Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Vers un benchmark pour une évaluation robuste de la catégorisation de contenus audiovisuels transcrits

Abdelkrim Beloued
Institut national de l'audiovisuel, 4 avenue de l'Europe, 94360 Bry-sur-Marne, France
abeloued@ina.fr

RÉSUMÉ

La catégorisation thématique des contenus audiovisuels constitue un enjeu central pour l'analyse des tendances éditoriales et l'exploration de grands corpus d'archives. Contrairement au texte écrit, pour lequel de nombreux frameworks et benchmarks existent, les contenus audiovisuels, et en particulier leurs transcriptions, disposent de peu de ressources dédiées, notamment en français. Dans cet article, nous proposons un benchmark en français adapté à cette tâche. Il repose sur un jeu de données issu de contenus audiovisuels transcrits et vise à évaluer les performances de modèles de langue, qu'ils soient génératifs ou discriminatifs. Nous décrivons une méthodologie de curation permettant de produire plusieurs variantes du jeu de données ainsi que différents niveaux de qualité, afin d'analyser le comportement des modèles face à des données de fiabilité variable. Les expérimentations menées évaluent à la fois la qualité du jeu de données, les méthodes de construction utilisées et les performances des modèles avant et après affinage. Ce travail contribue ainsi à combler le manque de ressources d'évaluation pour la catégorisation de contenus audiovisuels en français.

ABSTRACT

Towards a benchmark for robust evaluation of transcribed audiovisual content categorization

The thematic categorization of audiovisual content plays a crucial role in analyzing thematic patterns and trends within large archival corpora. Unlike written text, for which many frameworks and benchmarks are available, audiovisual content, and particularly its transcriptions, benefits from very limited dedicated resources, especially in French. In this article, we introduce a French benchmark tailored to this task. It is based on a dataset derived from transcribed audiovisual content and aims to assess the performance of language models, whether generative or discriminative. We describe a curation methodology that produces several variants of the dataset as well as different quality levels, enabling the analysis of model behavior under varying levels of data reliability. The experiments evaluate both the quality of the dataset, the construction methods employed, and model performance before and after fine-tuning. This work contributes to addressing the lack of evaluation resources for the categorization of audiovisual content in French.

MOTS-CLÉS : Classification thématique, Catégorisation, Contenu audiovisuel, Transcription, Jeu de données, Benchmark, Évaluation, Modèle de langue, LLM.

KEYWORDS: Thematic classification, Categorization, Audiovisual content, Transcription, Dataset, Benchmark, Evaluation, Language model, LLM.

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 1 : articles scientifiques originaux, pages 669–687.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

1 Introduction

La classification thématique constitue l'une des tâches fondatrices du traitement automatique du langage naturel. Largement explorée depuis l'essor des réseaux de neurones, la tâche a vu ses performances nettement progresser avec l'émergence des *Transformers* et des grands modèles de langue. L'évaluation de ces modèles constitue une étape déterminante pour garantir leur utilisation fiable dans un domaine d'application donné. Un modèle ne peut être considéré comme performant que si son évaluation est réalisée sur des données représentatives du contexte réel d'utilisation.

Au-delà des données elles-mêmes, la mise en place d'une méthodologie d'évaluation rigoureuse est indispensable. Celle-ci doit permettre d'analyser les performances selon plusieurs axes (précision, robustesse, généralisation), mais également d'évaluer la qualité des jeux de données ainsi que leurs modalités de constitution. Une telle approche contribue à garantir une évaluation véritablement représentative des conditions réelles d'utilisation.

En dépit des progrès significatifs enregistrés en classification thématique, son application à des domaines spécifiques soulève encore des difficultés importantes. C'est notamment le cas de la catégorisation de contenus audiovisuels à partir de leurs transcriptions, pour laquelle les ressources d'évaluation disponibles en français restent limitées. De plus, l'absence de protocoles d'évaluation adaptés complique la comparaison des modèles et des jeux de données. À cela s'ajoute un manque de standards d'annotation partagés. La multiplicité des référentiels réduit les possibilités de mutualisation des jeux de données et des modèles et freine la structuration d'un écosystème commun. Une harmonisation des référentiels pourrait ainsi faciliter le partage des ressources et améliorer la réutilisation des modèles au sein de la communauté scientifique.

Afin de répondre à ces défis, nous proposons plusieurs contributions. Premièrement, nous introduisons un nouveau jeu de données en français, issu des collections d'archives de l'INA (Institut National de l'Audiovisuel), annoté manuellement par des professionnels et consolidé au moyen de plusieurs procédures de curation et de validation. Deuxièmement, nous présentons un protocole expérimental d'évaluation permettant d'exploiter pleinement le potentiel de ce jeu de données. Enfin, nous défendons une démarche de normalisation des référentiels visant à faciliter la mutualisation des ressources (jeux de données, modèles affinés). L'ensemble de ces contributions vise à structurer un cadre cohérent et unifié pour l'affinage et l'évaluation de modèles de langue, qu'ils soient génératifs ou discriminatifs, intégrant différentes architectures de classification thématique.

2 État de l'art

L'évaluation des modèles dans leur contexte réel d'utilisation nécessite des jeux de données représentatifs et de qualité. Notre proposition cible spécifiquement la catégorisation de contenus audiovisuels à partir de transcriptions, alors que la majorité des benchmarks existants se concentrent sur la classification d'articles de presse pris dans leur intégralité, souvent en mono-label. Parmi ces benchmarks, ***AG News***(Zhang *et al.*, 2016) est l'un des benchmarks les plus utilisés pour la classification thématique. Il regroupe des articles d'actualité répartis en 4 catégories (sciences, sports, business, technologies) et sert principalement de référence pour des tâches de classification supervisée de base. ***Reuters-21578***(Lewis, 1987) constitue une collection historique de dépêches économiques et financières largement utilisée en catégorisation de textes. Il a longtemps servi de standard expérimental pour l'évaluation de modèles de classification de documents. Le ***News Category Dataset***(Misra, 2022)

rassemble plus de 210000 articles du site d'actualité *Huffington Post* publiés entre 2012 et 2022 et répartis en 42 catégories. Il constitue un cadre multi-classes réaliste et est utilisé pour des tâches variées telles que la classification thématique, l'analyse de biais médiatiques ou l'étude diachronique des tendances journalistiques. Le benchmark ***20 Newsgroups*** [1] (Mitchell, 1997) comprend environ 19000 messages provenant de forums de discussion répartis en 20 catégories mono-label couvrant des thématiques variées (informatique, politique, religion, sciences, loisirs).

Une caractéristique commune à ces benchmarks est que chacun définit son propre ensemble de catégories, alors même qu'une large partie de ces classes se recoupent d'un référentiel à l'autre. Une harmonisation permettrait ainsi de mutualiser plus efficacement les jeux de données et les modèles. Dans cette perspective, la taxonomie IPTC (*International Press Telecommunications Council*) [2] constitue une référence structurante dans le domaine de la presse et des médias et pourrait servir de socle commun pour la conception de benchmarks de catégorisation. Peu de travaux dans la littérature s'appuient explicitement sur cette taxonomie. Parmi eux, le jeu de données MN-DS (Petukhova & Fachada, 2023) qui est composé d'articles de presse en anglais et permet une classification multi-label selon les deux premiers niveaux de la taxonomie IPTC. Il constitue une ressource de référence pour l'apprentissage et l'évaluation selon cette taxonomie. Nous proposons dans cet article une comparaison directe entre ce corpus et notre jeu de données.

Ces travaux présentent néanmoins plusieurs limites : l'absence de ressources en français, le manque de jeux de données spécifiquement adaptés au domaine audiovisuel et l'inexistence de jeux de données issus de transcriptions audiovisuelles. C'est dans ce contexte que nous proposons un nouveau benchmark en français dédié à l'évaluation de la catégorisation de contenus audiovisuels transcrits. Contrairement aux benchmarks existants, centrés sur des articles écrits, notre approche vise une catégorisation directement à partir des transcriptions, sans recourir à des annotations supplémentaires ni à des étapes intermédiaires de résumé. L'objectif est ainsi d'évaluer la classification à partir d'une seule modalité issue du signal : la transcription, afin de mieux refléter les conditions réelles d'usage.

3 Corpus annoté

Le corpus annoté comprend 202 émissions, représentant environ 81 heures de programmes télévisés et radiophoniques, et couvre la période 1982-2025. Ces émissions proviennent de chaînes françaises parmi les plus suivies. Sa constitution repose sur deux critères principaux : la représentativité des formats présents dans les fonds d'archives de l'INA et la diversité des thématiques abordées. Le corpus inclut ainsi une large variété de formats, allant des journaux télévisés et radiophoniques aux interviews politiques, débats, émissions de divertissement et talk-show. Ces programmes offrent plusieurs niveaux de complexité et permettent ainsi d'évaluer les modèles de langage selon divers degrés de difficulté. L'objectif est de tester les modèles à la fois sur des programmes structurés (comme les journaux télévisés qui sont une succession de plateau/reportage) et sur des programmes non structurés (par exemple les talk-shows où les interruptions et changements de sujet et de thématique sont fréquents).

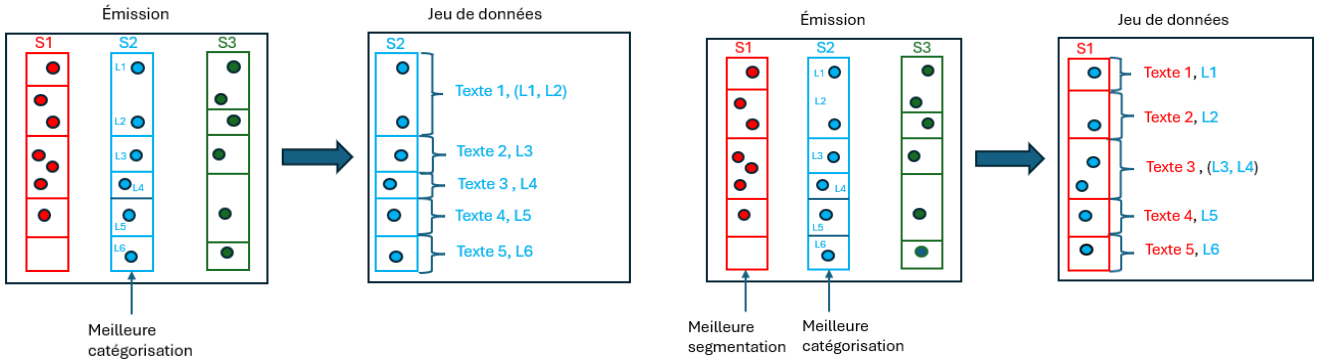
Ce corpus a été confié à des annotateurs professionnels chargés de réaliser les tâches de segmentation et de catégorisation, principalement à partir des transcriptions des programmes. Au total, 28 annotateurs experts ont participé à la campagne d'annotation. Chaque émission est annotée par plusieurs annotateurs. Le travail consiste d'abord à segmenter chaque émission en plusieurs sujets, puis à

1. https://archive.ics.uci.edu/dataset/113/twenty+newsgroups
2. https://www.iptc.org/std/NewsCodes/treeview/mediatopic/mediatopic-en-GB.html

attribuer à chacun d'eux une ou plusieurs catégories issues d'une taxonomie hiérarchique dérivée de l'IPTC, comprenant 115 catégories réparties sur trois niveaux. Ce processus d'annotation aboutit à un accord inter-annotateur de 0,6, mesuré à l'aide du coefficient de *Krippendorff* (Krippendorff, 1980).

4 Curation et constitution du jeu de données d'évaluation

Le jeu de données final, destiné à une tâche de classification thématique, doit donc être consolidé à partir de ces annotations multiples. Il s'agit d'identifier, pour chaque émission, les segments les plus pertinents ainsi que les catégories les plus fiables afin de définir respectivement les textes et leurs labels (cf. figure 1). Les stratégies de consolidation mises en œuvre sont détaillées dans les sections 4.1 et 4.2.



(a) Jeu de données basé sur la meilleure catégorisation (b) Jeu de données basé sur la meilleure segmentation

FIGURE 1 – Structure et variantes du jeu de données

S_1, S_2 et S_3 correspondent aux segmentations produites par plusieurs annotateurs pour une même émission.

4.1 Approches d'agrégation basées sur l'accord inter-annotateur

Meilleure catégorisation Afin d'identifier la meilleure catégorisation, cette approche ignore les frontières de segments et opère exclusivement au niveau des phrases de la transcription. Les catégories attribuées à chaque segment sont ainsi propagées aux phrases correspondantes. L'annotation retenue est celle qui maximise la moyenne du coefficient de *Krippendorff* (Krippendorff, 1980), calculé à partir des accords pair-à-pair entre les annotations d'une même émission. La segmentation associée à cette annotation est ensuite utilisée pour générer les textes du jeu de données (cf. figure 1a).

Cette approche se distingue par sa simplicité de mise en œuvre et par l'utilisation des labels issus de la catégorisation la plus consensuelle. Elle intègre indirectement une prise en compte partielle de la segmentation. En revanche, les textes retenus ne correspondent pas nécessairement aux segments les plus pertinents, car l'optimisation porte prioritairement sur la catégorisation et non sur la segmentation.

Meilleure segmentation Cette approche privilégie l'optimisation de la segmentation, supposée fournir la structuration la plus pertinente des textes constituant le jeu de données. La métrique *WindowDiff* (Pevzner & Hearst, 2002) est utilisée pour mesurer la distance entre les frontières de segmentation. La segmentation retenue pour une émission est celle qui présente la plus faible moyenne de *WindowDiff*, calculée à partir des comparaisons pair-à-pair avec les autres segmentations de la même émission. Les segments issus de cette segmentation sont alors sélectionnés pour constituer les

textes du jeu de données. Les labels sont ensuite déterminés, pour chaque segment, en appliquant la méthode décrite dans la section précédente, mais restreinte aux phrases appartenant à ce segment (cf. figure (1b)).

Cette approche permet de sélectionner exclusivement des textes issus de la segmentation la plus consensuelle. De plus, les labels sont déterminés au niveau de chaque segment, ce qui permet un choix plus fin que dans l'approche précédente, où la décision était prise à l'échelle de l'émission entière.

4.2 Approche d'agrégation probabilistes

Dawid & Skene (Dawid & Skene, 1979) est une méthode probabiliste visant à estimer une vérité latente à partir d'annotations multiples. Comme dans l'approche "meilleure catégorisation" 4.1, les frontières de segments sont ignorées et les catégories sont propagées aux phrases correspondantes. L'algorithme EM (*Expectation–Maximization*)(Dempster *et al.*, 1977) est ensuite appliqué pour estimer, pour chaque phrase, la catégorie la plus probable en intégrant la fiabilité des annotateurs. Cette fiabilité est apprise automatiquement en comparant les annotations individuelles aux labels latents inférés, ce qui permet d'attribuer un poids spécifique à chaque annotateur dans l'estimation finale. Les labels obtenus servent ensuite à reconstruire la segmentation en regroupant les phrases consécutives partageant la même catégorie.

L'atout principal de cette approche réside dans la prise en compte explicite de la fiabilité des annotateurs, ce qui améliore la pertinence des labels choisis. Le traitement au niveau de la phrase permet également un choix plus fin de ces labels. La taille des segments reconstruits dépend toutefois de la dynamique thématique des émissions : des changements fréquents conduisent à des segments courts, tandis que des émissions monothématiques produisent des segments plus longs.

HMM Cette approche est inspérée des travaux de (Nguyen *et al.*, 2017) et repose sur l'application d'un modèle de Markov caché (HMM) (Rabiner, 1989) aux annotations d'une même émission afin d'estimer les catégories les plus probables pour chaque phrase de transcription. L'aspect séquentiel des phrases se prête particulièrement bien à une modélisation par HMM. L'inférence des catégories peut en effet être formulé comme l'estimation d'une vérité latente correspondant aux labels réels des phrases, lesquels peuvent être modélisés par les états cachés du HMM. La probabilité qu'un même thème reste le même entre deux phrases successives est plus élevée que celle d'un changement de thématique, ce qui correspond aux probabilités de transition d'un HMM. Les annotations produites par les annotateurs sont considérées comme des observations, tandis que la probabilité d'observer une catégorie donnée sachant le label réel correspond au modèle d'émission du HMM. L'ensemble de ces éléments montre l'adéquation du cadre probabiliste des HMM à notre problématique. Comme pour l'algorithme Dawid & Skene, une phase de reconstruction de la segmentation est ensuite effectuée en regroupant les phrases successives partageant la même catégorie.

Les principaux avantages de cette approche résident dans la modélisation explicite des dépendances séquentielles entre phrases. Elle permet également de corriger certaines catégorisations erronées et d'attribuer des labels à des phrases initialement non catégorisées, renforçant ainsi la robustesse face au bruit. Comme pour l'approche Dawid & Skene, la longueur des segments reconstruits dépend toutefois de la structure thématique des émissions.

4.3 Répartition du jeu de données selon la qualité

L'accord inter-annotateur, calculé par émission, varie de $-0{,}45$ à $0{,}90$, révélant une forte hétérogénéité marquée de la qualité des annotations : certaines sont plus fiables que d'autres. Cette variabilité peut influencer directement l'évaluation des modèles, leurs performances dépendant de la fiabilité des annotations de référence. Un déséquilibre entre données fiables et données bruitées est ainsi susceptible de biaiser les résultats. Afin d'analyser la robustesse des modèles face à des niveaux de qualité contrastés, nous construisons plusieurs sous-jeux de données en fonction du degré d'accord inter-annotateur (IAA) observé pour chaque émission. Les émissions présentant un IAA supérieur à 0,8 constituent le sous-jeu ***Gold*** (qualité excellente). Celles dont l'IAA est compris entre 0,6 et 0,8 forment le sous-jeu ***Silver*** (bonne qualité). Les émissions avec un IAA compris entre 0,3 et 0,6 sont regroupées dans le sous-jeu ***Bronze*** (qualité intermédiaire). Enfin, les émissions dont l'IAA est inférieur à 0,3 composent le sous-jeu ***Unreliable***, correspondant aux données de faible qualité.

4.4 Approche de constitution du jeu de données

L'approche de curation adoptée repose sur un processus structuré visant à améliorer la qualité et la cohérence des annotations. À partir des données brutes produites par les annotateurs, nous mettons en œuvre une procédure itérative de détection et d'exclusion des valeurs aberrantes (*outliers*). Cette procédure (voir la figure 2) consiste à identifier les annotateurs présentant un comportement atypique, sur la base de la moyenne de leur coefficient de Krippendorff. Un seuil d'exclusion, fixé comme une petite proportion du premier quartile (Q1), permet d'écarter de manière itérative les annotateurs insuffisamment fiables jusqu'à stabilisation des mesures d'accord. Le résultat de cette phase est un corpus épuré, caractérisé par des annotations plus consensuelles, une réduction de l'impact des annotateurs peu fiables et une amélioration de la cohérence globale des données.

Ensuite, nous appliquons la méthode de répartition du jeu de données décrite dans la section 4.3, en le divisant en plusieurs sous-ensembles en fonction du niveau de qualité des annotations. Cette étape aboutit à la constitution de quatre sous-jeux : **Gold**, **Silver**, **Bronze** et **Unreliable**.

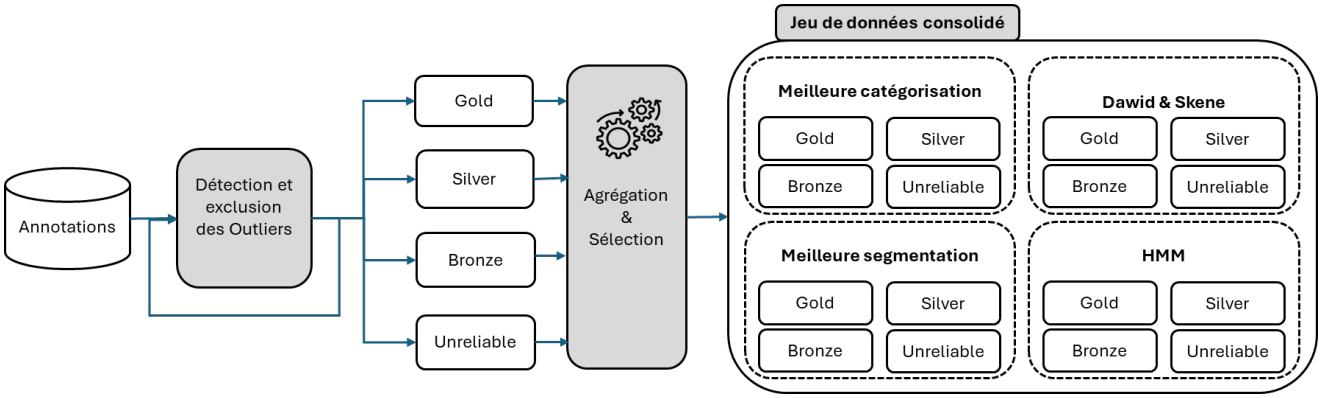


FIGURE 2 – Curation et constitution du jeu de données

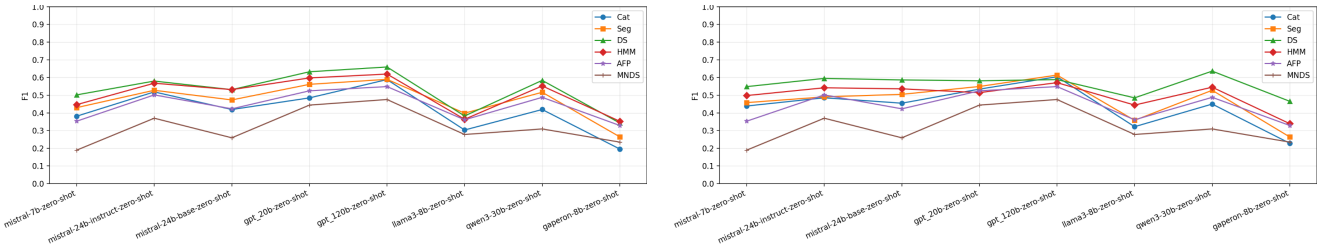
Enfin, nous appliquons aux différents sous-ensembles les quatre approches d'agrégation présentées dans les sections 4.1 et 4.2. Cette étape aboutit à la construction d'un jeu de données consolidé décliné en quatre variantes : (**CAT**) meilleure catégorisation, (**SEG**) meilleure segmentation, (**DS**) Dawid & Skene et (**HMM**) HMM. Chacune de ces variantes est elle-même déclinée en quatre sous-jeux : ***Gold***,

Silver, ***Bronze*** et ***Unreliable*** (cf. figure 2). Le volume total de ce jeu de données varie entre 3200 et 4100 exemples selon la méthode d'agrégation utilisée.

5 Évaluation

5.1 Évaluation de la qualité du jeu de données

L'évaluation de la qualité du jeu de données repose sur une comparaison expérimentale avec deux jeux de référence : celui constitué à partir de dépêches AFP, issu du projet OTMedia (Hervé *et al.*, 2013) et aligné sur le référentiel IPTC, ainsi que le jeu de données MN-DS présenté dans la section 2. L'hypothèse est que si plusieurs modèles de langue produisent un classement cohérent des différents jeux en termes de performances, ce consensus constitue un indicateur indirect de leur niveau de qualité relatif. Les expériences ont été menées en zero-shot à l'aide de plusieurs modèles, sélectionnés selon des critères liés à leur architecture, à leur niveau d'instruction et à leur taille. La liste des modèles utilisés est la suivante : Mistral-24b-instruct et Mistral-24b-base (deux modèles fournis par *Mistral AI* dans le cadre du projet ArGiMi [3]), Mistral-7b-Instruct [4], GPT-OSS-120b [5], GPT-OSS-20b [6], Llama-3-8b-Instruct [7], Qwen-3-30b-Instruct [8] et Gaperon-8b-SFT [9]. L'évaluation est réalisée sur les quatre sous-jeux de données (*Gold*, *Silver*, *Bronze* et *Unreliable*). Un ensemble d'instructions a été élaboré à partir des données annotées (voir en annexe la figure A.7 pour un exemple de prompt). La figure 3a montre que l'ensemble des modèles obtient de meilleures performances sur notre jeu de données, toutes variantes confondues, que sur les jeux AFP et MN-DS. Les résultats observés sur AFP sont globalement comparables à ceux obtenus avec la variante "**CAT**" (meilleure catégorisation). En revanche, les performances mesurées sur MN-DS sont sensiblement inférieures à celles observées sur les autres jeux en français. Ces observations suggèrent que notre jeu de données présente un niveau de qualité satisfaisant.



(a) évaluation sur l'ensemble du jeu de données (b) évaluation sur le sous-jeu de données "*unreliable*"

FIGURE 3 – évaluation de la qualité du jeu de données

Nous avons ensuite reproduit l'expérimentation en nous concentrant exclusivement sur le sous-ensemble "*unreliable*" de notre jeu de données (figure 3b). La tendance générale reste compa-

3. https://www.ina.fr/institut-national-audiovisuel/equipe-recherche/projet-argimi
4. https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
5. https://huggingface.co/openai/gpt-oss-120b
6. https://huggingface.co/openai/gpt-oss-20b
7. https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
8. https://huggingface.co/Qwen/Qwen3-30B-A3B-Instruct-2507
9. https://huggingface.co/almanach/Gaperon-1125-8B-SFT

rable : les meilleures performances sont majoritairement obtenues sur les variantes issues de notre jeu, bien que l'écart soit moins prononcé que dans l'expérimentation précédente. Les variantes fondées sur la meilleure catégorisation et la meilleure segmentation sont parfois rejointes, voire légèrement dépassées, par le jeu AFP. En revanche, MN-DS occupe systématiquement la dernière position.

5.2 Classement des modèles génératifs

Nous avons établi un classement des modèles génératifs (LLMs) en fonction de leurs performances en zero-shot sur les sous-jeux (*Gold*, *Silver*, *Bronze*, *Unreliable*) ainsi que sur l'AFP et MN-DS.

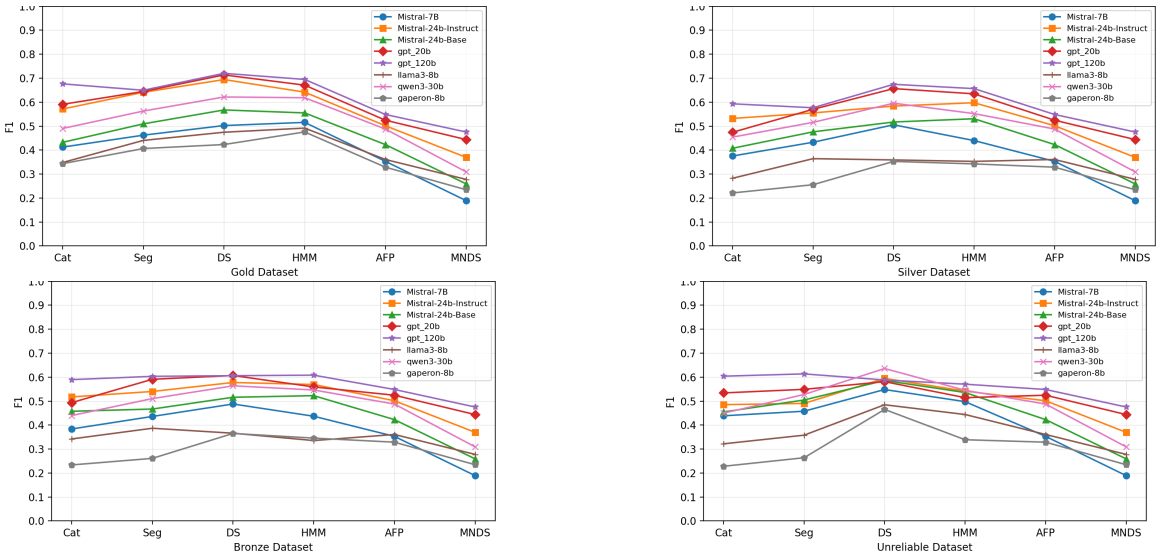


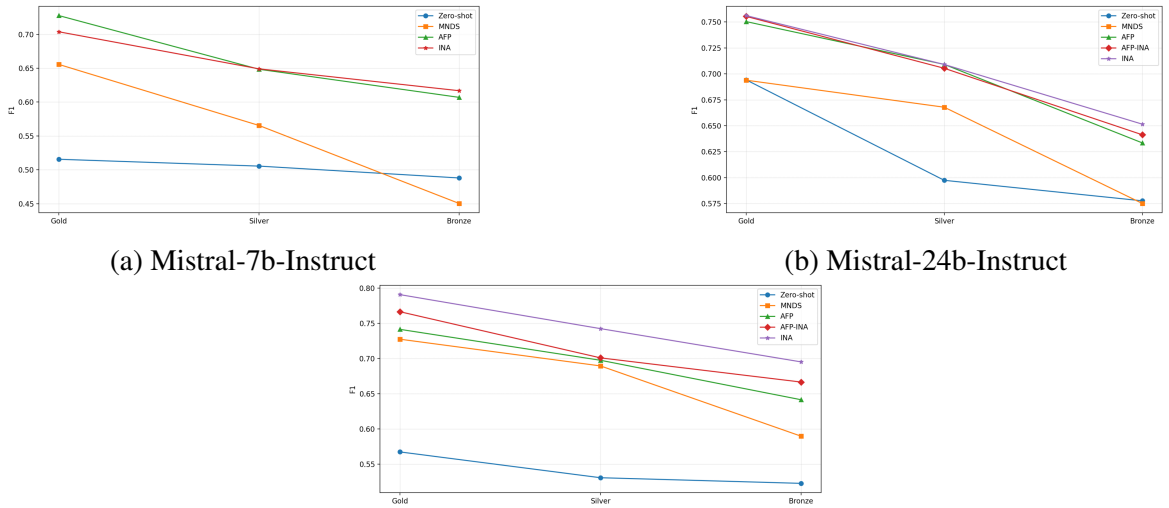
FIGURE 4 – Classement des modèles génératifs (zero-shot)

Les résultats (cf. figure 4) indiquent que GPT-OSS-120b occupe systématiquement la première position, suivi de GPT-OSS-20b, quel que soit le sous-ensemble considéré. Ces deux modèles, issus de la même architecture mais diffèrent par le nombre de paramètres, mettent en évidence l'effet de la taille du modèle sur les performances. Mistral-24b-Instruct se classe généralement en troisième position et dépasse notamment Qwen-3-30b, pourtant plus volumineux, ce qui suggère que l'architecture constitue également un facteur déterminant. Mistral-24b-Base, faiblement instruit, se situe en cinquième position, confirmant l'impact positif de l'instruction. Il surpasse néanmoins des modèles plus petits (Mistral-7b, LLaMA-3-8b et Gaperon-8b) confirmant à nouveau l'effet de la taille du modèle. Globalement, les familles de modèles se répartissent selon l'ordre suivant : GPT, Mistral, Qwen, LLaMA et Gaperon.

5.3 Évaluation des modèles fine-tunés

Nous avons procédé au fine-tuning des seuls modèles Mistral afin d'évaluer dans quelle mesure leurs versions adaptées peuvent surpasser les modèles de la famille GPT, qui dominent le classement précédent. Quatre jeux de données ont été utilisés : MN-DS, AFP, AFP-INA (combinaison du corpus AFP et du sous-ensemble *Unreliable*) ainsi que INA (le sous-ensemble *Unreliable*). Des informations complémentaires sur le processus de fine-tuning sont fournies dans la section A.2 en annexe.

Nous avons conduit une série d'évaluations des modèles fine-tunés afin d'analyser : 1) leur sensibilité et leur robustesse face au bruit, 2) l'écart de performance entre les configurations zero-shot et fine-tunées en fonction de la qualité des données, 3) classement des jeux de données d'entraînement, 4) l'impact de la langue du corpus d'entraînement, et 5) l'effet du niveau d'instruction sur l'efficacité du fine-tuning. Les résultats détaillés de ces évaluations sont présentés en annexe, section A.3, tableaux : *Gold* 1 ; *Silver* 2, *Bronze* 3, *Unreliable* 4, Ensemble du jeu de données 5.



(a) Mistral-7b-Instruct

(b) Mistral-24b-Instruct

(c) Mistral-24b-Base

FIGURE 5 – Évaluation des modèles génératifs fine-tunés

Les résultats (cf. figure 5) mettent en évidence une relation claire entre la qualité des données et les performances : celles-ci décroissent progressivement du niveau *Gold* vers *Silver* puis *Bronze*. Le sous-ensemble *Unreliable* n'apparaît pas dans cette évaluation, car il a également servi de données d'entraînement pour certains modèles (AFP-INA, INA). Cette tendance confirme que l'augmentation du bruit dans les annotations entraîne une dégradation systématique des performances. Par ailleurs, l'écart entre zero-shot et fine-tuning diminue à mesure que la qualité des données se dégrade. Sur *Gold*, le fine-tuning apporte un gain substantiel. Ce gain s'atténue sur *Silver* puis devient marginal sur *Bronze*, où les performances zero-shot peuvent parfois égaler, voire dépasser, celles des modèles adaptés, notamment pour Mistral-24b-Instruct.

Concernant le classement des jeux d'entraînement, celui issu du sous-ensemble *Unreliable* (INA) permet d'obtenir les meilleures performances. Ce résultat s'explique par la proximité entre ces données et le jeu d'évaluation, tous deux constitués de transcriptions de programmes audiovisuels. À l'inverse, les données issues de l'AFP et de MN-DS relèvent principalement de textes écrits, introduisant un décalage de domaine. Le jeu de données combinant l'AFP et la partie *Unreliable* (AFP-INA) se classe en deuxième position, devant les dépêches AFP utilisées seules. Cette amélioration suggère que l'intégration de données audiovisuelles apporte une information complémentaire pertinente, bien que l'ajout des dépêches AFP puisse également introduire un certain bruit par rapport à l'utilisation exclusive de *Unreliable*. Le jeu de données AFP se classe ensuite en troisième position, devant MN-DS. Cela peut s'expliquer en partie par le fait que MN-DS est majoritairement constitué de textes en anglais, ce qui limite son adaptation à une tâche de classification en français.

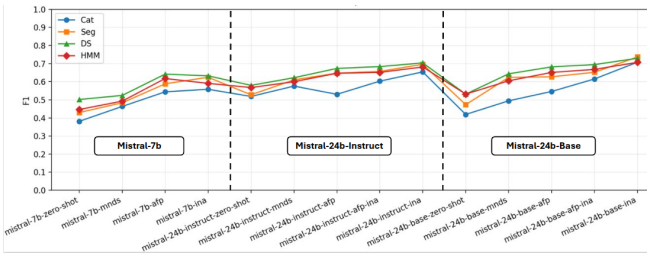
Concernant la langue d'entraînement, les modèles fine-tunés sur le corpus Anglais MN-DS obtiennent des performances correctes en français, mais demeurent inférieurs à ceux entraînés sur des données en Français. Le transfert entre langues de l'anglais vers le français apparaît donc possible, mais moins

efficace qu'un apprentissage réalisé directement dans la langue cible.

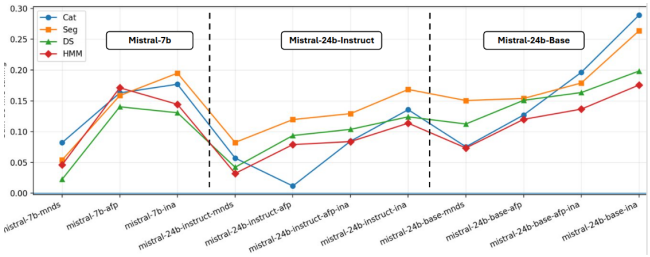
Enfin, l'analyse de l'effet de l'instruction montre que les modèles instruits (notamment Mistral-24b-Instruct) bénéficient d'un avantage en zero-shot. Toutefois, après fine-tuning, les modèles non instruits (Mistral-24b-Base) enregistrent souvent des gains plus importants et peuvent dépasser les modèles instruits dans plusieurs configurations, suggérant une plus grande capacité d'adaptation lors de l'apprentissage supervisé.

5.4 Évaluation des approches de constitution du jeu de données

Les différentes variantes du jeu de données ont été construites à partir de plusieurs approches de sélection et d'agrégation d'annotations issues de multiples annotateurs (cf. sections 4.1 et 4.2). Il est donc essentiel d'en évaluer l'impact afin de déterminer dans quelle mesure ces choix méthodologiques influencent les performances des modèles. Pour cette analyse, nous avons mobilisé les modèles fine-tunés. En effet, les méthodes d'agrégation ont pour objectif d'estimer une "vérité latente", correspondant aux labels les plus fiables parmi ceux fournis par les annotateurs. Les modèles fine-tunés ont un objectif similaire : en étant adaptés aux données, ils sont supposés mieux capturer leur structure et se rapprocher des labels les plus pertinents. Les approches d'agrégation et le fine-tuning partagent ainsi un objectif commun, à savoir l'estimation la plus fidèle possible des catégories sous-jacentes.



6a Évaluation des méthodes d'agrégation de données



6b Gain du fine-tuning

Les résultats (cf.figure 6a) indiquent que les méthodes probabilistes obtiennent globalement les meilleures performances, devant les approches fondées sur l'accord inter-annotateur. Parmi les méthodes probabilistes, Dawid & Skene surpasse le modèle HMM. Du côté des approches basées sur l'accord inter-annotateur, la stratégie reposant sur la meilleure segmentation (SEG) dépasse celle fondée sur la meilleure catégorisation (CAT). Dans l'ensemble, ces résultats suggèrent que les méthodes probabilistes permettent d'approcher plus efficacement la vérité latente (c'est-à-dire les labels de référence) que les stratégies basées sur l'accord inter-annotateur.

5.5 Gain du fine-tuning

À la suite de ces analyses, nous avons cherché à quantifier le gain apporté par le fine-tuning pour les trois modèles de la famille Mistral, en comparant leurs performances après fine-tuning à celles obtenues en configuration zero-shot. L'objectif est d'évaluer ce gain pour chaque modèle et d'en analyser la variation selon les différentes variantes du jeu de données.

La figure 6b illustre le gain de fine-tuning pour ces trois modèles. L'axe des abscisses correspond à la performance de référence obtenue en zero-shot. Une courbe est tracée pour chaque variante du jeu de données (CAT, SEG, DS, HMM), et chaque point correspond à l'écart de performance entre le

modèle en zero-shot et sa version fine-tunée. Les résultats indiquent que le modèle Mistral-24b-Base présente globalement le gain de fine-tuning le plus élevé, en particulier dans sa version fine-tunée sur le sous ensemble *Unreliable* (Mistral-24b-base-ina) pour la variante "meilleure catégorisation" (CAT). Ce comportement peut s'expliquer par le fait que la version Base, non instruite, dispose initialement de moins de connaissances spécifiques à la tâche que les modèles instruits. Le fine-tuning lui apporte donc un bénéfice relatif plus important. Par ailleurs, bien que les données INA (*Unreliable*) ne soient pas nécessairement de qualité optimale, elles demeurent plus proches du domaine d'évaluation que d'autres corpus tels que AFP ou MN-DS, ce qui peut également contribuer au gain observé. Pour les deux modèles instruits (Mistral-7b-Instruct et Mistral-24b-Instruct), le modèle Mistral-7b-Instruct présente un gain de fine-tuning plus marqué. Cette observation suggère qu'un modèle de plus petite taille bénéficie davantage d'un ajustement supervisé qu'un modèle plus volumineux déjà fortement paramétré et préalablement instruit. Dans le cas du modèle Mistral-24b-Instruct, une partie substantielle des connaissances pertinentes semble déjà intégrée, ce qui limite l'apport du fine-tuning.

Enfin, l'analyse selon les variantes du jeu de données montre que les variantes construites à partir de méthodes probabilistes (Dawid & Skene et HMM) présentent des gains de fine-tuning plus faibles que celles issues d'approches dérivées de l'accord inter-annotateur (Meilleure catégorisation et Meilleure segmentation). Ce résultat est en accord avec les observations présentées dans la section 5.4 selon laquelle les méthodes probabilistes fournissent déjà une estimation plus proche de la vérité sous-jacente ; le fine-tuning dispose alors d'une marge d'amélioration plus réduite que dans le cas des variantes issues de l'accord inter-annotateur.

6 Conclusion

Dans cet article, nous avons proposé un benchmark dédié à l'évaluation de la catégorisation de contenus audiovisuels à partir de leurs transcriptions. Nous avons introduit un jeu de données issu d'annotations manuelles réalisées sur un corpus diversifié d'émissions télévisées et radiophoniques diffusées entre 1982 et 2025. Nous avons également mené plusieurs niveaux d'évaluation portant à la fois sur la qualité du jeu de données, sur les méthodes de construction utilisées et sur les performances de modèles de langue génératifs appliqués à ce corpus.

En perspectives, nous prévoyons d'explorer de nouvelles stratégies d'agrégation et de sélection des annotations, notamment des approches basées sur l'apprentissage profond (par exemple des modèles neuronaux de type *LSTM*, *Transformers* ou modèles de langue), afin d'évaluer dans quelle mesure elles permettent de se rapprocher davantage de la vérité latente que les méthodes probabilistes. Nous envisageons également d'améliorer la qualité du jeu de données, soit par une intervention manuelle (correction et enrichissement des annotations à l'aide de l'outil utilisé par les annotateurs), soit par des approches automatiques consistant à entraîner un modèle sur les sous-ensembles *Gold* et *Silver* afin de corriger ou compléter les sous-ensembles *Bronze* et *Unreliable*. Du point de vue de l'évaluation, nous envisageons également d'expérimenter d'autres architectures, notamment des architectures discriminatives et contrastives ainsi que des architectures end-to-end opérant directement sur les annotations brutes, sans passer par la phase de curation d'un jeu de données consolidé.

Ce jeu de données sera mis à disposition de la communauté scientifique après libération des droits associés aux contenus. Il sera publié conjointement avec quatre autres jeux de données issus du même corpus, dédiés aux tâches suivantes : segmentation thématique, extraction et liage d'entités nommées, détection et extraction de citations, ainsi que résolution de coréférences.

Remerciements

Ces travaux sont partiellement financés par le programme France 2030 dans le cadre du projet ArGiMi (n° BPI DOS0238736). L'auteur remercie les documentalistes et les technicien·ne·s de gestion des données multimédia (TGDM) de l'INA pour leur précieuse contribution à la campagne d'annotation du projet.

Références

DAWID A. P. & SKENE A. M. (1979). Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Applied Statistics*, **28**(1), 20. DOI : 10.2307/2346806.

DEMPSTER A. P., LAIRD N. M. & RUBIN D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society : Series B (Methodological)*, **39**(1), 1–22. DOI : 10.1111/j.2517-6161.1977.tb01600.x.

HERVÉ N., VIAUD M.-L., THIÈVRE J., SAULNIER A., CHAMP J., LETESSIER P., BUISSON O. & JOLY A. (2013). OTMedia : the French TransMedia news observatory. In *Proceedings of the 21st ACM international conference on Multimedia*, p. 441–442, Barcelona Spain : ACM. DOI : 10.1145/2502081.2502260.

KRIPPENDORFF K. (1980). Content analysis : An introduction to its methodology.

LEWIS D. (1987). Reuters-21578 Text Categorization Collection. DOI : 10.24432/C52G6M.

MISRA R. (2022). News Category Dataset. arXiv :2209.11429 [cs], DOI : 10.48550/arXiv.2209.11429.

MITCHELL T. (1997). Twenty Newsgroups. DOI : 10.24432/C5C323.

NGUYEN A. T., WALLACE B., LI J. J., NENKOVA A. & LEASE M. (2017). Aggregating and Predicting Sequence Labels from Crowd Annotations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 299–309, Vancouver, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/P17-1028.

PETUKHOVA A. & FACHADA N. (2023). MN-DS : A Multilabeled News Dataset for News Articles Hierarchical Classification. *Data*, **8**(5), 74. arXiv :2212.12061 [cs], DOI : 10.3390/data8050074.

PEVZNER L. & HEARST M. A. (2002). A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, **28**(1), 19–36. DOI : 10.1162/089120102317341756.

RABINER L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, **77**(2), 257–286. DOI : 10.1109/5.18626.

ZHANG X., ZHAO J. & LECUN Y. (2016). Character-level Convolutional Networks for Text Classification. arXiv :1509.01626 [cs], DOI : 10.48550/arXiv.1509.01626.

A Annexe

A.1 Exemple de prompt

System : Tu es un assistant et tu dois réaliser une tâche de classification thématique en affectant au texte ci-après (texte qui suit l'instruction User) une ou plusieurs classes parmi les classes thématiques suivantes : ['**Environnement / Protection de la nature**', '**Société/Démographie**', **'Politique/Élections'**, '**Politique / Relations internationales / Organisation internationale**', **...**]

User : Nicolas Sarkozy est donc élu avec 64,5% des voix. En deuxième, Bruno Le Maire, 29,2% des suffrages. Et puis Hervé Mariton en troisième, 6,3%. Nicolas Sarkozy succède donc au trio Juppé-Fillon-Raffarin qui avait suivi la démission de Jean-François Copé.

Assistant : ['Politique/Élections']

FIGURE A.7 – Exemple de prompt

A.2 Fine-tuning des modèles génératifs

Le fine-tuning a été réalisé selon la méthode LoRA à l'aide du framework **Mistral-Finetune**[10]. L'entraînement a été effectué sur une machine équipée de 4 GPU H100, avec des hyperparamètres légèrement ajustés selon les modèles. Par exemple, le *weight decay* est fixé à 0,1 pour **Mistral-24b-Base** contre 0,001 pour **Mistral-24b-Instruct**. Le rang LoRA est de 64 pour ces deux modèles et de 16 pour **Mistral-7b-Instruct**, qui est par ailleurs quantifié en 4 bits. Le *learning rate* est fixé à $6*10^{-5}$ pour les deux premiers modèles et à $1*10^{-4}$ pour Mistral-7b. Enfin, le *warmup* est fixé à 0,05 pour l'ensemble des modèles.

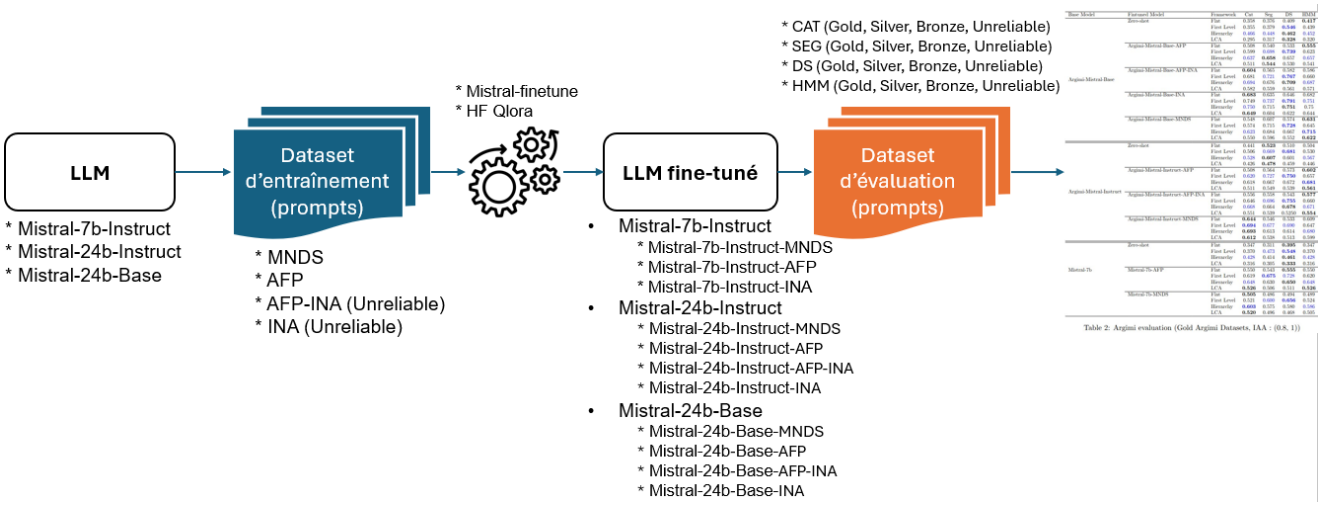


FIGURE A.8 – Fine-tuning des modèles Mistral

A.3 Résultats d'évaluation des modèles Mistral fine-tunés

Nous avons évalué ces modèles en calculant le F1-score selon 4 Frameworks :

— ***Flat*** : Toutes les catégories prédites et annotées sont traitées à plat, indépendamment de leur hiérarchie.

10. https://github.com/mistralai/mistral-finetune

— ***Hierarchical*** : Les relations hiérarchiques entre catégories sont intégrées dans le calcul du F1-score.
— ***First-Level*** : Seul le premier niveau de la hiérarchie qui est pris en compte pour le calcul de F1 score.
— ***LCA*** : Une variante du framework hiérarchique où la distance entre catégories est calculée à partir de leur plus petit ancêtre commun dans l'arborescence (**LCA, *Lowest Common Ancestor***).

Les tableaux 1, 2, 3, 4 et 5 montrent les résultats d'évaluation sur les sous ensembles : *Gold*, *Silver*, *Bronze*, *Unreliable* ainsi que l'ensemble du jeu de données. Les valeurs en **gras** correspondent aux meilleures performances pour chaque métrique et pour chaque modèle. Les valeurs en **bleu** indiquent la meilleure performance obtenue pour chaque modèle selon la méthode d'agrégation. Les valeurs en **gras et bleu** indiquent la meilleure performance globale pour un modèle donné, toutes métriques et méthodes d'agrégation confondues.

Base Model	Fintuned Model	Framework	Cat	Seg	DS	HMM
Mistral-24b-Base	Zero-shot	Flat	0.358	0.349	0.405	**0.407**
		First Level	0.417	0.509	**0.567**	0.555
		Hierarchy	0.432	0.426	0.470	**0.473**
		LCA	0.289	0.290	0.338	**0.341**
	Mistral-24b-Base-AFP	Flat	0.508	0.540	0.533	**0.551**
		First Level	0.599	0.698	0.739	**0.742**
		Hierarchy	0.637	0.658	0.657	**0.663**
		LCA	0.511	**0.544**	0.530	0.542
	Mistral-24b-Base-AFP-INA	Flat	**0.604**	0.565	0.582	0.575
		First Level	0.681	0.721	**0.767**	0.759
		Hierarchy	0.694	0.676	**0.709**	0.680
		LCA	**0.582**	0.559	0.561	0.539
	Mistral-24b-Base-INA	Flat	**0.683**	0.635	0.646	0.623
		First Level	0.749	0.737	**0.791**	0.745
		Hierarchy	0.750	0.715	**0.751**	0.713
		LCA	**0.649**	0.604	0.622	0.583
	Mistral-24b-Base-MNDS	Flat	0.548	0.607	**0.574**	0.564
		First Level	0.574	0.715	**0.728**	0.702
		Hierarchy	0.623	0.684	**0.667**	0.654
		LCA	0.550	**0.596**	0.552	0.545
Mistral-24b-Instruct	Zero-shot	Flat	0.456	0.463	**0.511**	0.470
		First Level	0.506	0.641	**0.694**	0.642
		Hierarchy	0.571	0.580	**0.605**	0.565
		LCA	0.437	**0.456**	0.453	0.419
	Mistral-24b-Instruct-AFP	Flat	0.508	0.564	**0.573**	0.551
		First Level	0.620	0.727	**0.750**	0.743
		Hierarchy	0.618	0.667	**0.672**	0.662
		LCA	0.511	**0.549**	0.539	0.533
	Mistral-24b-Instruct-AFP-INA	Flat	0.556	**0.558**	0.543	0.548
		First Level	0.646	0.696	**0.755**	0.740
		Hierarchy	0.668	0.664	**0.678**	0.673
		LCA	**0.551**	0.539	0.525	0.533
	Mistral-24b-Instruct-INA	Flat	**0.658**	0.594	0.630	0.622
		First Level	0.712	0.695	**0.756**	0.747
		Hierarchy	**0.722**	0.688	0.719	0.715
		LCA	**0.620**	0.571	0.582	0.576
	Mistral-24b-Instruct-MNDS	Flat	**0.644**	0.546	0.533	0.527
		First Level	**0.694**	0.677	0.690	0.644
		Hierarchy	**0.693**	0.613	0.614	0.599
		LCA	**0.612**	0.538	0.513	0.495
Mistral-7b-Instruct	Zero-shot	Flat	0.314	0.326	**0.361**	0.323
		First Level	0.349	0.462	0.502	**0.516**
		Hierarchy	0.412	0.432	**0.460**	0.441
		LCA	0.305	0.327	**0.347**	0.336
	Mistral-7b-Instruct-AFP	Flat	0.550	0.543	**0.555**	0.528
		First Level	0.619	0.675	**0.728**	0.708
		Hierarchy	0.648	0.630	**0.650**	0.640
		LCA	**0.526**	0.506	0.511	**0.512**
	Mistral-7b-Instruct-INA	Flat	**0.605**	0.561	0.576	0.548
		First Level	0.650	0.685	**0.704**	0.677
		Hierarchy	**0.672**	0.644	0.660	0.644
		LCA	**0.544**	0.511	0.515	0.511
	Mistral-7b-Instruct-MNDS	Flat	**0.505**	0.486	0.494	0.449
		First Level	0.521	0.600	**0.656**	0.588
		Hierarchy	**0.603**	0.575	0.580	0.528
		LCA	**0.520**	0.496	0.468	0.430

TABLE 1 – Évaluation des modèles Mistral sur le sous-ensemble ***Gold***

Base Model	Finetuned Model	Framework	Cat	Seg	DS	HMM
Mistral-24b-Base	Zero-shot	Flat	0.273	0.284	**0.336**	0.319
		First Level	0.408	0.476	0.517	**0.531**
		Hierarchy	0.352	0.370	0.407	**0.409**
		LCA	0.248	0.265	**0.290**	0.287
	Mistral-24b-Base-AFP	Flat	0.436	0.483	**0.548**	0.509
		First Level	0.540	0.619	**0.698**	0.678
		Hierarchy	0.545	0.576	**0.626**	0.605
		LCA	0.433	**0.544**	0.495	0.477
	Mistral-24b-Base-AFP-INA	Flat	0.492	0.476	**0.526**	0.498
		First Level	0.595	0.609	**0.701**	0.680
		Hierarchy	0.596	0.574	**0.624**	0.605
		LCA	0.485	0.460	**0.526**	0.471
	Mistral-24b-Base-INA	Flat	0.555	0.548	**0.596**	0.553
		First Level	0.621	0.609	**0.743**	0.717
		Hierarchy	0.647	0.654	**0.675**	0.643
		LCA	0.537	0.547	**0.548**	0.515
	Mistral-24b-Base-MNDS	Flat	0.455	0.500	**0.523**	0.486
		First Level	0.519	0.636	**0.690**	0.657
		Hierarchy	0.530	0.592	**0.619**	0.578
		LCA	0.451	0.508	**0.511**	0.478
Mistral-24b-Instruct	Zero-shot	Flat	0.357	0.354	**0.357**	0.341
		First Level	0.532	0.555	0.584	**0.597**
		Hierarchy	**0.500**	0.480	0.489	0.475
		LCA	0.362	**0.363**	0.357	0.351
	Mistral-24b-Instruct-AFP	Flat	0.444	0.515	**0.553**	0.493
		First Level	0.547	0.661	**0.709**	0.688
		Hierarchy	0.550	0.604	**0.634**	0.595
		LCA	0.456	0.490	**0.513**	0.470
	Mistral-24b-Instruct-AFP-INA	Flat	0.485	0.469	**0.528**	0.473
		First Level	0.606	0.614	**0.705**	0.671
		Hierarchy	0.589	0.583	**0.629**	0.577
		LCA	0.485	0.467	**0.487**	0.455
	Mistral-24b-Instruct-INA	Flat	0.485	0.475	**0.556**	0.516
		First Level	0.596	0.641	**0.709**	0.693
		Hierarchy	0.590	0.588	**0.629**	0.611
		LCA	0.495	0.497	**0.498**	0.476
	Mistral-24b-Instruct-MNDS	Flat	0.511	0.517	**0.524**	0.491
		First Level	0.592	0.658	**0.668**	0.647
		Hierarchy	**0.603**	0.591	**0.603**	0.572
		LCA	0.522	0.503	0.500	**0.468**
Mistral-7b-Instruct	Zero-shot	Flat	0.228	0.231	**0.259**	0.218
		First Level	0.329	0.433	**0.505**	0.440
		Hierarchy	0.375	0.381	**0.402**	0.357
		LCA	0.286	**0.300**	**0.304**	0.278
	Mistral-7b-Instruct-AFP	Flat	0.455	0.449	**0.472**	0.460
		First Level	0.558	0.599	**0.648**	0.647
		Hierarchy	0.558	0.545	**0.571**	0.555
		LCA	0.450	0.433	**0.448**	0.434
	Mistral-7b-Instruct-INA	Flat	0.448	0.463	**0.483**	0.441
		First Level	0.581	**0.649**	**0.649**	0.620
		Hierarchy	0.551	**0.574**	0.570	0.557
		LCA	0.441	**0.467**	0.429	0.421
	Mistral-7b-Instruct-MN-DS	Flat	0.425	**0.426**	0.424	0.415
		First Level	**0.479**	0.526	0.565	0.554
		Hierarchy	0.507	**0.517**	0.506	0.484
		LCA	**0.433**	0.437	0.412	0.389

TABLE 2 – Évaluation des modèles Mistral sur le sous-ensemble ***Silver***

Base Model	Finetuned Model	Framework	Cat	Seg	DS	HMM
Mistral-24b-Base	(Zero-shot	Flat	0.305	0.296	**0.317**	0.315
		First Level	0.458	0.467	0.516	**0.523**
		Hierarchy	0.409	0.395	**0.513**	0.438
		LCA	**0.272**	0.261	0.281	**0.303**
	Mistral-24b-Base-AFP	Flat	0.427	0.495	**0.506**	0.451
		First Level	0.522	0.627	**0.642**	0.615
		Hierarchy	0.548	0.593	0.591	0.564
		LCA	0.412	**0.453**	0.450	0.417
	Mistral-24b-Base-AFP-INA	Flat	0.475	0.478	**0.510**	0.449
		First Level	0.599	0.642	**0.667**	0.629
		Hierarchy	0.602	0.597	**0.616**	0.573
		LCA	0.468	**0.454**	0.451	0.411
	Mistral-24b-Base-INA	Flat	0.528	0.532	**0.546**	0.489
		First Level	0.610	**0.695**	0.682	0.652
		Hierarchy	**0.643**	0.637	0.629	0.597
		LCA	**0.512**	**0.512**	0.485	0.438
	Mistral-24b-Base-MNDS	Flat	0.361	**0.459**	0.453	0.411
		First Level	0.410	**0.590**	0.576	0.549
		Hierarchy	0.476	**0.542**	0.514	0.485
		LCA	0.359	**0.454**	0.426	0.383
Mistral-24b-Instruct	Zero-shot	Flat	**0.379**	0.332	0.357	0.319
		First Level	0.517	**0.540**	0.578	0.569
		Hierarchy	**0.517**	0.478	0.502	0.479
		LCA	**0.369**	0.338	0.344	0.335
	Mistral-24b-Instruct-AFP	Flat	0.426	0.468	**0.502**	0.436
		First Level	0.530	**0.633**	0.630	0.607
		Hierarchy	0.530	0.578	**0.589**	0.556
		LCA	0.404	**0.448**	0.442	0.410
	Mistral-24b-Instruct-AFP-INA	Flat	0.474	0.457	**0.504**	0.417
		First Level	0.579	0.629	**0.641**	0.608
		Hierarchy	**0.599**	0.576	0.596	0.548
		LCA	0.**445**	0.433	0.429	0.395
	Mistral-24b-Instruct-INA	Flat	0.459	0.483	**0.512**	0.457
		First Level	0.566	0.638	**0.652**	0.630
		Hierarchy	0.588	0.589	**0.608**	0.567
		LCA	**0.454**	0.445	0.450	0.408
	Mistral-24b-Instruct-MNDS	Flat	0.465	0.459	**0.467**	0.542
		First Level	0.508	0.572	**0.575**	0.556
		Hierarchy	**0.537**	0.527	0.518	0.487
		LCA	**0.452**	0.441	0.425	0.383
Mistral-7b-Instruct	Zero-shot	Flat	0.205	0.218	**0.241**	0.209
		First Level	0.310	0.436	**0.488**	0.437
		Hierarchy	**0.384**	0.388	**0.405**	0.400
		LCA	**0.286**	0.292	**0.297**	0.279
	Mistral-7b-Instruct-AFP	Flat	**0.412**	0.388	0.411	0.395
		First Level	0.522	0.579	**0.607**	0.575
		Hierarchy	**0.547**	0.527	**0.533**	0.513
		LCA	**0.424**	0.406	0.405	0.384
	Mistral-7b-Instruct-INA	Flat	**0.436**	0.431	0.412	0.369
		First Level	0.559	**0.617**	0.589	0.559
		Hierarchy	**0.567**	0.553	0.529	0.506
		LCA	**0.437**	0.418	0.381	0.347
	Mistral-7b-Instruct-MN-DS	Flat	**0.365**	0.315	0.348	0.321
		First Level	0.373	0.397	**0.450**	0.440
		Hierarchy	**0.430**	0.397	0.409	0.390
		LCA	**0.351**	0.327	0.328	0.306

TABLE 3 – Évaluation des modèles Mistral sur le sous-ensemble ***Bronze***

Base Model	Finetuned Model	Framework	Cat	Seg	DS	HMM
Mistral-24b-Base	Zero-shot	Flat	0.288	0.272	**0.338**	0.283
		First Level	0.455	0.505	**0.586**	0.536
		Hierarchy	0.447	0.424	**0.471**	0.445
		LCA	0.276	0.271	0.305	**0.310**
	Mistral-24b-Base-AFP	Flat	0.423	0.485	**0.514**	0.413
		First Level	0.451	0.586	**0.701**	0.624
		Hierarchy	0.559	0.620	**0.651**	0.556
		LCA	0.406	0.441	**0.475**	0.410
	Mistral-24b-Base-MNDS	Flat	0.346	**0.450**	0.446	0.357
		First Level	0.362	0.557	**0.574**	0.520
		Hierarchy	0.484	**0.586**	0.562	0.465
		LCA	0.346	**0.437**	0.420	0.342
Mistral-24b-Instruct	Zero-shot	Flat	0.343	0.324	**0.371**	0.361
		First Level	0.474	0.490	**0.595**	0.542
		Hierarchy	0.485	0.457	**0.507**	0.446
		LCA	0.327	0.298	**0.334**	0.303
	Mistral-24b-Instruct-AFP	Flat	0.393	0.422	**0.518**	0.386
		First Level	0.418	0.557	**0.648**	0.590
		Hierarchy	0.518	0.547	**0.617**	0.524
		LCA	0.363	0.400	**0.449**	0.374
	Mistral-24b-Instruct-MNDS	Flat	**0.438**	0.394	0.424	0.343
		First Level	**0.535**	0.524	0.531	0.508
		Hierarchy	**0.520**	0.475	0.495	0.437
		LCA	**0.394**	0.382	0.360	0.330
Mistral-7b-Instruct	Zero-shot	Flat	0.190	0.222	**0.278**	0.192
		First Level	0.439	**0.458**	0.548	0.498
		Hierarchy	0.402	0.387	**0.460**	0.401
		LCA	**0.303**	0.283	**0.326**	0.300
	Mistral-7b-Instruct-AFP	Flat	0.453	0.440	**0.490**	0.341
		First Level	0.541	0.594	**0.649**	0.584
		Hierarchy	0.606	0.588	0.604	0.477
		LCA	**0.441**	0.433	0.433	0.345
	Mistral-7b-Instruct-MN-DS	Flat	**0.359**	0.309	0.305	0.262
		First Level	0.417	0.425	**0.458**	0.400
		Hierarchy	**0.488**	0.413	0.402	0.361
		LCA	**0.366**	0.318	0.300	0.273

TABLE 4 – Évaluation des modèles Mistral sur le sous-ensemble ***Unreliable***

Base Model	Finetuned Model	Framework	Cat	Seg	DS	HMM
Mistral-24b-Base	Zero-shot	Flat	0.278	0.287	**0.333**	0.322
		First Level	0.419	0.474	**0.531**	**0.531**
		Hierarchy	0.378	0.389	0.423	**0.428**
		LCA	0.253	0.264	0.294	**0.300**
	Mistral-24b-Base-AFP	Flat	0.414	0.485	**0.518**	0.470
		First Level	0.520	0.627	**0.682**	0.651
		Hierarchy	0.530	0.588	**0.611**	0.580
		LCA	0.414	0.460	**0.471**	0.441
	Mistral-24b-Base-AFP-INA	Flat	0.495	0.493	**0.515**	0.468
		First Level	0.600	0.652	**0.694**	0.668
		Hierarchy	0.615	0.609	**0.626**	0.591
		LCA	**0.485**	0.468	0.474	0.440
	Mistral-24b-Base-INA	Flat	**0.616**	0.607	0.587	**0.535**
		First Level	0.692	**0.737**	0.729	0.707
		Hierarchy	**0.708**	0.703	0.678	0.644
		LCA	**0.585**	0.580	0.538	0.496
	Mistral-24b-Base-MNDS	Flat	0.408	0.484	**0.495**	0.443
		First Level	0.488	0.624	**0.643**	0.605
		Hierarchy	0.494	0.580	**0.577**	0.531
		LCA	0.422	**0.490**	0.473	0.426
Mistral-24b-Instruct	Zero-shot	Flat	0.372	0.335	**0.340**	0.318
		First Level	0.518	0.528	**0.580**	0.568
		Hierarchy	**0.504**	0.469	0.485	0.466
		LCA	**0.378**	0.335	0.339	0.329
	Mistral-24b-Instruct-AFP	Flat	0.415	0.474	**0.522**	0.458
		First Level	0.530	0.647	**0.673**	0.647
		Hierarchy	0.529	0.586	**0.605**	0.567
		LCA	0.409	0.460	**0.474**	0.429
	Mistral-24b-Instruct-AFP-INA	Flat	0.480	0.481	**0.511**	0.452
		First Level	0.596	0.657	**0.683**	0.651
		Hierarchy	0.603	0.599	**0.614**	0.572
		LCA	**0.471**	0.467	0.460	0.426
	Mistral-24b-Instruct-INA	Flat	**0.565**	0.558	0.554	0.512
		First Level	0.644	0.696	**0.704**	0.681
		Hierarchy	**0.654**	0.651	0.640	0.620
		LCA	0.523	**0.525**	0.494	0.466
	Mistral-24b-Instruct-MNDS	Flat	**0.495**	0.479	0.484	0.453
		First Level	0.557	0.610	**0.622**	0.600
		Hierarchy	**0.576**	0.559	0.552	0.524
		LCA	**0.488**	0.468	0.449	0.420
Mistral-7b-Instruct	Zero-shot	Flat	0.216	0.223	**0.242**	0.216
		First Level	0.332	0.430	**0.502**	0.446
		Hierarchy	0.381	0.386	0.394	**0.396**
		LCA	0.285	**0.292**	0.290	0.286
	Mistral-7b-Instruct-AFP	Flat	0.415	0.416	**0.440**	0.426
		First Level	0.544	0.588	**0.642**	0.618
		Hierarchy	0.539	0.536	**0.550**	0.533
		LCA	**0.428**	0.418	0.425	0.408
	Mistral-7b-Instruct-INA	Flat	0.438	0.435	**0.446**	0.406
		First Level	0.556	0.624	**0.632**	0.591
		Hierarchy	**0.558**	0.557	0.557	0.536
		LCA	**0.431**	0.428	0.411	0.381
	Mistral-7b-Instruct-MN-DS	Flat	0.378	0.387	**0.392**	0.351
		First Level	0.438	0.483	**0.524**	0.492
		Hierarchy	0.463	**0.472**	0.468	0.426
		LCA	**0.390**	**0.390**	0.377	0.339

TABLE 5 – Évaluation des modèles Mistral sur l'ensemble du jeu de données