: https://arxiv.org/abs/2308.11329.

[2] ByteDance Inc. *SALMONN Model Image*, 2024. Available at: https://github.com/bytedance/SALMONN/blob/main/resource/salmon.png, Accessed: 2024-09-23.

[3] ByteDance Inc. *SALMON Model Structure Image*, 2024. Available at: https://github.com/bytedance/SALMONN/blob/main/resource/structure.png, Accessed: 2024-09-23.

[4] Wikimedia Commons contributors. *File:Claude AI logo.png - Wikimedia Commons*, 2024. Available at: https://commons.wikimedia.org/wiki/File:Claude_AI_logo.png, Accessed: 2024-09-23, Licensed under CC BY-SA 4.0.

[5] Wikipedia contributors. *File:Stable Diffusion architecture.png - Wikipedia*, 2021. Available at: https://en.wikipedia.org/wiki/File:Stable_Diffusion_architecture.png, Accessed: 2024-09-23, Licensed under CC BY-SA 4.0.

[6] Aqeel Anwar. *Difference Between Autoencoder (AE) and Variational Autoencoder (VAE)*. https://towardsdatascience.com/difference-between-autoencoder-ae-and-variational-autoencoder-vae-ed7be1c038f2, 2021. Accessed: 2024-09-27.

[7] OpenAI. *Approach diagram of Whisper model*, 2022. Available at: https://raw.githubusercontent.com/openai/whisper/main/approach.png, Accessed: 2024-09-27.

[8] Yuan Gong. *Audio Spectrogram Transformer (AST) diagram*, 2021. Available at: https://raw.githubusercontent.com/YuanGongND/ast/refs/heads/master/ast.png, Accessed: 2024-09-27.

[9] Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghua Lin, Anton Ragni, Emmanouil Benetos, Norbert Gyenge, Roger Dannenberg, Ruibo Liu, Wenhu Chen, Gus Xia, Yemin Shi, Wenhao Huang, Yike Guo and Jie Fu. *MERT: Acoustic Music Understanding Model with Large-Scale Self-supervised Training*, 2023. arXiv:2306.00107, [cs.SD].

[10] Zihao Deng, Yinghao Ma, Yudong Liu, Rongchen Guo, Ge Zhang, Wenhu Chen, Wenhao Huang and Emmanouil Benetos. *MusiLingo: Bridging Music and Text with Pre-trained Language Models for Music Captioning and Query Response*. *Proceedings of the 2024 Annual Conference of the North American Chapter of the Association for*

Computational Linguistics (NAACL 2024). Association for Computational Linguistics, 2024.

[11] Shansong Liu. *MU-LLaMA Diagram*, 2023. Available at: https://github.com/shansongliu/MU-LLaMA/blob/LLaMA-2/assets/MU-LLaMA.png, Accessed: 2024-09-27.

[12] Shansong Liu. *M^2UGen Diagram*, 2023. Available at: https://github.com/shansongliu/M2UGen/blob/main/assets/M2UGen.png, Accessed: 2024-09-27.

[13] Haotian Liu, Chunyuan Li, Qingyang Wu and Yong Jae Lee. *Visual Instruction Tuning*, 2023. arXiv:2304.08485, [cs.CV]. Available at: https://arxiv.org/abs/2304.08485.

[14] NVIDIA. *Accelerating a Hugging Face Llama 2 and Llama 3 models with Transformer Engine*. https://docs.nvidia.com/deeplearning/transformer-engine/user-guide/examples/te_llama/tutorial_accelerate_hf_llama_with_te.html, 2024. Accessed: 2024-09-27.

[15] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diegode las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix and William El Sayed. *Mistral 7B*, 2023. arXiv:2310.06825, [cs.CL]. Available at: https://arxiv.org/abs/2310.06825.

[16] Sentence Transformers. *Cross-Encoders*. https://www.sbert.net/examples/applications/cross-encoder/README.html, 2020. Accessed: 2024-09-27.

[17] Lianghui Zhu, Xinggang Wang and Xinlong Wang. *JudgeLM: Fine-tuned Large Language Models are Scalable Judges*. 2023.

[18] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu and Enhong Chen. *A Survey on Multimodal Large Language Models*, 2024. arXiv:2306.13549, [cs.CV]. Available at: https://arxiv.org/abs/2306.13549.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin. *Attention Is All You Need*, 2023. arXiv:1706.03762, [cs.CL]. Available at: https://arxiv.org/abs/1706.03762.

[20] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain and Jianfeng Gao. *Large Language Models: A Survey*, 2024. arXiv:2402.06196, [cs.CL]. Available at: https://arxiv.org/abs/2402.06196.

[21] Mayank Mittal and Harkirat Behl. *Variational Autoencoders: A Brief Survey*. https://mayankm96.github.io/assets/documents/projects/cs698-report.pdf, 2018. Accessed: 2024-09-18.

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser and Björn Ommer. *High-Resolution Image Synthesis with Latent Diffusion Models*, 2021. arXiv:2112.10752, [cs.CV].

[23] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey and Ilya Sutskever. *Robust Speech Recognition via Large-Scale Weak Supervision*, 2022. arXiv:2212.04356, [eess.AS]. Available at: https://arxiv.org/abs/2212.04356.

[24] Yuan Gong, Yu An Chung and James Glass. *AST: Audio Spectrogram Transformer*, 2021. arXiv:2104.01778, [cs.SD]. Available at: https://arxiv.org/abs/2104.01778.

[25] Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA and Chao Zhang. *SALMONN: Towards Generic Hearing Abilities for Large Language Models*. *The Twelfth International Conference on Learning Representations*, 2024.

[26] Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun and Ying Shan. *Music Understanding LLaMA: Advancing Text-to-Music Generation with Question Answering and Captioning*. *arXiv preprint arXiv:2308.11276*, 2023.

[27] Atin Sakkeer Hussain, Shansong Liu, Chenshuo Sun and Ying Shan. *M^2UGen: Multi-modal Music Understanding and Generation with the Power of Large Language Models*. *arXiv preprint arXiv:2311.11255*, 2023.

[28] Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain and Aman Chadha. *Exploring the Frontier of Vision-Language Models: A Survey of Current Methodologies and Future Directions*, 2024. arXiv:2404.07214, [cs.CV]. Available at: https://arxiv.org/abs/2404.07214.

[29] Haotian Liu, Chunyuan Li, Yuheng Li and Yong Jae Lee. *Improved Baselines with Visual Instruction Tuning*, 2023.

[30] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei and Ilya Sutskever. *Language Models are Unsupervised Multitask Learners*. https://hayate-lab.com/wp-content/uploads/2023/05/61b1321d512410607235e9a7457a715c.pdf, 2019. Accessed: 2024-09-18.

[31] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave and Guillaume Lample. *LLaMA: Open and Efficient Foundation Language Models*, 2023. arXiv:2302.13971, [cs.CL]. Available at: https://arxiv.org/abs/2302.13971.

[32] Wei Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica and Eric P. Xing. *Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality*, 2023. Available at: https://lmsys.org/blog/2023-03-30-vicuna/.

[33] Wing Lian, Bleys Goodson, Guan Wang, Eugene Pentland, Austin Cook, Chanvichet Vong and "Teknium". *MistralOrca: Mistral-7B Model Instruct-tuned on Filtered OpenOrcaV1 GPT-4 Dataset*. https://huggingface.co/Open-Orca/Mistral-7B-OpenOrca, 2023. Accessed: 2024-09-18.

[34] Anthropic. *The Claude 3 Model Family: Opus, Sonnet, Haiku*. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf, 2024. Accessed: 2024-09-18.

[35] Carlos Gómez-Rodríguez and Paul Williams. *A Confederacy of Models: a Comprehensive Evaluation of LLMs on Creative Writing*, 2023. arXiv:2310.08433, [cs.CL]. Available at: https://arxiv.org/abs/2310.08433.

[36] Anthropic. *Claude 3.5 Sonnet*. https://www.anthropic.com/news/claude-3-5-sonnet, 2024. Accessed: 2024-09-18.

[37] Kishore Papineni, Salim Roukos, Todd Ward and Wei Jing Zhu. *Bleu: a method for automatic evaluation of machine translation*. *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[38] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao and Bill Dolan. *A Diversity-Promoting Objective Function for Neural Conversation Models*, 2016. arXiv:1510.03055, [cs.CL]. Available at: https://arxiv.org/abs/1510.03055.

[39] Zhen Xu, Bingquan Liu, Baoxun Wang, Chengjie Sun, Xiaolong Wang, Zhuoran Wang and Chao Qi. *Neural Response Generation via GAN with an Approximate Embedding Layer*. *Conference on Empirical Methods in Natural Language Processing*, 2017.

[40] Wei Zhao, Michael Strube and Steffen Eger. *DiscoScore: Evaluating Text Generation with BERT and Discourse Coherence*, 2023. arXiv:2201.11176, [cs.CL]. Available at: https://arxiv.org/abs/2201.11176.

[41] Nils Reimers and Iryna Gurevych. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*, 2019. arXiv:1908.10084, [cs.CL]. Available at: https://arxiv.org/abs/1908.10084.

[42] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu and Chenguang Zhu. *G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment*, 2023. arXiv:2303.16634, [cs.CL]. Available at: https://arxiv.org/abs/2303.16634.

[43] Lianmin Zheng, Wei Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez and Ion Stoica. *Judging LLM-as-a-judge with MT-Bench and Chatbot Arena*, 2023. arXiv:2306.05685, [cs.CL].

[44] Sher Badshah and Hassan Sajjad. *Reference-Guided Verdict: LLMs-as-Judges in Automatic Evaluation of Free-Form Text*, 2024. arXiv:2408.09235, [cs.CL]. Available at: https://arxiv.org/abs/2408.09235.

[45] Gabriel Meseguer-Brocal, Alice Cohen-Hadria and Geoffroy Peeters. *DALI: A Large Dataset of Synchronized Audio, Lyrics and notes, Automatically Created using Teacher-student Machine Learning Paradigm*. 2018.

[46] Romain Hennequin, Anis Khlif, Felix Voituret and Manuel Moussallam. *Spleeter: a fast and efficient music source separation tool with pre-trained models*. *Journal of Open Source Software*, 5(50):2154, 2020. Deezer Research.

[47] *xlm-roberta-base-language-detection*. huggingface.co/papluca/xlm-roberta-base-language-detection, 2022. Accessed: 2024-09-18.

[48] *pyspellchecker*. https://github.com/barrust/pyspellchecker, 2024.

[49] Fangyu Liu, Yunlong Jiao, Jordan Massiah, Emine Yilmaz and Serhii Havrylov. *Trans-Encoder: Unsupervised sentence-pair modelling through self- and mutual-distillations*, 2022. arXiv:2109.13059, [cs.CL]. Available at: https://arxiv.org/abs/2109.13059.

[50] *Cross-Encoder for MS Marco*. https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-12-v2, 2021. Accessed: 2024-09-18.

[51] Mistral AI. *Model Card for Mistral-7B-Instruct-v0.3*, 2024. Available at: https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3, Accessed: 2024-09-22.

[52] Meta AI. *Introducing Llama 3.1: Our most capable models to date*, 2024. Available at: https://ai.meta.com/blog/meta-llama-3-1/, Accessed: 2024-09-22.

[53] Ralph Allan Bradley and Milton E. Terry. *Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons*. *Biometrika*, 39(3/4):324–345, 1952.

[54] Wei Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez and Ion Stoica. *Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference*, 2024. arXiv:2403.04132, [cs.AI]. Available at: https://arxiv.org/abs/2403.04132.

[55] Wikipedia contributors. *Z-test*, 2024. Available at: https://en.wikipedia.org/wiki/Z-test, Accessed: 2024-10-12.

[56] Wikipedia contributors. *P-value*, 2024. Available at: https://en.wikipedia.org/wiki/P-value, Accessed: 2024-10-12.

[57] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman and Paul Lamere. *The Million Song Dataset*. *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

[58] Colin Raffel. *Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching*. PhD Thesis, Columbia University, 2016.

[59] Jean Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei,

Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman and Karen Simonyan. *Flamingo: a Visual Language Model for Few-Shot Learning*, 2022. arXiv:2204.14198, [cs.CV]. Available at: https://arxiv.org/abs/2204.14198.

[60] Thomas Melistas, Theodoros Giannakopoulos and Georgios Paraskevopoulos. *Lyrics and Vocal Melody Generation conditioned on Accompaniment*. *Proceedings of the 2nd Workshop on NLP for Music and Spoken Audio (NLP4MusA)*, Sergio Oramas, Elena Epure, Luis Espinosa-Anke, Rosie Jones, Massimo Quadrana, Mohamed Sordo and Kento Watanabe, editors, pages 11–16, Online, 2021. Association for Computational Linguistics.

[61] Paul Mou. *Listening with LLM*. https://paul.mou.dev/posts/2023-12-31-listening-with-llm/, 2023. Accessed: 2024-09-18.

[62] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser and Björn Ommer. *High-Resolution Image Synthesis With Latent Diffusion Models*. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.