Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Évaluation de la description automatique de scènes audio par la tâche d'Audio Question Answering

Marcel Gibier Raphaël Duroselle Pierre Serrano Olivier Boeffard
Jean-François Bonastre
Inria, Paris, France
prénom.nom@inria.fr

RÉSUMÉ

Nous explorons l'évaluation de la tâche de description automatique de scènes audio à travers une approche indirecte basée sur la réponse aux questions sur des documents audio. En l'absence de métriques d'évaluation robustes et automatiques pour la tâche de description automatique de scènes audio, nous nous appuyons sur le benchmark MMAU, un jeu de questions à choix multiple sur des extraits audio variés. Nous introduisons une architecture en cascade qui dépasse les performances de certains modèles de référence de taille comparable. Toutefois, nos résultats mettent en évidence des limitations du benchmark MMAU, notamment un biais textuel et une capacité limitée à évaluer l'intégration conjointe des informations relatives à la parole et aux événements sonores. Nous suggérons des pistes d'amélioration pour rendre les évaluations futures plus fidèles aux enjeux de la tâche de description automatique de scènes audio.

ABSTRACT

Towards the automatic description of audio scenes using the Audio Question Answering task

We explore the task of automatic audio scene description through a proxy task named Audio Question Answering. In the absence of robust and automatic evaluation metrics for audio scene description, we rely on the MMAU benchmark, a multiple-choice question set on diverse audio clips. We introduce a cascade architecture that outperforms some reference models of comparable size. However, our results highlight limitations of the MMAU benchmark, including textual bias and a limited capacity to evaluate the joint integration of speech and sound event information. We propose some possible directions for improvement to better address the challenges of automatic audio scene description in future evaluations.

MOTS-CLÉS : Grands modèles de langage, Modèles multimodaux, Évaluation des modèles, Audio Question Answering .

KEYWORDS: Large Language Models, Multimodal models, Model evaluation, Audio Question Answering.

1 Introduction

Une scène audio désigne la représentation sonore d'un environnement réel, telle qu'elle est perçue ou enregistrée par un ou plusieurs microphones. Ainsi l'objectif de la tâche de description automatique de scènes audio est de générer, à partir d'un signal audio, une représentation sémantique structurée de

Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM), pages 164–177.

l'environnement acoustique, reflétant sa dynamique globale et sa cohérence, plutôt qu'une simple énumération d'événements isolés comme la parole, la musique ou les bruits.

Le niveau le plus élémentaire de cette tâche consiste à générer une transcription enrichie à partir de l'audio (Gravier *et al.*, 2004). Cette transcription va au-delà de la reconnaissance de la suite de mots et intègre plusieurs dimensions complémentaires : indications temporelles, identification des locuteurs, détection des événements sonores, ainsi qu'extraction d'entités nommées (personnes, lieux, organisations, etc.). Chacune de ces sous-tâches peut être évaluée à l'aide de métriques automatiques : le Word Error Rate pour la reconnaissance automatique de la parole, le Slot Error Rate pour l'extraction d'entités (Makhoul *et al.*, 2007), et le F-score pour la détection d'événements ou l'attribution correcte des segments de parole aux locuteurs, en tenant compte d'une tolérance temporelle (Mesaros *et al.*, 2016). La description automatique de scènes audio s'inscrit dans un continuum allant de la transcription enrichie au résumé synthétique, ce dernier ne retenant que les éléments jugés saillants ou pertinents selon un objectif ou un profil utilisateur. Plus on s'oriente vers le résumé, plus le processus de sélection devient subjectif, car il implique des choix explicites sur les informations à conserver ou à omettre. Pour cette raison l'évaluation d'un résumé audio demeure complexe, d'autant plus qu'elle s'appuie majoritairement sur des métriques issues de la traduction automatique, peu adaptées à cette tâche (Elliott & Keller, 2014; Hodosh *et al.*, 2013; Peng *et al.*, 2017).

N'ayant pas identifié de méthodes d'évaluation satisfaisantes pour la tâche de description automatique de scènes audio, nous proposons d'adopter une tâche proxy plus facilement mesurable : la réponse aux questions sur des documents audio (ou "AQA" pour Audio Question Answering) (Lipping *et al.*, 2022). En concevant un questionnaire à choix multiple couvrant un large éventail de tâches audio, il devient alors possible d'évaluer les performances d'un modèle à partir de son taux de bonnes réponses.

Toutefois, l'utilisation de questions en langage écrit rend préférable le recours à un module de traitement automatique du langage naturel. C'est pourquoi nous avons retenu, pour cette étude, des modèles multimodaux combinant texte et audio, couramment désignés sous le terme de modèles de langage basés sur l'audio (ou ALM pour Audio Language Models) (Su *et al.*, 2025).

Notre étude s'articulera en deux sections. Nous commencerons par un état de l'art sur les ALM, avant de présenter les principaux benchmarks d'AQA utilisés pour évaluer ces modèles. Nous introduirons ensuite un modèle simple conçu pour la tâche de description automatique de scènes audio, dont nous analyserons les performances sur un benchmark d'AQA. Enfin, nous discuterons des limites potentielles de ce benchmark, afin d'ouvrir la voie à des perspectives d'amélioration pour l'évaluation de la tâche de description automatique de scènes audio.

2 État de l'art

Dans ce qui suit, nous proposons une revue ciblée des travaux relatifs en AQA, en nous intéressant à la fois aux modèles existants développés pour cette tâche, ainsi qu'aux méthodes d'évaluation employées pour mesurer leur performance.

2.1 Modèles

Nous distinguons ici deux grandes catégories de modèles : les modèles en cascade et les modèles *end-to-end*. Les premiers fonctionnent selon une architecture où chaque composant traite une sous-tâche spécifique tandis que les seconds adoptent une approche dans laquelle toutes les étapes du traitement sont apprises conjointement.

2.1.1 Utilisation d'un pipeline en cascade

Un pipeline en cascade désigne une architecture dans laquelle plusieurs étapes de traitement sont enchaînées de manière séquentielle. Dans notre cas, cela correspond à une succession de modèles experts, chacun spécialisé dans une sous-tâche, collaborant pour résoudre une tâche principale. Ce type d'architecture, bien que potentiellement performant, présente un inconvénient notable : l'ajout de nouveaux modules experts peut entraîner une augmentation significative du nombre de paramètres actifs, alourdissant le système et favorisant l'accumulation d'erreurs issues des différents modules.

Pour limiter cette explosion de complexité, une stratégie consiste à entraîner un module de routage capable de sélectionner dynamiquement les experts pertinents (Naveen *et al.*, 2024). Cette sélection peut se faire sur la base de la question posée, ou à partir des caractéristiques de l'entrée audio. Le routeur peut ainsi activer des modules tels que la reconnaissance automatique de la parole, l'identification du locuteur, la description musicale, etc. Les sorties produites par ces experts spécialisés forment alors un ensemble riche d'informations, qui peuvent être exploitées pour répondre à la question. Une approche consiste à fournir ces sorties à un LLM chargé de formuler la réponse.

Selon la nature des contenus audio impliqués, le recours à un seul modèle expert (en plus du LLM) peut s'avérer suffisant. Par exemple, un modèle de détection d'événements sonores peut suffire pour répondre aux questions portant sur des enregistrements dépourvus de parole et de musique (Bai *et al.*, 2024). De même, un modèle de sous titrage audio (ou captioning audio), générant une brève description résumant le contenu sonore, peut suffire si ce résumé contient assez d'informations pertinentes pour permettre de répondre à la question posée (Kuan & yi Lee, 2024).

2.1.2 Approche *end-to-end*

Un modèle de bout en bout (ou *end-to-end*) est un système dont les prédictions finales sont directement produites à partir des données d'entrée, sans qu'aucune représentation intermédiaire interprétable par un humain ne soit explicitement accessible ou exploitable. Cette définition s'oppose ainsi à celle d'un modèle en cascade, bien que ces deux approches puissent être toutes deux qualifiées de modèles multimodaux dès lors qu'elles intègrent et traitent plusieurs types de données.

Nous nous intéressons ici à une catégorie spécifique de modèles *end-to-end* multimodaux, appelés Audio Language Models (ALM), qui constituent une autre famille de modèles employés pour la tâche d'AQA. Un modèle multimodal est un système capable de traiter et d'intégrer plusieurs types de données (comme le texte, l'audio, ou les images) afin de produire une sortie cohérente tenant compte de toutes les modalités en jeu. Les ALM, en particulier, sont conçus pour faciliter l'analyse et la compréhension des données audio via une interface linguistique.

Un de ces premiers modèles est le modèle CLAP (Elizalde *et al.*, 2022). Il s'agit d'un modèle contrastif inspiré de CLIP (Radford *et al.*, 2021), conçu pour apprendre des représentations partagées

entre l'audio et le texte. Il associe un encodeur audio et un encodeur de texte, entraînés conjointement pour rapprocher dans un espace latent les paires audio/texte correspondantes, via une perte contrastive. Bien que CLAP obtienne d'excellents résultats sur des tâches standards telles que la classification audio en zero-shot, ses performances demeurent limitées sur des tâches plus complexes, notamment celles impliquant la compréhension de l'ordre temporel ou l'attribution d'attributs aux événements acoustiques. Pour pallier ces limitations, plusieurs améliorations ont récemment été proposées (Ghosh *et al.*, 2024c,a). Ce modèle est principalement employé pour des tâches d'AQA impliquant plusieurs réponses possibles. Chaque proposition est encodée avec l'audio, et celle dont la représentation est la plus similaire à celle de l'audio est sélectionnée comme réponse.

Avec l'essor des LLMs, les architectures à base de décodeurs ont gagné en popularité. Pengi (Deshmukh *et al.*, 2023) s'est distingué comme le premier modèle à atteindre des performances SOTA sur un large éventail de tâches de classification audio. Ce travail a ainsi favorisé l'émergence de nombreux grands modèles audio, parmi lesquels figurent LTU (Gong *et al.*, 2024), LTU-AS (Gong *et al.*, 2023), Qwen-Audio (Chu *et al.*, 2023) et Qwen-Audio 2 (Chu *et al.*, 2024). Ces modèles

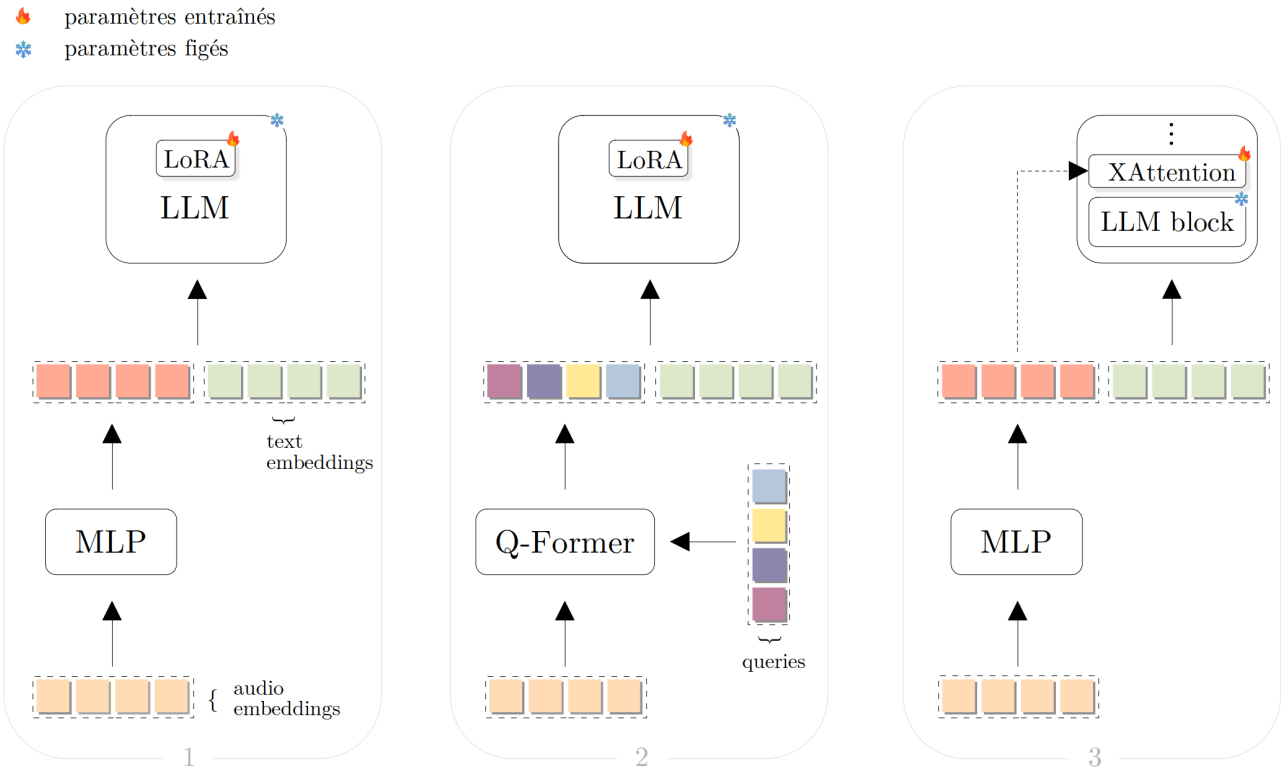

FIGURE 1 – Trois architectures pour intégrer des informations audio dans un modèle de langage : (1) les représentations audio sont simplement passés en tant que préfixe du LLM ; (2) les embeddings sont projetés via un module Q-Former composé de *queries* apprenables ; (3) les représentations audio et textuelles sont fusionnées à l'aide d'un mécanisme de cross-attention. Par ailleurs, le MLP d'entrée peut être étendu et, si nécessaire, intégrer des couches d'attention.

reposent sur une architecture visant à projeter les représentations issues d'un encodeur audio dans un espace de même dimension que celui des embeddings textuels utilisés en entrée du modèle de langage (schéma 1 de la Figure 1). D'autres modèles, tels que SALMONN (Tang *et al.*, 2024) ou GAMA (Ghosh *et al.*, 2024b), adoptent une approche différente en utilisant un Q-Former (Li *et al.*, 2023) chargé de transformer les représentations audio en un ensemble de vecteurs de queries appris (schéma 2). Dans ces approches, le modèle de langage n'est pas directement fine-tuné sur les données

audio. À la place, des *adapters* légers, tels que LoRA (Low-Rank Adaptation) (Hu *et al.*, 2021), sont intégrés afin de conditionner efficacement le modèle à la modalité audio, en facilitant l'interprétation et le traitement des représentations audio injectées en préfixe.

Enfin, des modèles récents tels qu'Audio Flamingo (Kong *et al.*, 2024) et sa dernière version (Ghosh *et al.*, 2025) proposent d'intégrer les informations audio directement au sein des représentations du modèle de langage, en exploitant des mécanismes de cross-attention pour permettre une fusion fine des modalités (Alayrac *et al.*, 2022) (schéma 3).

Cependant, les réponses générées par le modèle de langage ne sont pas systématiquement alignées avec les questions posées. En effet, ces modèles ne sont pas initialement conçus pour la tâche spécifique d'AQA, mais visent plus largement à extraire des informations à partir de requêtes textuelles. Pour cette raison, les réponses peuvent parfois nécessiter un meilleur alignement avec les objectifs de la tâche (Li *et al.*, 2025) ou même un post-traitement comme l'utilisation d'expressions régulières.

Désormais, l'enjeu est de pouvoir évaluer rigoureusement ces modèles multimodaux, afin de mesurer leur capacité réelle à exploiter l'information audio pour répondre à la tâche de description automatique.

2.2 Différents benchmarks pour la tâche d'AQA

Le question answering est une tâche qui consiste à soumettre une question à un modèle afin qu'il génère une réponse pertinente et cohérente. Ce type de tâche peut prendre plusieurs formes : les questions peuvent être ouvertes ou fermées, et les réponses peuvent être proposées sous forme de choix multiples ou non.

Le question answering s'est imposé comme un paradigme central pour l'évaluation des modèles de langage, en raison de sa capacité à couvrir un large éventail de compétences linguistiques et cognitives. En posant aux modèles des questions ciblées, on peut tester de manière fine des aptitudes telles que la compréhension de texte, la synthèse, le raisonnement, ou encore la capacité à mobiliser des connaissances factuelles ou implicites. Cette polyvalence a conduit de nombreux travaux à adopter le QA comme benchmark de référence pour évaluer de manière unifiée les performances sur des tâches aussi diverses que la lecture de documents (Rajpurkar *et al.*, 2016) ou la résolution de problèmes mathématiques simples nécessitant du raisonnement (Cobbe *et al.*, 2021). Enfin, la tâche de question answering a été naturellement élargie pour inclure des questions portant sur des données autres que du texte, telles que des données audio.

Le jeu de données DAQA (Fayek & Johnson, 2019) constitue l'une des premières initiatives à proposer des triplets (audio, question, réponse) afin d'évaluer des compétences en raisonnement temporel. Il se concentre sur des aspects spécifiques tels que la localisation, la comparaison, le comptage, ainsi que les relations temporelles entre les événements audio dans une séquence. Clotho-AQA (Lipping *et al.*, 2022) a ensuite adopté une approche basée sur des questions fermées, se concentrant principalement sur la compréhension générale de l'environnement avec des questions comme *"Y a-t-il des vagues dans l'audio ?"*.

Ce n'est que grâce à l'émergence des ALM, capables de traiter et comprendre l'audio via des interfaces de traitement du langage, que de nouveaux benchmarks pour l'évaluation de la tâche d'AQA ont été introduits (Yang *et al.*, 2024; Wang *et al.*, 2024). Une limitation majeure de ces benchmarks réside dans leur focalisation sur des tâches élémentaires telles que l'ASR, la détection d'événements sonores ou certaines formes de raisonnement simples, comme le raisonnement temporel ou compositionnel de

base. Par ailleurs, leur méthode d'évaluation repose sur une approche de type *Model-as-Judge* (Zheng *et al.*, 2023), dans laquelle un LLM attribue un score aux réponses générées. Cette méthode introduit un biais potentiel : le LLM peut privilégier des formulations stylistiquement plaisantes ou conformes à des préférences humaines, au détriment de la véracité de la réponse.

Dans la continuité de MMLU (Hendrycks *et al.*, 2021), un nouveau benchmark nommé MMAU (Sakshi *et al.*, 2024) a été proposé pour évaluer des tâches complexes nécessitant à la fois une perception audio fine (sons, musique, parole) et un raisonnement expert couvrant 27 compétences spécialisées, telles que la phonétique ou l'harmonie musicale. Ce benchmark contient 10 000 extraits audio soigneusement choisis, représentant de la parole, des sons environnementaux (simplement sons) et de la musique, ainsi que des questions en langage naturel, toutes annotées manuellement. Chaque question étant accompagnée de quatre options de réponse, l'évaluation d'un modèle sur ce benchmark repose alors sur son taux de bonnes réponses.

3 Expérimentations

Cette section présente la conception d'un modèle multimodal en cascade pour la description automatique de scènes audio. Nous évaluons les performances de notre modèle sur le benchmark MMAU avant d'analyser les limites potentielles de ce benchmark pour l'évaluation de notre tâche initiale.

3.1 Construction d'un modèle multimodal en cascade

Notre hypothèse est qu'un modèle de langage est capable de répondre correctement à une question portant sur un contenu audio, à condition de disposer d'une représentation suffisamment riche et structurée de ce signal. Dans cette perspective, nous cherchons à produire une transcription enrichie de l'audio en combinant plusieurs tâches de traitement automatique : la transcription de la parole, l'identification des locuteurs, et la détection d'événements sonores.

L'objectif est d'extraire, à partir d'un signal audio brut, un ensemble d'informations temporellement localisées et sémantiquement pertinentes : des segments de parole transcrits et attribués à chaque locuteur détecté, ainsi qu'une annotation des événements acoustiques présents dans l'enregistrement.

Pour cela, nous combinons plusieurs modèles spécialisés. Un modèle de détection de la voix (Ravanelli *et al.*, 2021) est utilisé afin d'isoler les segments de parole à partir de l'audio, échantillonné à 16 kHz. Ces segments sont ensuite transformés en représentations acoustiques (MFCCs), qui alimentent le modèle ECAPA-TDNN (Desplanques *et al.*, 2020), dédié à l'identification du locuteur. Ce dernier génère des embeddings vocaux, regroupés ensuite en clusters correspondant aux différents locuteurs via une classification hiérarchique ascendante. Les segments annotés sont finalement transcrits à l'aide du décodeur du modèle Whisper (Radford *et al.*, 2022), permettant d'obtenir une transcription horodatée et associée à chaque locuteur détecté. En parallèle, nous exploitons le modèle BEATS (Chen *et al.*, 2023) pour détecter les événements sonores présents dans l'ensemble de l'audio. Ce modèle auto-supervisé, basé sur une architecture ViT (Dosovitskiy *et al.*, 2021), est fine-tuné pour la classification d'événements à partir du corpus AudioSet (Gemmeke *et al.*, 2017).

Finalement la transcription enrichie est organisée dans prompt structuré, utilisé pour interroger un modèle de langage. Nous utilisons dans ce cadre le modèle Qwen-2.5 (Qwen *et al.*, 2025), en raison

de ses bonnes performances sur le benchmark MMLU, et de la disponibilité open source de ses variantes 7B et 32B.

3.2 Résultats sur la tâche d'AQA

Comme mentionné précédemment, nous proposons d'évaluer notre modèle de description automatique de scènes audio à travers la tâche d'AQA. Pour cette évaluation, nous avons sélectionné le benchmark MMAU. Une version *test-mini*, comprenant 1 000 échantillons et leurs réponses correctes, a été mise à disposition, respectant les statistiques du benchmark. Nos évaluations ont été réalisées sur cette version *test-mini*. Voici un exemple de question sur la parole :

Question (parole) : *Identify the role of the first and the second speaker in the conversation.*
A. Parent and child
B. Teacher and student
C. Doctor and patient
D. Coach and athlete

Modèles	**Taille**	**Sons**	**Musique**	**Parole**	**Moyenne**
Qwen2-Audio-Instruct	8.4B	54.95	50.98	42.04	49.20
Audio Flamingo 2	3B	65.10	**72.90**	**68.10**	**68.7**
Ours (Qwen2.5-7b)	8.6B	62.89	54.80	61.29	59.66
Ours (Qwen2.5-32b)	33.6B	**69.37** +6.5%	64.37	61.56	65.10

TABLE 1 – Taux de bonnes réponses (%) sur *test-mini* par catégorie de questions (sons, musique et paroles) pour différents modèles

Nous avons comparé nos résultats (Table 1) à ceux des deux meilleurs modèles de l'état de l'art sur ce benchmark. Notre modèle surpasse Qwen-2-Audio-Instruct à taille équivalente, ce qui peut s'expliquer par la composition déséquilibrée de son jeu de données de pré-entraînement, contenant plus de 70% de données issues de la parole contre seulement environ 2% de sons non verbaux. Ce déséquilibre peut limiter la capacité du modèle à généraliser sur des tâches auditives non linguistiques. En revanche, AudioFlamingo 2 obtient de meilleures performances malgré une taille de modèle significativement inférieure. Cela peut s'expliquer par le fait que c'est un des seuls ALM à être supervisé sur la tâche d'AQA. Par ailleurs, dans notre modèle, l'utilisation d'une tête de classification supervisée dans BEATs pourrait restreindre la richesse des représentations acoustiques extraites. En effet, la seule information binaire sur la présence d'événements sonores ne permet pas de capturer la totalité des attributs acoustiques tels que la hauteur ou l'intensité, et la réponse à des questions dans le benchmark portant sur ces attributs repose alors uniquement sur la connaissance a priori que le modèle de langage possède sur ces sons.

Cependant, l'augmentation de la taille du modèle de langage permet à notre système de surpasser nettement AudioFlamingo 2 sur les questions relatives aux sons, et ce, sans l'ajout explicite d'informations acoustiques. Cela suggère que l'augmentation de la capacité de raisonnement linguistique du modèle permet d'être plus performant sur le benchmark n'ayant aucune information supplémentaire sur les données audio.

3.3 Discussion sur les limites du benchmark MMAU

Pour comprendre pourquoi l'augmentation de la capacité de raisonnement linguistique du modèle suffisait à obtenir un meilleur score sur MMAU, nous avons décidé de l'examiner en détail. Nous avons observé certaines questions dont la réponse pouvait être déduite uniquement à partir de l'énoncé textuel, sans nécessiter le recours à l'information audio, comme en témoigne l'exemple suivant :

> **Question** (sons) : *What is the sound in the audio that is typically produced by small, flying insects that feed on blood, often causing itchy bites ?*
> **A. Mosquito buzzing**
> B. Birds chirping
> C. Wind blowing
> D. Rain Falling

Cette observation s'explique par le mode de construction de MMAU, qui repose sur l'exploitation d'un modèle de langage pour générer de nouvelles réponses candidates à partir de la question et de la réponse correcte. Cela permet de comprendre l'amélioration significative des performances observée avec l'augmentation de la taille du modèle de langage.

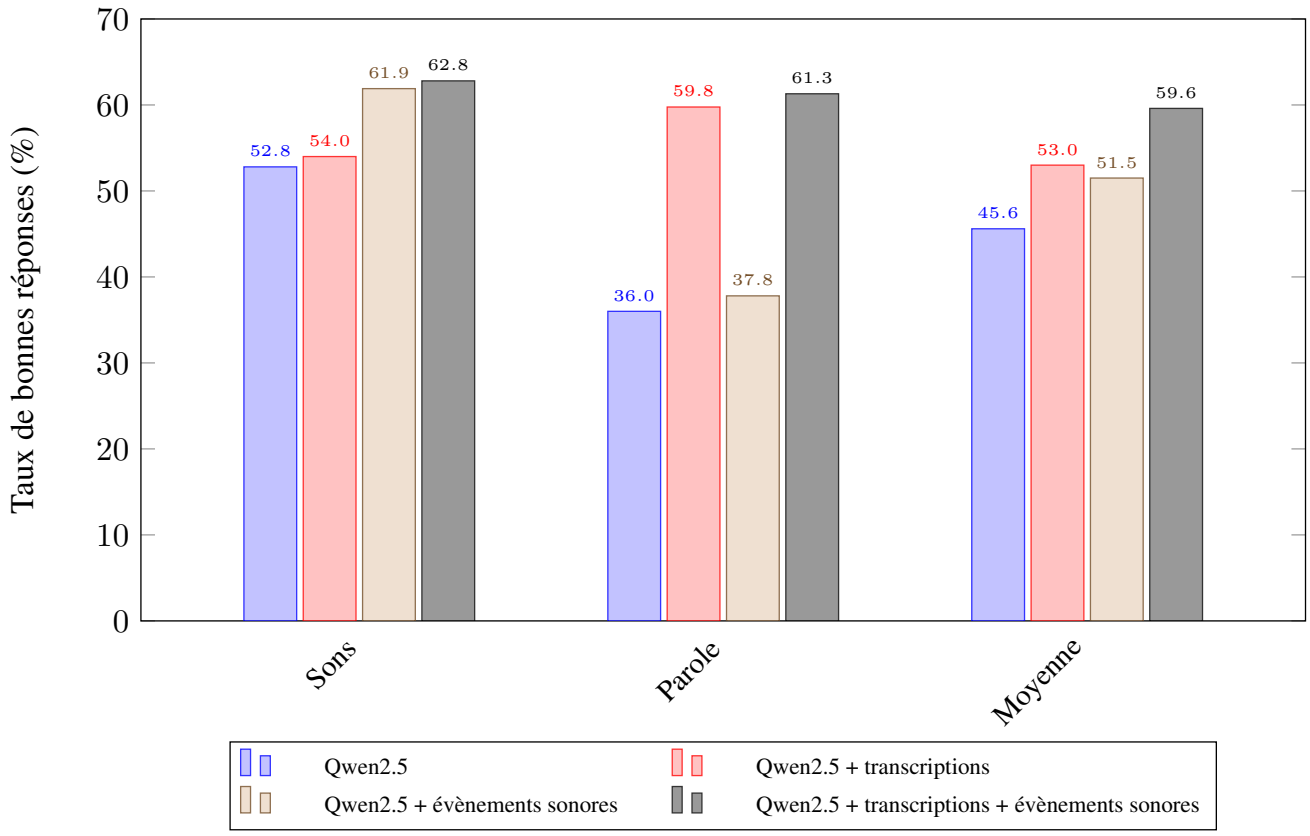

FIGURE 2 – Taux de bonnes réponses (%) sur *test-mini* par catégorie de questions (sons et paroles) du modèle Qwen2.5-7b selon la disponibilité des évènements sonores de BEATS et des transcriptions de Whisper.

Afin de nous assurer que notre évaluation n'a pas été biaisée par une sélection particulière de questions mal construites, nous avons mesuré les performances de notre modèle sans accès aux informations

audio (Figure 2 en bleu). Contrairement aux travaux des auteurs de MMAU, qui visent à montrer que remplacer l'audio par du bruit n'affecte pas les performances de certains modèles, notre objectif ici est de valider que le modèle de langage seul est performant sans accès à l'audio .

Nous remarquons que pour les questions portant sur la parole, Qwen-2.5 avec les transcription dépasse largement la version sans (59.8 % contre 36.0 %), ce qui suggère que l'audio contient une information pertinente. En revanche, pour les questions liées aux sons, l'écart de performance entre Qwen-2.5 avec ou sans accès aux événements sonores est nettement moins marqué (61.9 % contre 52.8 %). Ce résultat indique la présence d'un biais textuel significatif dans la formulation des questions ou des réponses du benchmark. À titre comparatif, ce taux de bonnes réponses dépasse souvent celui obtenu par les modèles multimodaux audio de taille équivalente, qui se concentrent pourtant spécifiquement sur l'audio.

Sur cette même figure, nous avons également reporté le taux de bonnes réponses du modèle pour les questions relatives aux sons et à la parole, selon qu'il dispose des informations issues de la transcription de la parole, des évènements sonores, ou des deux. Cette analyse révèle que :

— Pour les questions de type *sons*, les évènements sonores seuls sont généralement suffisants, et l'ajout des transcriptions n'apporte qu'une amélioration marginale des résultats.
— De même, pour les questions de type *parole*, les informations sur les événements ne semblent pas véritablement nécessaires pour obtenir la réponse correcte.

Ces observations suggèrent que le benchmark actuel ne couvre pas suffisamment de cas où l'intégration conjointe des informations relatives à la parole et aux événements sonores est essentielle pour répondre correctement aux questions. Bien que cela ne constitue pas un objectif du benchmark, cette limitation reste problématique pour la tâche de description automatique de scènes audio. En effet, la tâche de description automatique de scènes audio requiert parfois un raisonnement fondé sur l'interdépendance de ces deux types d'informations. C'est notamment le cas lorsque, par exemple, une personne adapte son discours en réaction immédiate à un événement sonore ou élève la voix pour se faire entendre dans un environnement bruyant (effet Lombard (Brumm & Zollinger, 2011)).

4 Conclusion

Dans cet article, nous étudions la pertinence de l'utilisation de la tâche d'Audio Question Answering comme proxy pour l'évaluation de la description automatique de scènes audio. À cette fin, nous proposons un modèle multimodal en cascade, simple et sans phase d'entraînement. L'évaluation de ce modèle est réalisée à l'aide du benchmark MMAU, un questionnaire à choix multiple conçu pour des enregistrements audio comprenant des sons, de la musique et de la parole, et couvrant un large éventail de tâches de compréhension audio.

Cependant, une analyse du benchmark MMAU révèle certaines limites méthodologiques susceptibles de biaiser l'évaluation comparative des approches basées sur les LLM. En effet, il apparait que plusieurs paires question/réponse se révèlent mal construites, notamment en raison du processus semi-automatique de leur génération via un LLM. Une amélioration possible consisterait à enrichir le prompt de génération des distracteurs en y intégrant des informations relatives au contenu audio, contrairement à MMAU qui se limite actuellement à la question et à la réponse. De même, un LLM pourrait être mobilisé pour reformuler les questions de manière à éliminer toute information

permettant d'identifier la bonne réponse de façon triviale.

Enfin, la tâche de description automatique de scènes audio implique un raisonnement fondé sur l'intégration conjointe des informations relatives à la parole et aux événements sonores. Ce type de raisonnement n'est pas encore pris en compte dans l'évaluation proposée par MMAU. Une extension du benchmark pourrait consister à introduire des questions nécessitant l'extraction et l'interprétation combinée de ces deux types d'informations (Wang *et al.*, 2025). Par ailleurs, explorer le rôle potentiel des agents pourrait offrir de nouvelles pistes intéressantes pour améliorer la tâche de description de scènes audio.

En conclusion, la tâche d'Audio Question Answering apparaît pertinente pour évaluer les systèmes de description automatique de scènes audio. Néanmoins, les benchmarks existants restent encore insuffisamment aboutis, tant en termes de rigueur d'annotation que de diversité et de réalisme des situations représentées, pour constituer une évaluation unique et pleinement fiable.

Références

ALAYRAC J.-B., DONAHUE J., LUC P., MIECH A., BARR I., HASSON Y., LENC K., MENSCH A., MILLICAN K., REYNOLDS M., RING R., RUTHERFORD E., CABI S., HAN T., GONG Z., SAMANGOOEI S., MONTEIRO M., MENICK J. L., BORGEAUD S., BROCK A., NEMATZADEH A., SHARIFZADEH S., BIŃKOWSKI M. A., BARREIRA R., VINYALS O., ZISSERMAN A. & SIMONYAN K. (2022). Flamingo : a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*. DOI : 10.48550/arXiv.2204.14198.

BAI J., YIN H., WANG M., SHI D., GAN W.-S., CHEN J. & RAHARDJA S. (2024). Audiolog : LLMs-Powered Long Audio Logging with Hybrid Token-Semantic Contrastive Learning . In *2024 IEEE International Conference on Multimedia and Expo (ICME)*. DOI : 10.1109/ICME57554.2024.10688214.

BRUMM H. & ZOLLINGER S. A. (2011). The evolution of the lombard effect : 100 years of psychoacoustic research. *Behaviour*, **148**, 1173–1198. DOI : 10.2307/41445240.

CHEN S., WU Y., WANG C., LIU S., TOMPKINS D., CHEN Z., CHE W., YU X. & WEI F. (2023). BEATs : Audio pre-training with acoustic tokenizers. In *Proceedings of the 40th International Conference on Machine Learning*. DOI : 10.48550/arXiv.2212.09058.

CHU Y., XU J., YANG Q., WEI H., WEI X., GUO Z., LENG Y., LV Y., HE J., LIN J., ZHOU C. & ZHOU J. (2024). Qwen2-audio technical report. DOI : 10.48550/arXiv.2407.10759.

CHU Y., XU J., ZHOU X., YANG Q., ZHANG S., YAN Z., ZHOU C. & ZHOU J. (2023). Qwen-audio : Advancing universal audio understanding via unified large-scale audio-language models. DOI : 10.48550/arXiv.2311.07919.

COBBE K., KOSARAJU V., BAVARIAN M., CHEN M., JUN H., KAISER L., PLAPPERT M., TWOREK J., HILTON J., NAKANO R., HESSE C. & SCHULMAN J. (2021). Training verifiers to solve math word problems. *CoRR*. DOI : 10.48550/arXiv.2110.14168.

DESHMUKH S., ELIZALDE B., SINGH R. & WANG H. (2023). Pengi : an audio language model for audio tasks. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. DOI : 10.48550/arXiv.2305.11834.

DESPLANQUES B., THIENPONDT J. & DEMUYNCK K. (2020). Ecapa-tdnn : Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In *21st Annual Conference of the International Speech Communication Association, Interspeech 2020*. DOI : 10.21437/interspeech.2020-2650.

DOSOVITSKIY A., BEYER L., KOLESNIKOV A., WEISSENBORN D., ZHAI X., UNTERTHINER T., DEHGHANI M., MINDERER M., HEIGOLD G., GELLY S., USZKOREIT J. & HOULSBY N. (2021). An image is worth 16x16 words : Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021*. DOI : 10.48550/arXiv.2010.11929.

ELIZALDE B., DESHMUKH S., ISMAIL M. A. & WANG H. (2022). Clap : Learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing*. DOI : 10.48550/arXiv.2206.04769.

ELLIOTT D. & KELLER F. (2014). Comparing automatic evaluation measures for image description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 452–457. DOI : 10.3115/v1/P14-2074.

FAYEK H. M. & JOHNSON J. (2019). Temporal reasoning via audio question answering. In *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* DOI : /10.48550/arXiv.1911.09655.

GEMMEKE J. F., ELLIS D. P. W., FREEDMAN D., JANSEN A., LAWRENCE W., MOORE R. C., PLAKAL M. & RITTER M. (2017). Audio set : An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. DOI : 10.1109/ICASSP.2017.7952261.

GHOSH S., KONG Z., KUMAR S., SAKSHI S., KIM J., PING W., VALLE R., MANOCHA D. & CATANZARO B. (2025). Audio flamingo 2 : An audio-language model with long-audio understanding and expert reasoning abilities. DOI : 10.48550/arXiv.2503.03983.

GHOSH S., KUMAR S., EVURU C. K. R., NIETO O., DURAISWAMI R. & MANOCHA D. (2024a). Reclap : Improving zero shot audio classification by describing sounds. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing*. DOI : 10.48550/arXiv.2409.09213.

GHOSH S., KUMAR S., SETH A., EVURU C. K. R., TYAGI U., SAKSHI S., NIETO O., DURAISWAMI R. & MANOCHA D. (2024b). Gama : A large audio-language model with advanced audio understanding and complex reasoning abilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. DOI : 10.48550/arXiv.2406.11768.

GHOSH S., SETH A., KUMAR S., TYAGI U., EVURU C. K., RAMANESWARAN S., SAKSHI S., NIETO O., DURAISWAMI R. & MANOCHA D. (2024c). Compa : Addressing the gap in compositional reasoning in audio-language models. In *12th International Conference on Learning Representations, ICLR 2024*. DOI : 10.48550/arXiv.2310.08753.

GONG Y., LIU A. H., LUO H., KARLINSKY L. & GLASS J. (2023). Joint audio and speech understanding. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, p. 1–8. DOI : 10.1109/ASRU57964.2023.10389742.

GONG Y., LUO H., LIU A. H., KARLINSKY L. & GLASS J. (2024). Listen, think, and understand. In *The 12th International Conference on Learning Representations, ICLR 2024*. DOI : 10.48550/arXiv.2305.10790.

GRAVIER G., BONASTRE J.-F., GEOFFROIS E., GALLIANO S., MCTAIT K. & CHOUKRI K. (2004). The ESTER evaluation campaign for the rich transcription of French broadcast news. In M. T. LINO, M. F. XAVIER, F. FERREIRA, R. COSTA & R. SILVA, Éds., *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*.

HENDRYCKS D., BURNS C., BASART S., ZOU A., MAZEIKA M., SONG D. & STEINHARDT J. (2021). Measuring massive multitask language understanding. In *The 9th International Conference on Learning Representations, ICLR 2021*. DOI : 10.48550/arXiv.2009.03300.

HODOSH M., YOUNG P. & HOCKENMAIER J. (2013). Framing image description as a ranking task : Data, models and evaluation metrics. *J. Artif. Intell. Res.*, **47**, 853–899. DOI : 10.1613/jair.3994.

HU E. J., SHEN Y., WALLIS P., ALLEN-ZHU Z., LI Y., WANG S., WANG L. & CHEN W. (2021). Lora : Low-rank adaptation of large language models. In *The 10th International Conference on Learning Representations, ICLR 2022*. DOI : 10.48550/arXiv.2106.09685.

KONG Z., GOEL A., BADLANI R., PING W., VALLE R. & CATANZARO B. (2024). Audio flamingo : A novel audio language model with few-shot learning and dialogue abilities. In *Proceedings of the 41st International Conference on Machine Learning*. DOI : 10.48550/arXiv.2402.01831.

KUAN C.-Y. & YI LEE H. (2024). Can large audio-language models truly hear ? tackling hallucinations with multi-task assessment and stepwise audio reasoning. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing*. DOI : 10.48550/arXiv.2410.16130.

LI G., LIU J., DINKEL H., NIU Y., ZHANG J. & LUAN J. (2025). Reinforcement learning outperforms supervised fine-tuning : A case study on audio question answering. DOI : 10.48550/arXiv.2503.11197.

LI J., LI D., SAVARESE S. & HOI S. (2023). Blip-2 : Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*. DOI : /10.48550/arXiv.2301.12597.

LIPPING S., SUDARSANAM P., DROSSOS K. & VIRTANEN T. (2022). Clotho-aqa : A crowdsourced dataset for audio question answering. In *2022 30th European Signal Processing Conference (EUSIPCO)*. DOI : /10.48550/arXiv.2204.09634.

MAKHOUL J., KUBALA F., SCHWARTZ R. E. & WEISCHEDEL R. M. (2007). Performance measures for information extraction. In *Proceedings of DARPA Broadcast News Workshop*.

MESAROS A., HEITTOLA T. & VIRTANEN T. (2016). Metrics for polyphonic sound event detection.

Applied Sciences, **6**(6). DOI : 10.3390/app6060162.

NAVEEN V., SRIDHAR A. K., GUO Y. & VISSER E. (2024). Comprehensive audio query handling system with integrated expert models and contextual understanding. DOI : 10.48550/arXiv.2412.03980.

PENG B., LI X., LI L., GAO J., CELIKYILMAZ A., LEE S. & WONG K.-F. (2017). Composite task-completion dialogue policy learning via hierarchical deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* : Association for Computational Linguistics. DOI : 10.18653/v1/d17-1237.

QWEN, :, YANG A., YANG B., ZHANG B., HUI B., ZHENG B., YU B., LI C., LIU D., HUANG F., WEI H., LIN H., YANG J., TU J., ZHANG J., YANG J., YANG J., ZHOU J., LIN J., DANG K., LU K., BAO K., YANG K., YU L., LI M., XUE M., ZHANG P., ZHU Q., MEN R., LIN R., LI T., TANG T., XIA T., REN X., REN X., FAN Y., SU Y., ZHANG Y., WAN Y., LIU Y., CUI Z., ZHANG Z. & QIU Z. (2025). Qwen2.5 technical report. DOI : 10.48550/arXiv.2407.10759.

RADFORD A., KIM J. W., HALLACY C., RAMESH A., GOH G., AGARWAL S., SASTRY G., ASKELL A., MISHKIN P., CLARK J., KRUEGER G. & SUTSKEVER I. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. DOI : 10.48550/arXiv.2103.00020.

RADFORD A., KIM J. W., XU T., BROCKMAN G., MCLEAVEY C. & SUTSKEVER I. (2022). Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*. DOI : 10.48550/arXiv.2212.04356.

RAJPURKAR P., ZHANG J., LOPYREV K. & LIANG P. (2016). Squad : 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. DOI : 10.18653/v1/D16-1264.

RAVANELLI M., PARCOLLET T., PLANTINGA P., ROUHE A., CORNELL S., LUGOSCH L., SUBAKAN C., DAWALATABAD N., HEBA A., ZHONG J., CHOU J.-C., YEH S.-L., FU S.-W., LIAO C.-F., RASTORGUEVA E., GRONDIN F., ARIS W., NA H., GAO Y., MORI R. D. & BENGIO Y. (2021). Speechbrain : A general-purpose speech toolkit. DOI : 10.48550/arXiv.2106.04624.

SAKSHI S., TYAGI U., KUMAR S., SETH A., SELVAKUMAR R., NIETO O., DURAISWAMI R., GHOSH S. & MANOCHA D. (2024). Mmau : A massive multi-task audio understanding and reasoning benchmark. In *The 13th International Conference on Learning Representations, ICLR 2025*. DOI : 10.48550/arXiv.2410.19168.

SU Y., BAI J., XU Q., XU K. & DOU Y. (2025). Audio-language models for audio-centric tasks : A survey. DOI : 10.48550/arXiv.2501.15177.

TANG C., YU W., SUN G., CHEN X., TAN T., LI W., LU L., MA Z. & ZHANG C. (2024). Salmonn : Towards generic hearing abilities for large language models. In *The 12th International Conference on Learning Representations, ICLR 2024*. DOI : 10.48550/arXiv.2310.13289.

WANG B., ZOU X., LIN G., SUN S., LIU Z., ZHANG W., LIU Z., AW A. & CHEN N. F. (2024).

Audiobench : A universal benchmark for audio large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics : Human Language Technologies (Volume 1 : Long Papers)*. DOI : 10.48550/arXiv.2406.16020.

WANG Y., MOUSAVI P., PLOUJNIKOV A. & RAVANELLI M. (2025). What are they doing ? joint audio-speech co-reasoning. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing*. DOI : 10.1109/ICASSP49660.2025.10889092.

YANG Q., XU J., LIU W., CHU Y., JIANG Z., ZHOU X., LENG Y., LV Y., ZHAO Z., ZHOU C. & ZHOU J. (2024). AIR-bench : Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*. DOI : 10.18653/v1/2024.acl-long.109.

ZHENG L., CHIANG W.-L., SHENG Y., ZHUANG S., WU Z., ZHUANG Y., LIN Z., LI Z., LI D., XING E. P., ZHANG H., GONZALEZ J. E. & STOICA I. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*. DOI : 10.48550/arXiv.2306.05685.