Évaluation de la qualité de rapport des essais cliniques avec des larges modèles de langue

Mathieu Laï-king* — **Patrick Paroubek***

* *Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, 91400 Orsay, France*
mathieu.lai-king@lisn.upsaclay.fr, patrick.paroubek@lisn.upsaclay.fr

RÉSUMÉ. *La qualité de rapport est un sujet important dans les articles de recherche sur les essais cliniques car elle peut avoir un impact sur les décisions cliniques prises. Nous testons la capacité des larges modèles de langue à évaluer la qualité de rapport de ce type d'article en utilisant les standards fusionnés pour la rédaction d'essais thérapeutiques (CONSORT). Nous créons un corpus d'évaluation à partir de deux études sur la vérification de la qualité de rapport de résumés d'articles avec les standards CONSORT définis pour les résumés. Nous évaluons ensuite la capacité de différents larges modèles de langue génératifs (du domaine général ou adaptés au domaine biomédical) à correctement évaluer chaque critère CONSORT avec différentes méthodes de requêtage (*prompting*) connues. Notre meilleure association de modèle et de méthode de requêtage obtient 85 % d'exactitude.*

MOTS-CLÉS : *articles biomédicaux, qualité des données textuelles, larges modèles de langue.*

TITLE. *Evaluation of Clinical Trials Reporting Quality using Large Language Models*

ABSTRACT. *Reporting quality is an important topic in clinical trial research articles, as it can have an impact on the clinical decisions made. In this article, we test the ability of large language models to assess the reporting quality of this type of article using the Consolidated Standards of Reporting Trials (CONSORT). We create an evaluation corpus from two studies on abstract reporting quality with CONSORT-abstract standards. We then evaluate the ability of different large generative language models (from the general domain or adapted to the biomedical domain) to correctly assess CONSORT criteria with different known prompting methods. Our best combination of model and prompting method achieves 85 % accuracy.*

KEYWORDS: *Biomedical Articles, Text Data Quality, Large Language Models.*

1. Introduction

Les essais cliniques, et plus particulièrement les essais contrôlés randomisés ou ECR (*Randomized Controlled Trials* ou *RCT* en anglais), sont la référence en matière d'évaluation de l'efficacité d'un traitement. Ces essais cliniques sont généralement rapportés après leur clôture dans des articles scientifiques. Ces articles doivent mentionner toutes les caractéristiques, les méthodes et les résultats obtenus au cours de l'essai en question, de manière claire et précise, afin que la communauté médicale puisse prendre des décisions cliniques correctes et éclairées.

Des standards indiquant quels éléments doivent être rapportés et comment ils doivent l'être ont été créés afin de garantir la qualité de rapport des ECR lorsqu'ils sont publiés dans un article. Ces standards établissent des règles pour correctement rapporter tous les aspects importants d'un essai clinique. Les standards largement utilisés sont les « Standards fusionnés pour la rédaction d'essais thérapeutiques » (ou *Consolidated Standards of Reporting Trials* en anglais) : CONSORT. Ces standards ont d'abord été établis en 1996 (Begg *et al.*, 1996) puis ont connu deux mises à jour en 2001 (Altman *et al.*, 2001) et en 2010 (Moher *et al.*, 2010). Nous nous référons dorénavant aux derniers standards sous l'abréviation *CONSORT-2010*. Il existe également de nombreuses extensions pour les différentes formes que peuvent prendre les essais cliniques et également pour différentes parties d'un article scientifique, comme les standards CONSORT pour les résumés (Hopewell *et al.*, 2008). Enfin, il est notable que depuis la création de ces normes, la qualité des rapports reste encore insuffisante dans plusieurs domaines médicaux, bien qu'elle se soit améliorée au fil des ans, comme le soulignent plusieurs études (Turner *et al.*, 2012 ; Warrier et Jayanthi, 2022 ; Wang *et al.*, 2021).

La réalisation de ces contrôles de qualité nécessite une analyse humaine experte et est coûteuse. Au fil des années, une augmentation du nombre d'essais cliniques a été constatée (Niforatos *et al.*, 2019), et plus encore lors de la pandémie de COVID-19 (Park *et al.*, 2021). C'est pourquoi il est de plus en plus important d'étudier si nous pouvons tirer parti des méthodes de traitement automatique des langues pour rendre l'évaluation de la qualité des publications d'essais cliniques plus efficace, et aussi pour permettre aux auteurs de rédiger des articles de meilleure qualité.

Du point de vue du traitement automatique des langues, l'arrivée des *Transformers* (Vaswani *et al.*, 2017) a ouvert de nouvelles possibilités d'applications. En particulier, l'utilisation de modèles génératifs préentraînés basés sur les *Transformers* a connu une croissance rapide ces dernières années. Ces modèles permettent de résoudre de nombreuses tâches d'analyse et de génération de texte (Brown *et al.*, 2020 ; Touvron *et al.*, 2023a ; Touvron *et al.*, 2023b). De plus, l'adaptation de ce type de modèle dans le domaine biomédical montre de nouvelles performances sur des tâches spécifiques à ce domaine (Luo *et al.*, 2022 ; Singhal *et al.*, 2023a) comme le corpus de questions-réponses MedQA (Jin *et al.*, 2020), comprenant des questions de médecine du style de

l'examen USMLE[1]. Aujourd'hui, il est même possible d'atteindre des performances de niveau expert dans les tâches de questions-réponses dans le domaine biomédical avec de très larges modèles comme Med-PalM 2 (Singhal *et al.*, 2023b).

C'est pourquoi nous étudions la capacité des larges modèles de langue génératifs à produire automatiquement une évaluation CONSORT des articles faisant état d'un essai clinique, en nous intéressant dans cette première étude à la partie de CONSORT spécifique au résumés d'articles. Nos contributions comprennent :

– l'adaptation à la tâche de question-réponse de 2 corpus anglais provenant d'évaluations CONSORT réalisées par des annotateurs experts pour des ECR qui concernent les interventions sur le COVID-19 d'une part, et d'autre part la prévention de la dépression chez les enfants et les adolescents ;

– l'évaluation de différentes variantes de larges modèles de langue génératifs publics sur cette tâche en utilisant différentes méthodes de requêtage (aussi appelées méthodes d'invite ou *prompting* en anglais). L'étude repose sur des modèles génératifs sans entraînement spécifique pour la tâche du fait de la petite taille des corpus, les corpus ne sont donc utilisés que pour l'évaluation.

2. Travaux existants

Plusieurs travaux ont été réalisés dans le domaine du traitement automatique des langues pour des tâches liées aux articles d'essais cliniques. Nous pouvons tout d'abord noter des travaux sur l'extraction d'informations. Par exemple, les entités PICO (*Population, Intervention, Comparator, Outcome*) sont des entités clés à détecter dans un essai clinique (Jin et Szolovits, 2018 ; Mutinda *et al.*, 2022 ; Wang *et al.*, 2022).

Nous notons ensuite quelques études qui se sont concentrées sur l'évaluation des risques de biais dans les articles d'ECR. Il s'agit de critères de qualité devant être vérifiés pour chaque article inclus lors d'une revue systématique d'essais cliniques, réalisée par des annotateurs humains en suivant un guide tel que celui de Cochrane : *Risk of Bias 2* (Sterne *et al.*, n.d.). Un exemple de critère de risque de biais est celui sur le processus d'assignation aléatoire des groupes, ce critère vérifie : si l'allocation de la séquence est aléatoire, si elle à été correctement cachée, si les différences entre les groupes suggèrent un problème avec le processus d'assignation aléatoire. Marshall *et al.* (2015) et Marshall *et al.* (2016) produisent un système complet basé sur des méthodes d'apprentissage automatique pour effectuer l'évaluation des risques de biais dans un article. Plus récemment, des méthodes basées sur des réseaux de neurones *Transformers*, et plus spécifiquement BERT (Devlin *et al.*, 2019), ont été appliquées à cette même tâche d'évaluation du risque de biais dans la littérature préclinique (Wang *et al.*, 2020). Cependant, il est encore difficile d'adopter ces méthodes automatiques

1. Examen de médecine des États-Unis.

dans une procédure standard d'évaluation du risque de biais pour les personnes pratiquant l'analyse d'articles (Jardim *et al.*, 2022).

Nous pouvons également noter des travaux sur la détection du « spin » (présentation inadéquate du résultat de la recherche) (Koroleva, 2020), avec différentes méthodes d'extraction d'information permettant notamment de mettre en évidence la relation entre la présentation d'un résultat par les auteurs et sa significativité statistique. En effet, ce phénomène est souvent observé et étudié dans les recherches évaluant la qualité des articles rapportant les essais cliniques (Bero *et al.*, 2021 ; Wang *et al.*, 2021).

Pour notre tâche spécifique d'évaluation automatique des critères CONSORT dans les articles d'ECR, Kilicoglu *et al.* (2021) ont crée le corpus CONSORT-TM composé de phrases d'articles annotées avec les éléments CONSORT-2010. Cependant, ce corpus a été annoté uniquement dans le but d'extraire les phrases liées à un critère, et non pas d'évaluer si le critère est vérifié ou non. Par ailleurs, Wang *et al.* (2020) ont produit un logiciel basé sur plusieurs méthodes de traitement automatique des langues et qui comprend une application graphique pour la génération automatique d'éléments CONSORT. Mais à notre connaissance, ils ne partagent pas publiquement leur code ni leurs données d'évaluation ; seule leur application graphique est disponible.

Avec l'essor des modèles de langue génératifs, la recherche sur les méthodes de requêtage a commencé à donner des résultats surprenants, comme l'étude de Kojima *et al.* (2022), montrant que l'ajout d'une simple phrase « *Let's think step by step* » à la requête, suscitant un raisonnement par étapes dans le modèle, améliore considérablement sa capacité à effectuer de nombreuses tâches liées au raisonnement logique. Des travaux récents ont même amélioré cette méthode en optimisant les requêtes qui sont décrites directement en langage naturel (Yang *et al.*, 2023). Enfin, Wei *et al.* (2022) montrent que l'ajout d'un exemple comprenant le raisonnement pour parvenir à la réponse avant de donner le résultat peut également améliorer les performances de plusieurs modèles génératifs. L'utilisation de ce type de méthode se révèle également intéressante pour les tâches de questions-réponses dans le domaine biomédical (Singhal *et al.*, 2023b).

3. Corpus

3.1. *Adaptation de corpus*

Nous avons d'abord cherché à constituer un corpus contenant plusieurs résumés d'articles, avec leur évaluation selon les critères CONSORT pour les résumés réalisée par des experts. Dans la suite de l'article, nous désignons ces critères CONSORT spécifiques aux résumés par CONSORT-RÉSUMÉ.

À partir d'une recherche dans la base de données publique PubMedCentral[2], nous avons identifié les études portant sur l'adhésion aux critères CONSORT-RÉSUMÉ (la requête exacte en anglais étant *adherence to CONSORT-abstract*). Nous avons ensuite relu chaque étude pour identifier dans le texte complet, dans la section « ressources supplémentaires »(*Supplementary Material*) et dans la section « déclaration de disponibilité des données » (*Data availability statement*) si les auteurs fournissaient publiquement le détail de l'évaluation CONSORT-RÉSUMÉ pour chaque résumé évalué par l'étude. Nous avons limité notre revue manuelle aux deux premières pages de la recherche (vingt articles par page = quarante articles au total). Sur les quarante articles examinés, seuls deux satisfaisaient nos critères de sélection, soit 5 % des articles examinés (2/40). Nous avons donc récupéré les données de ces deux études pour produire nos corpus.

Sachant que ces études n'ont pas été conçues en vue d'une analyse automatique, elles ne fournissent en général qu'un DOI[3], voire un titre et des métadonnées, mais pas le texte du résumé considéré pour l'étude. Nous avons donc cherché à le retrouver pour chaque entrée. Pour cela, nous avons tenté de récupérer son identifiant PubMed (PMID) : s'il n'était pas disponible, nous avons alors utilisé l'interface IDConverter[4] de PubMed pour l'obtenir à partir du DOI. Ensuite, une fois le PMID obtenu, nous avons utilisé une interface de la suite *Entrez Programming Utilities Help*[5] qui permet de récupérer le résumé d'un article à partir de son PMID.

3.2. *Description du corpus*

Nous obtenons finalement le corpus composé des deux sous-corpus suivants, que nous nommons CONSORT-QA et dont les statistiques sont présentées dans le tableau 1 :

– **CONSORT-QA-COVID** : Wang *et al.* (2021) publient une étude sur l'adhésion aux critères CONSORT-RÉSUMÉ sur des ECR concernant le COVID-19. Cette étude porte sur quarante résumés que nous parvenons tous à récupérer via l'interface *Entrez*. Pour chaque résumé, 16 points de la liste CONSORT sont considérés. Les critères sont évalués avec une mesure booléenne (critère vérifié ou non) ;

– **CONSORT-QA-DÉPRESSION** : Wiehn *et al.* (2022) publient à leur tour une étude sur l'adhésion aux critères CONSORT-RÉSUMÉ dans les rapports d'ECR sur la prévention de la dépression chez les enfants et les adolescents. Cette étude comprend également, en plus des ECR, des essais randomisés par grappes (*Cluster Randomized trials* ou *CRT* en anglais). Pour les essais par grappes, nous récupérons automatiquement les 63 résumés évalués. Pour les ECR, sur les 103 résumés évalués par les auteurs, nous en avons récupéré 84 automatiquement à l'aide de l'interface PubMed *Entrez*,

2. `https://www.ncbi.nlm.nih.gov/pmc/`
3. Le DOI (*Digital Object Identifier*) correspond à l'identifiant numérique d'un objet (physique, numérique ou abstrait) :`https://www.doi.org/`
4. `https://www.ncbi.nlm.nih.gov/pmc/tools/idconv/`
5. `https://www.ncbi.nlm.nih.gov/books/NBK25501/`

puis quinze autres en utilisant des recherches Google via leur DOI et/ou leur titre pour un total de 99 ECR ; les quatre articles que nous ne parvenons pas à récupérer n'ont pas de résultat avec la méthode précédente. Ce qui nous donne 162 résumés au total. Les auteurs adoptent une annotation différente des critères par rapport à l'étude précédente, différence que nous tentons d'homogénéiser. Pour le critère sur les sources de financement, les annotateurs se servent du texte complet et ajoutent à l'espace des valeurs possibles pour ce critère la mention « dans une autre section » (*in another section*) pour mentionner le fait que le critère est vérifié, mais dans le corps de l'article. Ainsi, pour ces valeurs, nous considérons que le critère n'est pas vérifié sachant que nous ne fournissons que le résumé à nos modèles de génération. Enfin, leur annotation contient trois niveaux de précision : « non rapporté », « rapporté de manière inadéquate », et « conforme ». Nous transformons donc ce schéma en une annotation booléenne, où un critère non vérifié englobe les annotations originales « non rapporté » et « rapporté de manière inadéquate ». Cela reste pertinent car dans ces deux cas le critère est en effet non vérifié. Enfin, pour ce corpus, les auteurs fournissent les valeurs d'accords inter-annotateurs pour chaque critère, sur lesquels nous nous basons afin de déterminer la difficulté du critère.

sujet du corpus	**nombre total de résumés**	**mots par résumé (moy.)**	**sections par résumé (moy.)**	**ratio total de réponses oui/non**
Dépression	99	225.90	3.69	0.30/0.70
COVID-19	40	339.50	5.58	0.48/0.52

TABLEAU 1. *Statistiques générales des corpus utilisés (moy. = moyenne). Les moyennes sont calculées sur tous les résumés du corpus.*

En ce qui concerne les critères CONSORT-RÉSUMÉ évalués dans les deux corpus, bien que ces critères proviennent du même standard, ils sont pour certains légèrement adaptés selon les besoins des évaluateurs, par exemple en subdivisant certains critères, voire en ajoutant un sous-critère pour avoir une annotation plus claire et précise. De plus, l'identifiant d'un critère spécifique (numéroté via un nombre et, optionnellement, une lettre : par exemple 13a) peut également varier selon les évaluateurs. Afin d'éviter une confusion, nous ajoutons une lettre au début de cet identifiant pour différencier les deux corpus. Les critères ainsi définis (sous forme de question pour la formulation de la tâche, voir section 4.1) sont présentés dans le tableau 2. Il faut noter que par rapport aux critères originaux CONSORT-RÉSUMÉ de Hopewell *et al.* (2008), certains ont été divisés en plusieurs sous-critères : par exemple, dans CONSORT-QA-DÉPRESSION, le critère vérifiant le mécanisme d'assignation secrète est divisé en trois. Cela montre que certains critères ont en fait besoin de plusieurs étapes de raisonnement afin d'arriver à la réponse.

Nous récupérons de l'étude sur la dépression les valeurs d'accord inter-annotateur pour chaque critère (via une métrique de type Kappa). En utilisant les correspondances entre les critères définis dans le corpus CONSORT-QA-COVID et CONSORT-

d_id	c_id	d_question	c_question
D01	C01	Is the study identified as randomized in the title ?	
D02	C02	Is there a structured summary of the trial design (e.g., parallel, crossover, cluster, non-inferiority) ?	
D03a	C03a	Are the eligibility criteria for participants mentioned ?	
D03b	C03b	Are the settings or locations where the data were collected stated in the abstract ?	
D04a	C04	Do the authors report essential features of the experimental intervention (if needed) ?	Are the interventions sufficiently detailed for each group (eg, when, how) ?
D04b		Do the authors report essential features of the comparison (= control) intervention (if needed) ?	
D05	C05	Are there specific objectives or hypothesis stated ?	Are there specific objectives or hypothesis stated ?
D06a	C06	Do the authors explicitly state the primary outcome as such (eg, primary / main / principal) ?	Are the primary outcomes clearly described for this trial in methods ?
D06b		Do the authors explicitly state when the primary outcome was assessed (time frame) ?	
	C07a		Is the random assignment declared (eg, random, randomized, randomization, random allocation) ?
D07	C07b	If they declared a random allocation to interventions (if they did not answer no), do the authors correctly report how they were allocated (e.g., computer-generated, random numbers, coin toss, etc.) ?	
	C07c		Are they referring to allocation concealment ?
D08a	C08	Do authors describe if participants were blinded ? (answer yes only if participants are blinded, do not care about caregivers or outcome assessors)	Are they mentioning whether or not participants, trial providers and data collectors were blinded ?
D08b		Do authors describe if the program deliverer (caregiver) were blinded ? (answer yes only if caregivers are blinded, do not care about participants or outcome assessors)	
D08c		Do authors describe if data collectors (outcome assessors, analysts) were blinded ? (answer yes only if data collectors are blinded, do not care about participants or caregivers)	
	C08a		Is there only a brief description of blinding (eg, single-blind, double-blind, triple-blind) ?
D09	C09	Are the numbers of participants randomized to each group clearly stated ?	
	C10		Is the Trial status (eg, on-going, closed to recruitment, closed to follow-up, etc.) mentionned ?
D10	C11	Are the numbers of participants analyzed for each group clearly stated (not the number randomized but the patients included in the analysis of the primary outcome) ?	
	C11a		Are the numbers of participants analyzed in accordance with the original grouping (eg, intention-to-treat analysis or pre-protocol analysis) ?
D11	C12a	For the primary outcome, is there a result for each group and the estimated effect size and its precision (e.g., 95% CI) ? (if one of them is missing, answer no)	For the primary outcome(s) , is there a summary report of results for each group ?
	C12b		For the primary outcome(s), is the estimated effect size clearly stated ?
	C12c		For the primary outcome(s), is the precision of the estimate (eg, 95%CI) clearly stated ?
D12	C13	Do the authors correctly mention the presence or absence of adverse events or side effects ?	
D13a		Do the authors state the conclusions of the trial ?	
D13b		Do the authors state implications for further research or clinical practice ?	
	C14		Are the general interpretations corresponding to the results ?
	C14a		Are the benefits and harms balanced in the conclusion ?
D14a	C15	Do the authors provide the trial registration number ?	Are the trial registration number and the name of trial register clearly stated ? Answer no if one of them is missing
D14b		Do the authors provide the name of the trial register ?	
D15	C16	Do the authors declare the source of funding ?	Do the authors declare the source of funding ?

TABLEAU 2. *Correspondance entre les questions/critères CONSORT pour le corpus* CONSORT-QA-DÉPRESSION *(préfixe « d ») et pour le corpus* CONSORT-QA-COVID *(préfixe « c »)*

QA-DÉPRESSION, nous assimilons les valeurs d'accord pour les critères définis dans le CONSORT-QA-COVID. Nous utilisons ces valeurs d'une part pour évaluer la difficulté de chaque critère et d'autre part pour déterminer si les modèles de langue génératifs

rencontrent plus de difficulté sur certains critères que sur d'autres. Nous présentons ces valeurs dans le tableau 3.

id	kappa
D01	0.96
D02	0.38
D03a	0.77
D03b	0.81
D04a	0.80
D04b	0.76
D05	0.73
D06a	0.91
D06b	0.69
D07	0.49

id	kappa
D08a	0.77
D08b	0.77
D08c	0.66
D09	0.95
D10	0.88
D11	0.94
D13a	0.75
D13b	0.74
D14a	1.00
D14b	0.98
D15	0.88

id	kappa
C01	0.96
C02	0.38
C03a	0.77
C03b	0.81
C04	0.78
C05	0.73
C06	0.80

id	kappa
C07b	0.49
C08	0.73
C09	0.95
C12a	0.94
C12b	0.94
C12c	0.94
C15	0.99
C16	0.88

TABLEAU 3. *Valeurs de kappa définies dans le corpus* CONSORT-QA-DÉPRESSION *pour chaque critère. Les valeurs pour le corpus* CONSORT-QA-COVID *ont été adaptées à partir des valeurs de l'autre corpus (c'est pourquoi certains identifiants sont manquants).*

Les corpus que nous obtenons sont relativement petits. C'est la raison pour laquelle nous avons décidé d'effectuer nos expériences sur des modèles génératifs en inférence uniquement, car un ajustement (aussi appelé affinage, et en anglais *fine-tuning*) aurait nécessité plus de données. Ces corpus nous servent donc uniquement à l'évaluation de nos méthodes.

3.3. *Filtrage des phrases des résumés*

Nous souhaitons aussi étudier l'influence des parties du résumé incluses dans la requête d'entrée du modèle. Pour cela, nous utilisons les travaux réalisés pour le corpus CONSORT-TM (Kilicoglu *et al.*, 2021). Ce corpus est constitué de cinquante articles complets sur des essais contrôlés randomisés, qui ont été annotés au niveau des phrases pour savoir si elles contiennent une information sur un critère CONSORT particulier (une phrase peut être utile à zéro, un ou plusieurs critères).

Il faut noter une limite importante du fait que ce corpus ne contient ni les titres, ni les résumés et n'évalue que les critères CONSORT sur la méthodologie. Nous définissons donc un tableau de conversion entre les critères de nos corpus et ceux pris en compte par CONSORT-TM.

Les modèles que nous utilisons sont entraînés par les auteurs du corpus CONSORT-TM[6]. Il s'agit de modèles BioBERT (Lee *et al.*, 2020) ajustés sur leur corpus entier (modèles servant uniquement à l'inférence) sur une tâche de classification multi-labels. Ces modèles existent en deux variantes : l'une entraînée sur les phrases du corpus seules (*text-only-50*) et l'autre entraînée sur les phrases précédées de la section du texte auxquelles elles appartiennent (*section-text-50*).

Sachant que certains des résumés de nos corpus sont structurés en sections, nous découpons ces derniers d'abord selon les sections identifiées. Puis pour séparer nos résumés en phrases, nous utilisons la librairie Python Scispacy (Neumann *et al.*, 2019) avec leur modèle *en_core_sci_lg*. Nous obtenons donc des phrases et optionnellement leur section associée si elle est disponible. Enfin, pour obtenir les critères pertinents pour chaque phrase, nous utilisons les modèles ajustés présentés ci-dessus et assignons les critères adéquats à chaque phrase selon les prédictions des modèles. Le modèle *text-only-50* est utilisé pour les phrases n'ayant pas de section définie et le modèle *section-text-50* est utilisé pour les phrases appartenant à une section.

Si la classification ne donne pas de résultat pour un critère dans un résumé (aucune des phrases n'est associée au critère), alors le contexte sera composé du résumé entier, de même que pour la méthode sans filtrage.

4. Méthodes

4.1. *Formulation de la tâche*

Nous avons choisi de considérer le problème de l'évaluation d'un résumé selon les critères CONSORT comme une tâche de question-réponse individuelle pour chaque critère avec une réponse booléenne (critère vérifié/critère rejeté). Pour cela, nous avons d'abord manuellement redéfini chaque critère sous forme de question, le développement de ces questions est précisé dans la section 4.4. Les questions pour les deux corpus sont fournies dans le tableau 2.

Ainsi, nous obtenons 4 272 paires de résumé/question au total pour nos deux corpus, que nous fournirons aux différents larges modèles génératifs choisis.

4.2. *Larges modèles de langue*

Nous utilisons plusieurs larges modèles de langue sur notre tâche. Ces modèles sont de tailles différentes et ont pour certains une version générale (entraînée sur des données non spécialisées) et une version adaptée au domaine biomédical (car notre tâche comporte des résumés d'articles de recherche du domaine biomédical).

6. Disponibles sur le lien suivant : `https://github.com/kilicogluh/CONSORT-TM/tree/master/bert`

Pour les modèles généraux, nous choisissons une gamme de modèles populaires récents dont les poids sont disponibles publiquement : Llama-2 (Touvron *et al.*, 2023b), Llama-3 (AI@Meta, 2024), Mistral (Jiang *et al.*, 2023), Mixtral (Jiang *et al.*, 2024), Bloomz (Muennighoff *et al.*, 2023), Phi-3 (Abdin *et al.*, 2024), Gemma (Team *et al.*, 2024) et Command-R-Plus (CohereForAI, 2024). Pour tous ces modèles, nous utilisons la version ajustée sur un jeu de données d'instruction lorsqu'elle est disponible. En effet, l'amélioration des modèles de langue en utilisant un ajustement par instruction après le préentraînement a déjà été démontrée sur de nombreuses tâches de question-réponse (Ouyang *et al.*, 2022) et également dans le domaine biomédical (Singhal *et al.*, 2023a). Nous utilisons aussi (lorsque c'est possible) la version ayant la plus grande taille de contexte d'entrée (car pour certaines de nos méthodes de requêtage, la taille d'entrée peut dépasser les 5 000 *tokens*).

Pour les modèles adaptés au domaine biomédical, nous utilisons trois variantes de modèles généraux (présentés ci-dessus) ajustées sur un corpus biomédical : Meditron (Chen *et al.*, 2023) ajusté à partir de Llama-2, BioMistral (Labrak *et al.*, 2024) ajusté à partir de Mistral-v0.1 et OpenBioLLM (Ankit Pal, 2024) ajusté à partir de Llama-3. Il ne s'agit pas des mêmes données et processus d'entraînement entre ces différents modèles car il n'existe pas forcément de variante biomédicale pour tous les modèles généraux et ils sont entraînés indépendamment les uns des autres.

4.3. *Stratégies de requêtage (ou* prompting*)*

Comme le montrent de nombreuses études, travailler sur l'entrée d'un modèle génératif peut grandement améliorer sa performance dans la résolution de différentes tâches (Wei *et al.*, 2022 ; Singhal *et al.*, 2023a). Ayant plusieurs critères et donc plusieurs questions pour chaque article, nous demandons au modèle de répondre à chaque question séparément. Nous testons les quatre stratégies suivantes.

– ***0-shot*** (ou zéro-exemple) : cette stratégie est la plus simple et la base des autres. Nous donnons d'abord un passage d'instruction générale, qui demande notamment de répondre à une question par oui ou non, puis le contexte, qui est le résumé de l'article à évaluer, et enfin la question du critère spécifique à évaluer. Pour cette méthode, nous limitons la génération à un unique *token*[7] en choisissant la première occurrence d'une des deux réponses possibles (*yes* ou *no* vu que les résumés et les requêtes sont écrits en anglais) dans la distribution fournie par le modèle, parmi les vingt *tokens* les plus probables. Si aucun des *tokens* de réponse n'est trouvé, nous considérons que le modèle hallucine (et cela sera donc compté comme une mauvaise réponse lors de l'évaluation).

– ***few-shot*** (ou quelques-exemples) : pour cette stratégie, nous ajoutons aux requêtes de la stratégie précédente quelques exemples provenant des corpus (CONSORT-

7. Mot anglais faisant référence aux éléments qui composent le vocabulaire d'un modèle de langue. Un *token* peut être une sous-partie de mot, un mot, un signe de ponctuation, un symbole, un espace, etc.

QA-COVID et CONSORT-QA-DÉPRESSION) et comprenant (pour chaque exemple) le résumé complet de l'article, la question du critère évalué et la réponse. Nous excluons pour cela cinq exemples par corpus. Nous testons trois configurations : *1-shot*, *3-shot* et *5-shot*, notamment car il s'agit des valeurs habituellement utilisées et obtenant généralement les meilleures performances pour les tâches de questions-réponses avec du *prompting* (Singhal *et al.*, 2023b). Ces exemples proviennent du même corpus et contiennent la même question que le résumé à évaluer (ils sont placés avant celui-ci dans la requête). Le *token* final de réponse est généré de la même manière que précédemment.

– ***1-shot-cot-orig*** (pour *Chain-of-Thought*, soit « chaîne de pensée » en français) : cette stratégie est similaire à la précédente, mais nous ajoutons aux exemples de la requête une explication avant la réponse et nous demandons également au modèle de générer une explication avant sa réponse (Wei *et al.*, 2022), cela permet en général d'améliorer les capacités des modèles génératifs sur les problèmes nécessitant plusieurs étapes de raisonnement. Pour générer l'explication, nous le faisons via un décodage glouton (pas d'échantillonnage) jusqu'au prochain saut de ligne (fin de l'explication) ou lorsque la longueur de l'explication dépasse 200 *tokens*. Le modèle génère ensuite le *token* de réponse comme précédemment. Il n'y a ici qu'un exemple par requête. Nous prenons un seul exemple car nous récupérons ceux donnés dans l'article CONSORT-RÉSUMÉ original de Hopewell *et al.* (2008), comprenant un seul exemple par critère (toujours positif, où le critère est donc vérifié), et nous adaptons manuellement les explications selon les sous-critères différemment définis entre nos 2 sources (tableau 2).

– ***few-shot-cot*** : cette méthode est la même que la précédente mais avec plusieurs exemples dans la requête. Étant donné que nos corpus ne fournissent pas d'annotation textuelle expliquant le choix de la réponse par l'annotateur, nous décidons d'annoter automatiquement les exemples avec un modèle de langue. Pour cela, nous choisissons le modèle *Llama-3* à 70 milliards de paramètres ajusté par instructions. Pour générer ces explications, nous fournissons la vraie réponse au modèle avant de générer l'explication afin qu'il soit guidé vers une explication plus cohérente. Nous le faisons pour les cinq exemples par corpus, comme cela est défini dans la stratégie *few-shot* [8].

Nous donnons un exemple de requête fourni au modèle dans la figure 1. Nos expériences sont effectuées sur des cartes graphiques NVIDIA A100 (nombre dépendant de la mémoire requise pour chaque modèle). Nous utilisons la librairie Python vLLM pour la génération (Kwon *et al.*, 2023).

4.4. *Développement des requêtes*

Le format de nos requêtes est similaire à celui utilisé dans l'article du modèle Med-PaLM 2 (Singhal *et al.*, 2023b). Pour redéfinir les critères en tant que questions, nous

8. Nous n'annotons pas davantage d'exemples, nous n'effectuons pas de validation ni n'appliquons de méthodes de sélection d'exemple afin de réduire le coût des expériences.

Instructions : The task is to verify a criterion from the Consolidated Standards of Reporting Trial (CONSORT) for a given abstract. The output should be yes or no (whether the criterion is met or not).

Context : """Title : A randomized [...]""".
Question : Are the numbers of participants analyzed for each group clearly stated (not the number randomized but the patients included in the analysis of the primary outcome) ?
Explanation : In the Results section, the authors report the number of participants analyzed for the primary outcome, saying : '300 were included in the analysis of the primary outcome'. They also precise the numbers for each group '(100 in the acetaminophen group, 100 in the ibuprofen group, and 100 in the codeine group)'. So the numbers of participants analyzed for each group is clearly stated.
Answer : Yes

Context : """Title : Online attentional [...]""".
Question : Are the numbers of participants analyzed for each group clearly stated (not the number randomized but the patients included in the analysis of the primary outcome) ?
Explanation :

FIGURE 1. *Exemple de requête pour la stratégie* 1-shot-cot-orig *dans le corpus* CONSORT-QA-DÉPRESSION*; ici, les résumés d'articles ont été abrégés par souci de place et nous ne fournissons qu'un seul exemple (mais pour les stratégies* few-shot*, nous en fournissons plusieurs).*

nous basons sur les formulations originales utilisées dans chaque étude[9] sur l'évaluation CONSORT-RÉSUMÉ (présentées dans la section 3.2).

De plus, certaines questions ont été ajustées, soit car nous jugions qu'elles n'étaient pas assez claires pour résoudre la tâche (même pour un annotateur humain), soit car nous avions observé des erreurs récurrentes en *0-shot* sur les exemples donnés dans la déclaration CONSORT-RÉSUMÉ originale de Hopewell *et al.* (2008) (qui ne se trouvent pas dans les données de test). En particulier, cela a abouti à donner aux modèles génératifs des instructions plus précises que les formulations mentionnées par les auteurs des études, qui, pour certaines, nous semblaient incomplètes (notamment car les auteurs des études ne fournissent que des tableaux avec de courtes descriptions de chaque

9. Pour CONSORT-QA-COVID, il s'agit du tableau 2 de leur article et pour CONSORT-QA-DÉPRESSION, le tableau S7 donné dans leur matériel supplémentaire.

critère, et non un guide d'annotation précis pour chaque critère). Nous utilisons également les définitions originales des critères CONSORT-RÉSUMÉ et CONSORT-2010 pour ajouter ces précisions (Hopewell *et al.*, 2008 ; Moher *et al.*, 2010).

4.5. *Évaluation des performances de nos méthodes*

Pour apprécier la capacité du modèle à évaluer un critère à partir d'un résumé, nous mesurons l'exactitude au niveau du critère et aussi au niveau du corpus entier. Nous considérons chaque exemple, contenant un résumé et un critère, de manière égale : il s'agit donc d'une micromoyenne, accordant plus d'importance au corpus CONSORT-QA-DÉPRESSION contenant un plus grand nombre d'exemples.

Pour les modèles n'ayant pas une taille de contexte suffisante pour les stratégies de type *few-shot*, nous les évaluons uniquement en *0-shot*.

Par ailleurs, pour évaluer l'influence qu'à le filtrage des phrases du résumé sur l'efficacité de la requête (section 3.3), nous réduisons l'espace des critères pour l'évaluation. En effet, comme mentionné précédemment, les modèles que nous utilisons pour le filtrage ne comprennnent pas tous les critères (et les sous-critères) de nos deux corpus. Ainsi, pour chaque corpus, on considère uniquement l'intersection entre l'ensemble des critères prédits par les modèles de classification de phrase et l'ensemble des critères du corpus auquel appartient l'exemple de test considéré. De plus, on ne considère que le modèle *Llama-2* afin de réduire les coûts de calcul[10].

5. Résultats

Nous comparons d'abord les performances globales des modèles (soit la micromoyenne telle que définie dans la section 4.5). Nous discutons ensuite des performances détaillées par critère pour le modèle obtenant les meilleures performances globales. Nous étudions ensuite la corrélation entre ces performances au niveau du critère et la difficulté de chaque critère. Enfin, nous analysons manuellement la cohérence des réponses correctes pour les générations de type *Chain-of-Thought* (à la fois les réponses justes et les réponses fausses du meilleur modèle).

5.1. *Comparaison de la performance globale des modèles*

Nous comparons l'exactitude des différents modèles et stratégies de requêtage testés dans la figure 2.

Nous observons les effets suivants :

10. Nous nous sentons d'autant plus confortés dans ce choix par les faibles différences de performances observées dans nos premières expérimentations avec le filtrage.

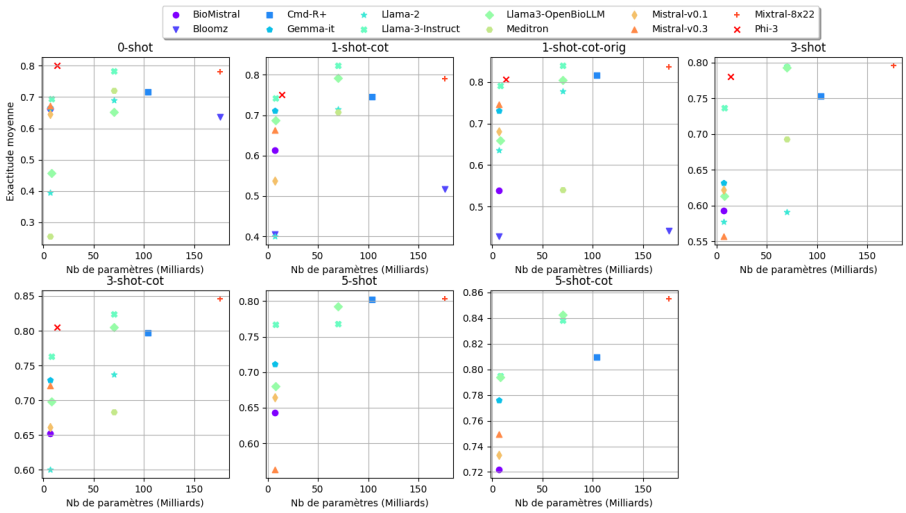


FIGURE 2. *Exactitude moyenne (sur les exemples concaténés de nos 2 corpus) de tous les modèles testés pour chaque stratégie de requêtage*

– effet de la taille des modèles : ici, comme attendu, pour un même modèle disponible en différentes tailles (tel que *Llama-3*, que nous évaluons avec sa version à huit milliards de paramètres et sa version à 70 milliards de paramètres), les modèles plus grands obtiennent de meilleures performances. Néanmoins, certains modèles plus petits en taille obtiennent des performances très proches de modèles plus grands, voire meilleures pour la stratégie *zero-shot*, où *Phi-3* obtient les meilleures performances (surpassant donc des modèles jusqu'à 10 fois plus grands en nombre de paramètres). Cependant, la raison de ces différences reste difficile à déterminer sachant que les auteurs de ces modèles ne précisent pas systématiquement la quantité de données utilisées pour le préentraînement et les critères de qualité utilisés pour sélectionner leurs données ;

– effet de l'utilisation d'annotations automatiques : en comparant les méthodes *1-shot-cot* et *1-shot-cot-orig*, nous remarquons une réduction des performances en utilisant les annotations générées automatiquement. Néanmoins, ces annotations automatiques permettent une amélioration globale car elles nous donnent notamment l'accès à la méthode *5-shot-cot* qui obtient les meilleures performances ;

– effet des différentes méthodes de requêtage : de manière globale, les stratégies contenant des exemples donnent de meilleures performances ainsi que l'ajout d'explications du modèle pour les stratégies COT. Nous analysons qualitativement les explications générées dans les sections 5.4 et 5.5 ;

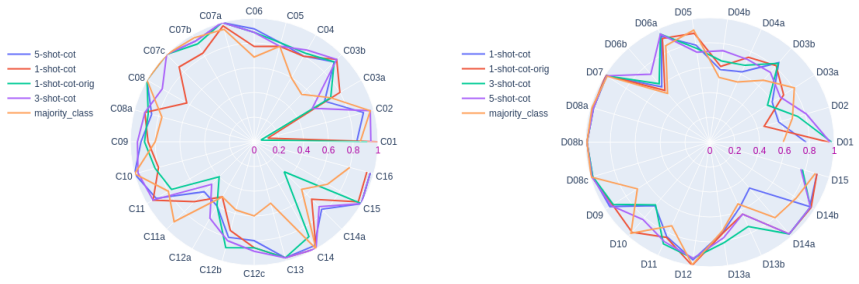
– effet de l'ajustement de modèles au domaine biomédical : globalement, on observe que l'équivalent biomédical des modèles généraux (BioMistral, Meditron et OpenBioLLM) obtient de moins bonnes performances que son équivalent général.

Cela peut être dû à la nature de la tâche qui ne nécessite pas forcément de connaissances biomédicales spécifiques (sauf pour certains critères). En effet, pour la majorité des critères, ils peuvent être vérifiés simplement via de la recherche d'information dans le contexte fourni, tâche pour laquelle les modèles de langue généraux sont déjà performants. Néanmoins cette différence reste surprenante car ces modèles biomédicaux sont entraînés sur des données très similaires aux articles fournis en entrée (soit des articles de recherche biomédicaux).

Enfin, le modèle obtenant les meilleurs résultats est le modèle Mixtral-8x22B ajusté par instructions avec la méthode *5-shot-cot*, avec une exactitude de 85 %. Ces performances globales sont encourageantes pour ce type d'approche, mais montrent que cette tâche reste encore difficile même pour les meilleurs modèles disponibles publiquement.

5.2. *Performances détaillées par critère CONSORT*

Nous montrons dans la figure 3 les performances au niveau de chaque critère afin d'observer les critères CONSORT-RÉSUMÉ pour lesquels les méthodes de requêtage plus sophistiquées améliorent la performance. Nous n'étudions ici que notre meilleur modèle, soit *Mixtral-8x22B*, sur nos deux corpus, nous affichons également la classe majoritaire pour chaque critère (néanmoins il s'agit de la classe majoritaire sur le jeu de test, auquel nos modèles n'ont pas accès, ils n'ont que 5 exemples au maximum dans leur contexte).



(a) CONSORT-QA-COVID **(b)** CONSORT-QA-DÉPRESSION

FIGURE 3. *Exactitude évaluée au niveau de chaque critère sur nos corpus pour le modèle Mixtral-8x22B avec nos stratégies de requêtage les plus performantes. La performance de la classe majoritaire sur le jeu de test est également affichée (en orange).*

Nous pouvons remarquer que le modèle parvient à obtenir des performances très élevées principalement pour les critères ayant des classes déséquilibrées (tels que D07, D08a, D08b, D08c, C07c, C08). En effet, certains critères sont toujours vérifiés dans nos corpus, les modèles peuvent donc être biaisés par rapport aux exemples fournis qui peuvent également être biaisés en faveur de la classe majoritaire (mais nous

n'analysons pas ce phénomène ici). Nous remarquons aussi que l'ajout d'exemples est majoritairement bénéfique, bien que ces exemples soient générés automatiquement.

5.3. *Corrélation avec la difficulté des critères*

Nous calculons la corrélation entre les performances du meilleur modèle (*Mixtral-8x22B* avec la stratégie *5-shot-cot*) et la difficulté des critères, que nous définissons par l'accord inter-annotateur défini dans l'étude sur la dépression (section 3.2), nous supposons donc que s'il y a un plus grand désaccord entre les annotateurs humains, cela signifie que la description du critère est imprécise et donc plus difficile à évaluer. Pour cela, nous mesurons la corrélation de Pearson (Benesty *et al.*, 2009) entre la performance (exactitude moyenne) sur chaque critère et les valeurs de kappa présentées dans le tableau 3. Nous calculons ces corrélations pour les différentes méthodes de requêtage avec les 3 meilleurs modèles et les présentons dans la figure 4

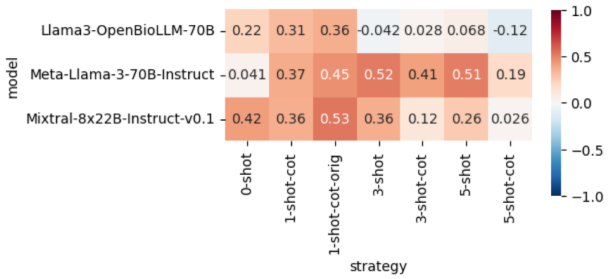


FIGURE 4. *Corrélation de Pearson entre les performances des modèles par critère et la difficulté de ces critères*

Nous remarquons qu'en moyenne, les performances utilisant dans la requête des explications manuelles provenant des critères originaux CONSORT-RÉSUMÉ sont plus corrélées au kappa que celles utilisant des explications générées automatiquement. Les annotations manuelles influencent ainsi le modèle à générer un comportement plus corrélé à celui d'un annotateur humain. Les méthodes sans explication (*0-shot*, *3-shot* et *5-shot*) ont également une plus grande corrélation que leur équivalent avec explication. Enfin, plus on a d'exemples annotés automatiquement, plus la corrélation est faible. Et ayant vu précédemment que ces méthodes ont de meilleures performances, cela signifie que les modèles sont davantage capables de généraliser sur des exemples plus diversifiés et de mieux comprendre la tâche (ils ont donc moins de difficultés sur les critères plus difficiles).

5.4. *Analyse qualitative de l'explication du modèle pour les réponses correctes*

Pour la méthode *few-shot-cot*, nous souhaitons vérifier si les réponses correctes trouvées par le modèle sont corrélées avec l'explication fournie. Nous échantillonnons

donc 50 vrais positifs et 50 vrais négatifs des générations du modèle *Mixtral-8x22B* (Jiang *et al.*, 2024) avec la méthode *5-shot-cot*. Nous annotons ensuite manuellement chacun de ces exemples de manière binaire : soit nous jugeons l'explication est jugée correcte (donc cohérente avec le contexte, la question CONSORT fournie en entrée ainsi que la réponse donnée par le modèle), soit elle est incorrecte (donc incohérente avec le contexte, la question ou la réponse donnée). Nous obtenons ainsi 99 réponses sur cent où l'explication fournie par le modèle est cohérente. Les explications fournies via la méthode *few-shot-cot* semblent donc en grande majorité cohérentes avec le contexte et la réponse donnée par le modèle lorsque le modèle répond juste. Nous ne réalisons cependant cette annotation qu'avec un seul annotateur (M. L.-K.), étudiant en thèse n'étant pas spécialisé dans le domaine biomédical, mais ayant acquis les connaissances suffisantes sur les essais cliniques randomisés contrôlés pour vérifier les critères CONSORT au niveau des résumés. Nous fournissons quelques exemples d'explication correcte dans la figure 5.

Question : For the primary outcome, is there a result for each group and the estimated effect size and its precision (e.g., 95% CI) ? (if one of them is missing, answer no) **Explanation :** The results section does not provide a clear report of the estimated effect size and its precision for the primary outcome. It only mentions that 'No significant group differences were observed on primary or secondary emotional outcomes' but does not provide the specific effect sizes and their precision. **Answer :** No
Question : Are the primary outcomes clearly described for this trial in methods ? **Explanation :** The primary outcome is clearly described in the methods section as 'the incidence of either laboratory-confirmed Covid-19 or illness compatible with Covid-19 within 14 days'. **Answer :** Yes
Question : Do authors describe if the program deliverer (caregiver) were blinded ? (answer yes only if caregivers are blinded, do not care about participants or outcome assessors) **Explanation :** The text does not mention whether the program deliverers (school clinicians) were blinded to the group assignment. **Answer :** No

texte surligné = parties générées par le modèle.

FIGURE 5. *Exemples corrects générés par le modèle Mixtral-8x22B avec la méthode 5-shot-cot. Seules les questions et les explications sont pris en compte afin de limiter la taille des exemples. Cependant ces exemples d'explications ont bien été validés manuellement avec leur contexte complet. Leur justification a ici été annotée comme correcte.*

5.5. *Analyse des erreurs*

Nous analysons également les parties générées par le modèle lorsqu'il commet une erreur. Nous échantillonnons 50 faux positifs et 50 faux négatifs pour le modèle

Mixtral-8x22B avec la méthode *5-shot-cot*. Nous annotons manuellement ces erreurs avec trois critères qualitatifs :

– cohérence de l'explication avec la question : ce critère vérifie à quel point l'explication fournie par le modèle montre une compréhension de la question. Nous définissons trois niveaux pour l'annotation : cohérent, partiellement cohérent et incohérent. Pour le niveau partiel, cela signifie que le modèle fournit une explication logique par rapport à la question, mais soit il ne prend pas en compte une partie de la question, soit il ajoute des éléments non pertinents ;

– cohérence de l'explication avec le contexte : nous avons les mêmes niveaux d'annotation que pour le critère précédent mais cette fois-ci nous observons lorsque le modèle cite une partie du contexte pour justifier sa réponse, ou lorsqu'il dit qu'un critère est manquant nous vérifions s'il est bien manquant dans le résumé. Si le modèle ajoute des parties n'existant pas (hallucination du modèle), nous considérons directement cela comme une incohérence (même si la partie non existante n'est pas importante pour la réponse) ;

– cohérence de la réponse finale avec l'explication : ici nous regardons si la tournure de l'explication va dans le sens de la réponse finale (critère vérifié ou non, soit génération du *token* « *Yes* » ou « *No* »). L'annotation « partiellement cohérent » est ici spécifiquement pour le cas où l'explication du modèle exprime les deux points de vue (il trouve la réponse mais la nuance ensuite et se trompe par exemple). L'annotation incohérente fait référence à une explication correcte mais à une génération de réponse finale incorrecte.

Nous fournissons quelques exemples d'erreurs annotées dans la figure 6 puis montrons les résultats de l'analyse d'erreurs dans la figure 7. Nous observons que la majorité des erreurs du modèle proviennent d'une mauvaise compréhension de la question. En effet, ayant fait le choix d'utiliser les définitions fournies par les auteurs, nous remarquons que certaines sont incomplètes et peuvent être imprécises, même pour un évaluateur humain. Une définition plus précise pourrait selon nous améliorer les performances sur les critères en question. Pour les erreurs sur le contexte, il existe encore parfois des hallucinations, où le modèle cite des parties n'étant pas dans le résumé. Il serait également possible de limiter ces erreurs en contrôlant la génération des parties citées par le modèle.

5.6. *Effet du filtrage des phrases selon le critère*

Nous observons l'influence du filtrage des phrases du résumé uniquement pour le modèle *Llama-2-70b-chat*, afin de limiter les quantités de calcul à effectuer. Les différences obtenues sont minimes, le filtrage étant globalement moins efficace sur les différentes stratégies de requêtage. Nous pouvons alors en déduire que le modèle parvient à retrouver l'information dans un résumé entier malgré le bruit introduit par certaines phrases non pertinentes pour le critère en cours de génération, ou que l'erreur

	Q	C	R
Question : Is there only a brief description of blinding (eg, single-blind, double-blind, triple-blind) ? **Explanation :** The authors report : 'Design : Parallel pilot randomized open label, double-masked clinical trial.' This means that the trial is not blinded at all, so there is a clear description of blinding (or lack thereof). **Answer :** No	✗	✓	≈
Context : Title : Structured Sensory Therapy (SITCAP-ART) for Traumatized Adjudicated Adolescents in Residential Treatment [...] **Question :** Is the study identified as randomized in the title ? **Explanation :** The title of this abstract ends with 'randomized controlled study'. It mentions the word 'randomized'. Hence, the study is correctly identified as randomized in the title **Answer :** Yes	✓	✗	✓
Question : For the primary outcome(s), is the precision of the estimate (eg, 95%CI) clearly stated ? **Explanation :** The primary outcome is not explicitly stated in the abstract. However, the abstract does provide the 95% CI for the recovery rate ratio (1.87 ; 95% CI, 0.72, 4.89) and the absolute risk reduction for invasive mechanical ventilation (32% ; 95% CI, -0.07, 0.71). **Answer :** Yes	✓	≈	✗

Texte surligné = parties générées par le modèle.
Les colonnes Q, C et R correspondent respectivement à question, contexte et réponse.
Les symboles signifient : ✗= incohérent, ✓= cohérent et ≈ = partiellement cohérent.

FIGURE 6. *Exemples incorrects générés par le modèle Mixtral-8x22B avec la méthode 5-shot-cot, ainsi que les annotations qualitatives réalisées*

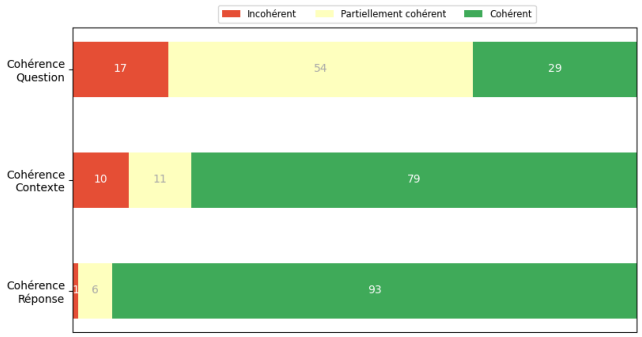


FIGURE 7. *Analyse qualitative des erreurs pour les explications du modèle Mixtral-8x22B*

introduite par le modèle de filtrage vient compenser le bénéfice ajouté (mais nous n'analysons pas ce phénomène ici).

Il est difficile d'introduire cette méthode dans notre évaluation complète car les modèles servant à filtrer les phrases n'ont pas été entraînés exactement sur le même type de données (textes complets alors que nous évaluons des résumés) ni pour les mêmes critères (critères CONSORT-2010 alors que nous utilisons les critères CONSORT-RÉSUMÉ, bien qu'il existe une intersection). Néanmoins cette technique peut être intérressante pour les stratégies *few-shot* lorsque la taille de la requête dépasse la taille maximale de contexte du modèle, car elle permet de réduire le nombre de phrases à inclure dans la requête (sachant que les résumés occupent la majeure partie de la taille de la requête, d'autant plus lorsqu'on utilise plusieurs exemples).

6. Discussion

6.1. *Vers une évaluation automatisée de la qualité des rapports*

Notre travail est le premier à évaluer la performance des larges modèles de langue pour évaluer la qualité des rapports sur les résumés d'essais cliniques. Il est important d'assurer la qualité des résumés qui sont généralement le premier passage que les praticiens cliniques vont lire. Ils peuvent parfois même se limiter à cette lecture.

Un éditorial récent (Nashwan *et al.*, 2023) souligne la possibilité pour les larges modèles de langue d'améliorer le processus d'évaluation de la qualité des essais cliniques. Ils précisent cependant que nous ne devrions pas compter uniquement sur les méthodes d'apprentissage automatique, mais qu'elles pourraient réduire la charge de travail si elles étaient combinées à l'expertise humaine.

Nous fournissons également deux corpus provenant de différents types d'essais cliniques évaluant les critères CONSORT-RÉSUMÉ. Nous obtenons de bonnes performances pour les meilleurs modèles testés avec nos stratégies de requêtage de type *Chain-of-Thought*, et qui permettent également d'avoir une réponse plus transparente du modèle (grâce à l'explication fournie), certains critères spécifiques sont même parfaitement évalués par les larges modèles de langue.

Les larges modèles de langue nous permettent ici, avec le même modèle et sans entraînement supplémentaire, de répondre à toute la gamme de critères définis par les standards CONSORT. Nous mettons en lumière l'importance de la taille du modèle et de la méthode de requêtage employée pour la résolution de cette tâche. L'utilité de l'adaptation de ces modèles au domaine biomédical n'a cependant pas encore pu être établie de manière certaine par nos expériences.

6.2. *Limitations*

Taille et variété des corpus. Les corpus présentés sont petits et ne concernent que deux domaines d'essais cliniques différents. Disposer de corpus plus importants et de spécialités médicales plus variées permettrait d'améliorer nos méthodes, d'affiner les modèles et d'avoir une évaluation plus complète. De plus, il s'agit ici des critères

uniquement pour les résumés. Pour une évaluation plus complète et détaillée d'un article, il peut être intéressant de l'effectuer sur le texte complet de l'article.

Utilisation des larges modèles génératifs. Bien qu'en général, les modèles génératifs se montrent très efficaces pour les tâches de questions-réponses, ils ne sont pas forcément pertinents pour tous les critères évalués ici. En effet, certains critères assez simples peuvent très probablement être détectés avec des modèles de plus petite taille, et donc moins coûteux.

Requêtage. Les larges modèles de langue génératifs sont également connus pour halluciner et dans notre cas ils peuvent fournir une explication fausse ou même citer des passages du résumé n'existant pas pour les méthodes *Chain-of-Thought*. Notons qu'il serait possible de contrôler ce dernier point automatiquement, du moins pour ce qui concerne les verbatim, en contraignant le modèle à ne pas générer des séquences de *tokens* déjà présentes dans son contexte.

7. Conclusion

Dans cet article, nous avons rapporté nos premières expériences utilisant les larges modèles de langue pour évaluer automatiquement la qualité des résumés d'articles de recherche faisant suite à un essai clinique. Nous avons ciblé les critères de qualité CONSORT-RÉSUMÉ, car ils sont largement acceptés par la communauté concernée.

Pour évaluer nos modèles, nous avons extrait deux corpus d'évaluation produits par des évaluateurs humains experts, comportant au total 139 résumés, portant sur deux domaines cliniques différents.

Nous utilisons une des collections de larges modèles de langue disponible publiquement et comparons les effets de la taille de ces modèles et de l'utilisation de différentes stratégies de requêtage. Nous obtenons nos meilleures performances en utilisant le modèle le plus large (70 milliards de paramètres), ajusté par instruction, et avec la stratégie de chaîne de pensée (COT). Cependant, l'intérêt de modèles ajustés pour le domaine biomédical reste encore à prouver pour notre tâche. Il en va de même pour l'utilisation d'un filtrage des phrases du résumé en amont de la génération.

Nos modèles atteignent près de 85 % en exactitude pour les deux corpus, ceci montre qu'il existe de nouvelles perspectives de recherche prometteuses pour l'évaluation de la qualité de rapport des articles d'essais cliniques par ce type de modèles. D'autant plus qu'ils ne nécessitent ici aucun entraînement supplémentaire et sont capables de justifier leurs réponses.

Il reste néanmoins encore de la marge pour améliorer ces méthodes, notamment parce que certains critères restent plus difficiles à évaluer que d'autres et que ces modèles sont toujours sujets aux hallucinations. Les pistes pour le travail futur sont nombreuses : l'ajout de plus de données pour entraîner ces modèles, l'ajout de plus de critères à évaluer (notamment les critères utilisant les textes complets d'articles), ou

encore l'ajout de contraintes pour limiter les hallucinations (par exemple contrôler la génération lorsque le modèle cite un passage du résumé).

Remerciements

Ces travaux ont été financés par le Centre National de la Recherche Scientifique (CNRS) sur un projet MITI 80 PRIME. Nos expériences ont été réalisées à l'aide de l'infrastructure de calcul et de stockage de GENCI à l'IDRIS dans le projet AD011014707, sur la partition A100 du supercalculateur Jean-Zay.

8. Bibliographie

Abdin M., Jacobs S. A., Awan A. A., Aneja J., Awadallah A., Awadalla H., Bach N., Bahree A., Bakhtiari A., Bao J., Behl H., Benhaim A., Bilenko M., Bjorck J., Bubeck S., Cai Q., Cai M., Mendes C. C. T., Chen W., Chaudhary V., Chen D., Chen D., Chen Y.-C., Chen Y.-L., Chopra P., Dai X., Del Giorno A., de Rosa G., Dixon M., Eldan R., Fragoso V., Iter D., Gao M., Gao M., Gao J., Garg A., Goswami A., Gunasekar S., Haider E., Hao J., Hewett R. J., Huynh J., Javaheripi M., Jin X., Kauffmann P., Karampatziakis N., Kim D., Khademi M., Kurilenko L., Lee J. R., Lee Y. T., Li Y., Li Y., Liang C., Liden L., Liu C., Liu M., Liu W., Lin E., Lin Z., Luo C., Madan P., Mazzola M., Mitra A., Modi H., Nguyen A., Norick B., Patra B., Perez-Becker D., Portet T., Pryzant R., Qin H., Radmilac M., Rosset C., Roy S., Ruwase O., Saarikivi O., Saied A., Salim A., Santacroce M., Shah S., Shang N., Sharma H., Shukla S., Song X., Tanaka M., Tupini A., Wang X., Wang L., Wang C., Wang Y., Ward R., Wang G., Witte P., Wu H., Wyatt M., Xiao B., Xu C., Xu J., Xu W., Yadav S., Yang F., Yang J., Yang Z., Yang Y., Yu D., Yuan L., Zhang C., Zhang C., Zhang J., Zhang L. L., Zhang Y., Zhang Y., Zhang Y., Zhou X., « Phi-3 Technical Report : A Highly Capable Language Model Locally on Your Phone », May, 2024. ArXiv.

AI@Meta, « Llama 3 Model Card », 2024.

Altman D. G., Schulz K. F., Moher D., Egger M., Davidoff F., Elbourne D., Gøtzsche P. C., Lang T., « The Revised CONSORT Statement for Reporting Randomized Trials : Explanation and Elaboration », *Annals of Internal Medicine*, vol. 134, n° 8, p. 663-694, April, 2001.

Ankit Pal M. S., « OpenBioLLMs : Advancing Open-Source Large Language Models for Healthcare and Life Sciences », 2024. Hugging Face.

Begg C., Cho M., Eastwood S., Horton R., Moher D., Olkin I., Pitkin R., Rennie D., Schulz K. F., Simel D., Stroup D. F., « Improving the Quality of Reporting of Randomized Controlled Trials : The CONSORT Statement », *JAMA*, vol. 276, n° 8, p. 637-639, August, 1996.

Benesty J., Chen J., Huang Y., Cohen I., « Pearson Correlation Coefficient », *in* I. Cohen, Y. Huang, J. Chen, J. Benesty (eds), *Noise Reduction in Speech Processing*, Springer, Berlin, Heidelberg, p. 1-4, 2009.

Bero L., Lawrence R., Leslie L., Chiu K., McDonald S., Page M. J., Grundy Q., Parker L., Boughton S., Kirkham J. J., Featherstone R., « Cross-Sectional Study of Preprints and Final Journal Publications from COVID-19 Studies : Discrepancies in Results Reporting and Spin in Interpretation », *BMJ open*, vol. 11, n° 7, p. e051821, July, 2021.

Brown T., Mann B., Ryder N., Subbiah M., Kaplan J. D., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., Agarwal S., Herbert-Voss A., Krueger G., Henighan T., Child R., Ramesh A., Ziegler D., Wu J., Winter C., Hesse C., Chen M., Sigler E., Litwin M., Gray S., Chess B., Clark J., Berner C., McCandlish S., Radford A., Sutskever I., Amodei D., « Language Models Are Few-Shot Learners », *Advances in Neural Information Processing Systems*, vol. 33, Curran Associates, Inc., p. 1877-1901, 2020.

Chen Z., Hernández-Cano A., Romanou A., Bonnet A., Matoba K., Salvi F., Pagliardini M., Fan S., Köpf A., Mohtashami A., Sallinen A., Sakhaeirad A., Swamy V., Krawczuk I., Bayazit D., Marmet A., Montariol S., Hartley M.-A., Jaggi M., Bosselut A., « MEDITRON-70B : Scaling Medical Pretraining for Large Language Models », 2023. ArXiv.

CohereForAI, « Command R Plus », , https ://huggingface.co/CohereForAI/c4ai-command-r-plus, 2024. Accessed on June 2, 2023.

Devlin J., Chang M.-W., Lee K., Toutanova K., « BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding », *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, p. 4171-4186, June, 2019.

Hopewell S., Clarke M., Moher D., Wager E., Middleton P., Altman D. G., Schulz K. F., Group a. t. C., « CONSORT for Reporting Randomized Controlled Trials in Journal and Conference Abstracts : Explanation and Elaboration », *PLOS Medicine*, vol. 5, n° 1, p. e20, 2008.

Jardim P. S. J., Rose C. J., Ames H. M., Echavez J. F. M., Van de Velde S., Muller A. E., « Automating Risk of Bias Assessment in Systematic Reviews : A Real-Time Mixed Methods Comparison of Human Researchers to a Machine Learning System », *BMC Medical Research Methodology*, vol. 22, n° 1, p. 167, June, 2022.

Jiang A. Q., Sablayrolles A., Mensch A., Bamford C., Chaplot D. S., de las Casas D., Bressand F., Lengyel G., Lample G., Saulnier L., Lavaud L. R., Lachaux M.-A., Stock P., Scao T. L., Lavril T., Wang T., Lacroix T., Sayed W. E., « Mistral 7B », October, 2023. ArXiv.

Jiang A. Q., Sablayrolles A., Roux A., Mensch A., Savary B., Bamford C., Chaplot D. S., de las Casas D., Hanna E. B., Bressand F., Lengyel G., Bour G., Lample G., Lavaud L. R., Saulnier L., Lachaux M.-A., Stock P., Subramanian S., Yang S., Antoniak S., Scao T. L., Gervet T., Lavril T., Wang T., Lacroix T., Sayed W. E., « Mixtral of Experts », , arXiv, January, 2024.

Jin D., Pan E., Oufattole N., Weng W.-H., Fang H., Szolovits P., « What Disease Does This Patient Have ? A Large-scale Open Domain Question Answering Dataset from Medical Exams », *arXiv :2009.13081*, 2020.

Jin D., Szolovits P., « PICO Element Detection in Medical Text via Long Short-Term Memory Neural Networks », *Proceedings of the BioNLP 2018 Workshop*, Association for Computational Linguistics, Melbourne, Australia, p. 67-75, 2018.

Kilicoglu H., Rosemblat G., Hoang L., Wadhwa S., Peng Z., Malički M., Schneider J., ter Riet G., « Toward Assessing Clinical Trial Publications for Reporting Transparency », *Journal of Biomedical Informatics*, vol. 116, p. 103717, April, 2021.

Kojima T., Gu S. S., Reid M., Matsuo Y., Iwasawa Y., « Large Language Models Are Zero-Shot Reasoners », *in* S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (eds), *Advances in Neural Information Processing Systems*, vol. 35, Curran Associates, Inc., p. 22199-22213, 2022.

Koroleva A., Assisted Authoring for Avoiding Inadequate Claims in Scientific Reporting, PhD thesis, Universiteit von Amsterdam, 2020.

Kwon W., Li Z., Zhuang S., Sheng Y., Zheng L., Yu C. H., Gonzalez J. E., Zhang H., Stoica I., « Efficient Memory Management for Large Language Model Serving with PagedAttention », *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Labrak Y., Bazoge A., Morin E., Gourraud P.-A., Rouvier M., Dufour R., « BioMistral : A Collection of Open-Source Pretrained Large Language Models for Medical Domains », , arXiv, February, 2024.

Lee J., Yoon W., Kim S., Kim D., Kim S., So C. H., Kang J., « BioBERT : a pre-trained biomedical language representation model for biomedical text mining », *Bioinformatics*, vol. 36, n° 4, p. 1234-1240, 2020.

Luo R., Sun L., Xia Y., Qin T., Zhang S., Poon H., Liu T.-Y., « BioGPT : generative pre-trained transformer for biomedical text generation and mining », *Briefings in Bioinformatics*, vol. 23, n° 6, p. bbac409, 09, 2022.

Marshall I. J., Kuiper J., Wallace B. C., « Automating Risk of Bias Assessment for Clinical Trials », *IEEE journal of biomedical and health informatics*, vol. 19, n° 4, p. 1406-1412, July, 2015.

Marshall I. J., Kuiper J., Wallace B. C., « RobotReviewer : Evaluation of a System for Automatically Assessing Bias in Clinical Trials », *Journal of the American Medical Informatics Association : JAMIA*, vol. 23, n° 1, p. 193-201, January, 2016.

Moher D., Hopewell S., Schulz K. F., Montori V., Gøtzsche P. C., Devereaux P. J., Elbourne D., Egger M., Altman D. G., « CONSORT 2010 Explanation and Elaboration : Updated Guidelines for Reporting Parallel Group Randomised Trials », *BMJ*, vol. 340, p. c869, 2010.

Muennighoff N., Wang T., Sutawika L., Roberts A., Biderman S., Scao T. L., Bari M. S., Shen S., Yong Z.-X., Schoelkopf H., Tang X., Radev D., Aji A. F., Almubarak K., Albanie S., Alyafeai Z., Webson A., Raff E., Raffel C., « Crosslingual Generalization through Multitask Finetuning », May, 2023. arXiv.

Mutinda F. W., Liew K., Yada S., Wakamiya S., Aramaki E., « Automatic Data Extraction to Support Meta-Analysis Statistical Analysis : A Case Study on Breast Cancer », *BMC Medical Informatics and Decision Making*, vol. 22, p. 158, June, 2022.

Nashwan A. J., Jaradat J. H., Nashwan A. J., Jaradat J. H., « Streamlining Systematic Reviews : Harnessing Large Language Models for Quality Assessment and Risk-of-Bias Evaluation », *Cureus*, August, 2023.

Neumann M., King D., Beltagy I., Ammar W., « ScispaCy : Fast and Robust Models for Biomedical Natural Language Processing », *Proceedings of the 18th BioNLP Workshop and Shared Task*, Association for Computational Linguistics, Florence, Italy, p. 319-327, August, 2019.

Niforatos J. D., Weaver M., Johansen M. E., « Assessment of Publication Trends of Systematic Reviews and Randomized Clinical Trials, 1995 to 2017 », *JAMA Internal Medicine*, vol. 179, n° 11, p. 1593-1594, November, 2019.

Ouyang L., Wu J., Jiang X., Almeida D., Wainwright C. L., Mishkin P., Zhang C., Agarwal S., Slama K., Ray A., Schulman J., Hilton J., Kelton F., Miller L., Simens M., Askell A., Welinder P., Christiano P., Leike J., Lowe R., « Training Language Models to Follow Instructions with Human Feedback », March, 2022. arXiv.

Park J. J. H., Mogg R., Smith G. E., Nakimuli-Mpungu E., Jehan F., Rayner C. R., Condo J., Decloedt E. H., Nachega J. B., Reis G., Mills E. J., « How COVID-19 Has Fundamentally Changed Clinical Research in Global Health », *The Lancet Global Health*, vol. 9, nº 5, p. e711-e720, May, 2021.

Singhal K., Azizi S., Tu T., Mahdavi S. S., Wei J., Chung H. W., Scales N., Tanwani A., Cole-Lewis H., Pfohl S., Payne P., Seneviratne M., Gamble P., Kelly C., Babiker A., Schärli N., Chowdhery A., Mansfield P., Demner-Fushman D., Agüera y Arcas B., Webster D., Corrado G. S., Matias Y., Chou K., Gottweis J., Tomasev N., Liu Y., Rajkomar A., Barral J., Semturs C., Karthikesalingam A., Natarajan V., « Large Language Models Encode Clinical Knowledge », *Nature*, vol. 620, nº 7972, p. 172-180, 2023a.

Singhal K., Tu T., Gottweis J., Sayres R., Wulczyn E., Hou L., Clark K., Pfohl S., Cole-Lewis H., Neal D., Schaekermann M., Wang A., Amin M., Lachgar S., Mansfield P., Prakash S., Green B., Dominowska E., y Arcas B. A., Tomasev N., Liu Y., Wong R., Semturs C., Mahdavi S. S., Barral J., Webster D., Corrado G. S., Matias Y., Azizi S., Karthikesalingam A., Natarajan V., « Towards Expert-Level Medical Question Answering with Large Language Models », *arXiv :2305.09617*, May, 2023b.

Sterne J. A. C., Savović J., Page M. J., Elbers R. G., Blencowe N. S., Boutron I., Cates C. J., Cheng H.-Y., Corbett M. S., Eldridge S. M., Emberson J. R., Hernán M. A., Hopewell S., Hróbjartsson A., Junqueira D. R., Jüni P., Kirkham J. J., Lasserson T., Li T., McAleenan A., Reeves B. C., Shepperd S., Shrier I., Stewart L. A., Tilling K., White I. R., Whiting P. F., Higgins J. P. T., « RoB 2 : A Revised Tool for Assessing Risk of Bias in Randomised Trials », , vol. 366, p. l4898, n.d.

Team G., Mesnard T., Hardin C., Dadashi R., Bhupatiraju S., Pathak S., Sifre L., Rivière M., Kale M. S., Love J., Tafti P., Hussenot L., Sessa P. G., Chowdhery A., Roberts A., Barua A., Botev A., Castro-Ros A., Slone A., Héliou A., Tacchetti A., Bulanova A., Paterson A., Tsai B., Shahriari B., Lan C. L., Choquette-Choo C. A., Crepy C., Cer D., Ippolito D., Reid D., Buchatskaya E., Ni E., Noland E., Yan G., Tucker G., Muraru G.-C., Rozhdestvenskiy G., Michalewski H., Tenney I., Grishchenko I., Austin J., Keeling J., Labanowski J., Lespiau J.-B., Stanway J., Brennan J., Chen J., Ferret J., Chiu J., Mao-Jones J., Lee K., Yu K., Millican K., Sjoesund L. L., Lee L., Dixon L., Reid M., Mikuła M., Wirth M., Sharman M., Chinaev N., Thain N., Bachem O., Chang O., Wahltinez O., Bailey P., Michel P., Yotov P., Chaabouni R., Comanescu R., Jana R., Anil R., McIlroy R., Liu R., Mullins R., Smith S. L., Borgeaud S., Girgin S., Douglas S., Pandya S., Shakeri S., De S., Klimenko T., Hennigan T., Feinberg V., Stokowiec W., Chen Y.-h., Ahmed Z., Gong Z., Warkentin T., Peran L., Giang M., Farabet C., Vinyals O., Dean J., Kavukcuoglu K., Hassabis D., Ghahramani Z., Eck D., Barral J., Pereira F., Collins E., Joulin A., Fiedel N., Senter E., Andreev A., Kenealy K., « Gemma : Open Models Based on Gemini Research and Technology », , arXiv, April, 2024.

Touvron H., Lavril T., Izacard G., Martinet X., Lachaux M.-A., Lacroix T., Rozière B., Goyal N., Hambro E., Azhar F., Rodriguez A., Joulin A., Grave E., Lample G., « LLaMA : Open and Efficient Foundation Language Models », , arXiv, February, 2023a.

Touvron H., Martin L., Stone K., Albert P., Almahairi A., Babaei Y., Bashlykov N., Batra S., Bhargava P., Bhosale S., Bikel D., Blecher L., Ferrer C. C., Chen M., Cucurull G., Esiobu D., Fernandes J., Fu J., Fu W., Fuller B., Gao C., Goswami V., Goyal N., Hartshorn A., Hosseini S., Hou R., Inan H., Kardas M., Kerkez V., Khabsa M., Kloumann I., Korenev A., Koura P. S., Lachaux M.-A., Lavril T., Lee J., Liskovich D., Lu Y., Mao Y., Martinet X., Mihaylov T., Mishra P., Molybog I., Nie Y., Poulton A., Reizenstein J., Rungta R., Saladi

K., Schelten A., Silva R., Smith E. M., Subramanian R., Tan X. E., Tang B., Taylor R., Williams A., Kuan J. X., Xu P., Yan Z., Zarov I., Zhang Y., Fan A., Kambadur M., Narang S., Rodriguez A., Stojnic R., Edunov S., Scialom T., « Llama 2 : Open Foundation and Fine-Tuned Chat Models », , arXiv, July, 2023b.

Turner L., Shamseer L., Altman D. G., Schulz K. F., Moher D., « Does Use of the CONSORT Statement Impact the Completeness of Reporting of Randomised Controlled Trials Published in Medical Journals ? A Cochrane Review », *Systematic Reviews*, vol. 1, p. 60, 2012.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I., « Attention Is All You Need », *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.

Wang D., Chen L., Wang L., Hua F., Li J., Li Y., Zhang Y., Fan H., Li W., Clarke M., « Abstracts for Reports of Randomized Trials of COVID-19 Interventions Had Low Quality and High Spin », *Journal of Clinical Epidemiology*, vol. 139, p. 107-120, 2021.

Wang F., Schilsky R. L., Page D., Califf R. M., Cheung K., Wang X., Pang H., « Development and Validation of a Natural Language Processing Tool to Generate the CONSORT Reporting Checklist for Randomized Clinical Trials », *JAMA Network Open*, vol. 3, n° 10, p. e2014661, October, 2020.

Wang Q., Liao J., Lapata M., Macleod M., « PICO Entity Extraction for Preclinical Animal Literature », *Systematic Reviews*, vol. 11, n° 1, p. 209, September, 2022.

Warrier K., Jayanthi C. R., « Completeness of Reporting and Outcome Switching in Trials Published in Indian Journals from 2017 to 2019 : A Cross-Sectional Study », *Perspectives in Clinical Research*, vol. 13, n° 2, p. 77-81, 2022.

Wei J., Wang X., Schuurmans D., Bosma M., Ichter B., Xia F., Chi E., Le Q. V., Zhou D., « Chain-of-Thought Prompting Elicits Reasoning in Large Language Models », *Advances in Neural Information Processing Systems*, vol. 35, p. 24824-24837, 2022.

Wiehn J., Nonte J., Prugger C., « Reporting Quality for Abstracts of Randomised Trials on Child and Adolescent Depression Prevention : A Meta-Epidemiological Study on Adherence to CONSORT for Abstracts », *BMJ Open*, vol. 12, n° 8, p. e061873, 2022.

Yang C., Wang X., Lu Y., Liu H., Le Q. V., Zhou D., Chen X., « Large Language Models as Optimizers », *arXiv :2309.03409*, September, 2023.