QUARTZ : Approche abstractive non supervisée par question-réponse pour le résumé de dialogue orienté tâche

Mohamed Imed Eddine Ghebriout[1] Gaël Guibon[2, 3] Ivan Lerner[4, 5, 6]
Emmanuel Vincent[1]
(1) Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
(2) Université de Lorraine, CNRS, LORIA, F-54000 Nancy, France
(3) Université Sorbonne Paris Nord, CNRS, Laboratoire d'Informatique de Paris Nord, LIPN, F-93430 Villetaneuse, France
(4) Inserm, Centre de Recherche des Cordeliers, Université Paris Cité, Sorbonne Université, F-75006 Paris, France
(5) HeKA, Inria Paris, F-75012 Paris, France
(6) Department of Medical Informatics, Assistance Publique Hôpitaux de Paris (AP-HP), Georges Pompidou European Hospital, Paris, France
{imed-eddine.ghebriout,gael.guibon}@loria.fr
{ivan.lerner,emmanuel.vincent}@inria.fr

RÉSUMÉ

Le résumé de dialogues condense les conversations en un texte concis, réduisant la complexité des applications riches en interactions. Les approches existantes reposent souvent sur l'entraînement de modèles de langue à imiter des résumés humains. Cependant, cette approche est coûteuse et les résumés obtenus manquent souvent de pertinence, entraînant des performances sous-optimales, notamment en médecine. Dans cet article, nous introduisons *QUARTZ*, une méthode non supervisée pour le résumé de dialogues orienté tâche. *QUARTZ* génère plusieurs résumés et paires de questions-réponses à l'aide de grands modèles de langue (LLMs). Les résumés sont évalués en demandant aux LLMs de répondre à ces questions avant *(i)* de sélectionner les meilleures réponses et *(ii)* d'identifier le résumé le plus informatif. Enfin, nous affinons le meilleur LLM sur les résumés générés sélectionnés. Validé sur plusieurs ensembles de données, *QUARTZ* atteint des performances compétitives en zéro-shot, rivalisant avec les approches supervisées de pointe.

ABSTRACT

QUARTZ : QA-based Unsupervised Abstractive Refinement for Task-oriented Dialogue Summarization

Dialogue summarization condenses conversations into concise text, reducing dialogue complexity in dialogue-heavy applications. Existing approaches heavily rely on costly human-written data, and the resulting summaries often lack task-specific focus, leading to suboptimal performance for downstream tasks, such as medical ones. In this paper, we introduce *QUARTZ*, a framework for task-oriented unsupervised dialogue summarization. *QUARTZ* starts by generating multiple summaries and task-specific question-answer pairs using large language models (LLMs). Summaries are evaluated by having the LLMs respond to task-related questions before *(i)* selecting the best candidate responses and *(ii)* identifying the most informative summary. Finally, we finetune the best LLM on the selected summaries. When validated on multiple datasets, *QUARTZ* achieves competitive zero-shot performance, rivaling fully-supervised State-of-the-Art (SoTA) approaches.

MOTS-CLÉS : Résumé de dialogues, Apprentissage non supervisé, Adaptation aux tâches spécifiques, Grand Modèles de langue (LLMs).

KEYWORDS: Dialogue summarization, Unsupervised learning, Task-oriented adaptation, Large Language Models (LLMs).

ARTICLE : **Accepté à** TALN.

1 Introduction

Le résumé automatique de texte (RA) vise à condenser l'information d'un texte source en un résumé bref, soit par extraction de phrases existantes (*extractif*), soit par génération de nouvelles phrases capturant l'essentiel (*abstractif*), imitant ainsi le raisonnement humain (Lin & Ng, 2019). Depuis les années 1950 (Nenkova *et al.*, 2011), les approches ont évolué, passant des méthodes basées sur les graphes (Mihalcea & Tarau, 2004) ou sur la fréquence des termes (Alsaedi *et al.*, 2016) aux grands modèles de langue (LLMs), dont la capacité à représenter finement le contexte textuel a permis des avancées majeures, notamment dans les applications cliniques (Van Veen *et al.*, 2024). Le résumé de dialogues, sous-discipline du RA, se concentre sur l'extraction d'informations clés à partir d'une conversation. Il s'impose dans les applications concrètes comme le service client (Feigenblat *et al.*, 2021), les réunions d'affaires (Rennard *et al.*, 2023) ou le domaine médical (Abacha *et al.*, 2023), où les échanges sont souvent moins structurées que des textes formels. Dans ces contextes, résumer les conversations permet de réduire leur complexité et optimiser les tâches en aval comme la prise de décision ou l'automatisation des processus. Toutefois, cette tâche pose des défis uniques en raison de la structure moins rigide des conversations, incluant souvent des répétitions, verbosités et des disfluences —rendant les méthodes classiques de RA, conçues pour des textes plus structurés, moins efficaces (Feng *et al.*, 2022). De plus, contrairement aux textes monologués, les dialogues impliquent plusieurs locuteurs aux styles, rôles et objectifs variés (Zechner, 2002). Récemment, le résumé de dialogues par LLMs a été abordé comme une tâche de type séquence-à-séquence (Van Veen *et al.*, 2024; Tian *et al.*, 2024). Bien que l'affinage de ces modèles pour le résumé orienté tâche donne de bons résultats —parfois supérieurs à ceux des humains (Van Veen *et al.*, 2024), cette approche reste coûteuse car dépendante de données annotées manuellement. Par ailleurs, malgré leur polyvalence acquise via le pré-entraînement massif, les LLMs peinent parfois à maintenir la cohérence thématique et factuelle, générant des digressions ou des contradictions (Tonmoy *et al.*, 2024). Dans cet article, nous introduisons *QUARTZ*, une nouvelle méthode non supervisée pour le résumé abstractif de dialogues orienté tâche. Son approche orientée tâche guide les LLMs vers des résumés cohérents et factuels, tandis que son caractère non supervisé signifie qu'elle ne repose ni sur des résumés humains de référence ni sur des connaissances spécifiques à la tâche, hormis celles nécessaires pour concevoir des prompts adaptés. *QUARTZ* introduit les contributions suivantes :

1. À notre connaissance, *QUARTZ* est la première méthode visant à améliorer la capacité des LLMs à résumer des dialogues de manière totalement non supervisée.
2. Nous générons plusieurs résumés par dialogue et proposons une approche de sélection en deux niveaux : *(i)* Identifier automatiquement les meilleures réponses pour chaque résumé et question, puis *(ii)* Sélectionner le résumé le plus informatif en se basant sur ces réponses.
3. Applicabilité à des cas concrets, notamment l'assistance et le résumé de conversations cliniques, avec une efficacité démontrée par l'affinage du modèle sur les résumés sélectionnés.

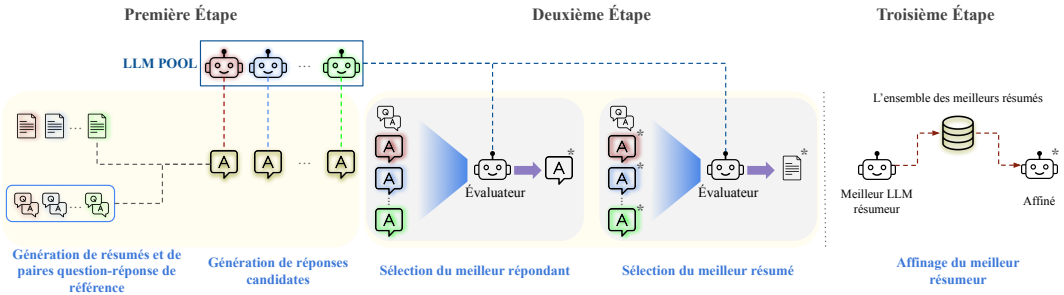

FIGURE 1 – Un aperçu de *QUARTZ* pour le résumé non supervisé de dialogues orienté tâche.

2 Travaux connexes

Résumé de dialogues. Le résumé de dialogues présente des défis liés aux changements sémantiques fréquents, aux transitions de sujet, aux redondances et aux interactions multi-parties. Pour y répondre, Zhao *et al.* (2021) ont exploité les états du dialogue et Tian *et al.* (2024) ont utilisé un modèle LLM Mixture-of-Experts (MoE) (Jacobs *et al.*, 1991). Malgré cela, la cohérence du résumé, la couverture de l'information et la compréhension globale du dialogue (Zhao *et al.*, 2021) restent souvent insuffisants. Le résumé de dialogues orienté tâche offre une solution à ces limitations. Wang *et al.* (2023) ont utilisé des LLMs pour générer des requêtes et leurs résumés à partir du résumé de référence pour produire des résumés guidés par instruction. La plupart des approches orientées tâche nécessitant une quantité massive de données annotées (Zou *et al.*, 2021), les situations à faibles ressources ont été abordées par augmentation de données (AD) (Liu *et al.*, 2022; Ouyang *et al.*, 2023). Cependant, le résumé de dialogues non supervisé, que nous abordons dans cet article, reste largement inexploré.

LLMs pour l'augmentation de données. L'AD (Feng *et al.*, 2021) est utilisée notamment pour la traduction automatique de langues peu dotées (Xia *et al.*, 2019) et la génération de conversations ancrées sur des résumés (Gunasekara *et al.*, 2021). Concernant la réponse aux questions (QA) (Guo *et al.*, 2023), Yang *et al.* (2019) ont proposé une méthode d'AD pour l'affinage de BERT et Riabi *et al.* (2021) ont enrichi la QA multilingue par la génération de questions. L'essor des LLMs a révolutionné l'AD (Tang *et al.*, 2023), avec GPT-3 générant des résumés de dialogues médicaux pour entraîner des modèles sur des données humaines et synthétiques (Chintagunta *et al.*, 2021) et GPT-4 produisant un ensemble de données d'instructions pour affiner le modèle Llama (Peng *et al.*, 2023).

Évaluation par QA utilisant les LLMs. L'évaluation par QA consiste à comparer les réponses à une question à un ensemble de réponses de référence (Wang *et al.*, 2024). Appliquée à un résumé, elle garantit que les informations clés du texte original restent accessibles dans le résumé. Historiquement, l'évaluation QA reposait sur des métriques de correspondance lexicale, comme l'Exact-Match (Izacard & Grave, 2021) ou la correspondance N-gram (Chen *et al.*, 2017), ou des métriques de similarité comme BERT Matching (Bulian *et al.*, 2022). Les LLMs conditionnés par un *prompt* système, une question, une réponse de référence et une réponse candidate (Wang *et al.*, 2024; Kamalloo *et al.*, 2024) sont récemment apparus comme une solution de précision comparable aux évaluateurs humains (Törnberg, 2023; Bavaresco *et al.*, 2024), qui peut être encore amélioré par des techniques d'ingénierie de prompt (Brown *et al.*, 2020; Wei *et al.*, 2022). Pour limiter le risque accru d'évaluation erronée en présence d'un contexte long, une solution efficace consiste à interroger plusieurs fois les LLMs et à agréger les évaluations obtenues (Tang *et al.*, 2024).

3 Notre approche

Notre approche est illustrée dans la Figure 1. Inspirée des récentes avancées en RA, où les résumés générés par les LLMs sont privilégiés par les évaluateurs humains (Pu *et al.*, 2023), *QUARTZ* commence par **(Première Étape)** solliciter un ensemble de LLMs pour générer des résumés orientés tâche et des paires de questions-réponses de référence à partir d'un dialogue d'entrée. L'utilisation d'un ensemble de LLMs garantit une couverture plus large et réduit les biais spécifiques. Pour garantir l'attention aux informations pertinentes pour la tâche, les LLMs sont conditionnés par le contenu du dialogue et des prompts personnalisés (cf. Annexe D). Nous utilisons également les LLMs pour répondre aux questions générées à partir des résumés (plutôt qu'à partir des dialogues originaux). Nous **(Deuxième Étape)** jugeons la qualité des résumés générés en évaluant les réponses candidates via un processus d'évaluation en deux niveaux. Dans le premier niveau, les LLMs sont utilisés comme classifieurs pour sélectionner la meilleure réponse candidate pour chaque question et résumé. Dans le deuxième niveau, le classement des LLMs détermine le meilleur résumé pour chaque dialogue. Enfin, **(Troisième Étape)** le meilleur LLM est affiné sur les résumés sélectionnés. Nous détaillons les différentes étapes de *QUARTZ* et son évaluation ci-dessous.

3.1 Première Étape : Génération

Génération de résumés orientés tâche. Un dialogue orienté tâche est présenté comme une conversation multi-tours entre deux individus ou plus, centrée sur l'atteinte d'un objectif spécifique. Nous sollicitons chaque modèle l_{S} de l'ensemble L pour générer un résumé $S_{i,l_{\mathrm{S}}}$ pour chaque dialogue D_i, en conditionnant ce processus par un prompt soigneusement élaboré T pour extraire les informations spécifiques à la tâche.

Génération de paires de questions-réponses de référence. De manière similaire, nous sollicitons chaque modèle $l_{\mathrm{Q}} \in L$ pour générer un ensemble de paires de questions-réponses (QA) de référence à partir de chaque dialogue D_i, en nous concentrant sur les questions pouvant être répondues en utilisant le résumé du dialogue orienté tâche. Les paires QA de référence générées par tous les LLMs sont fusionnées en un ensemble unique de J_i paires QA de référence $(Q_{i,j}, A_{i,j})$ indexées par j.

Génération des réponses candidates. Après avoir généré les résumés et les questions, nous cherchons la réponse à chaque question dans les résumés générés. En effet, un résumé factuel qui conserve toutes les informations nécessaires doit inclure les bonnes réponses à toutes les questions. Pour ce faire, nous sollicitons chaque modèle $l_{\mathrm{R}} \in L$ pour générer une réponse candidate $\hat{A}_{i,j,l_{\mathrm{S}},l_{\mathrm{R}}}$ pour chaque question $Q_{i,j}$ à partir de chaque résumé $S_{i,l_{\mathrm{S}}}$. En raison du phénomène de biais de soi (*self-bias*) (Xu *et al.*, 2024), selon lequel un LLM peut facilement répondre à une question qu'il a lui-même générée et avoir des difficultés à répondre à des questions générées par d'autres modèles, nous utilisons tous les LLMs de l'ensemble comme répondants, y compris le LLM générateur de question lui-même. Cela réduit le biais et permet de répondre aux questions de manière équitable.

3.2 Deuxième étape : Évaluation en deux niveaux

Évaluation de premier niveau *(Sélection du meilleur répondant).* À cette étape, notre objectif est d'identifier les meilleures réponses candidates pour chaque question et chaque résumé. Pour ce faire, nous utilisons chaque modèle $l_{\mathrm{E}} \in L$ en tant qu'évaluateur. L'évaluateur reçoit la question $Q_{i,j}$, la

réponse de référence $A_{i,j}$, ainsi que l'ensemble des $|L|$ réponses candidates $\{\hat{A}_{i,j,l_S,l_R}\}_{l_R \in L}$ obtenues à partir du résumé S_{i,l_S}, et doit fournir un classement (représenté par une permutation σ) des réponses candidates, selon leur pertinence et leur exactitude. Pour surmonter le biais de préférence propre (*self-preference*) (Panickssery *et al.*, 2024) selon lequel un LLM évaluateur attribue une note plus élevée à ses propres sorties par rapport aux autres, nous utilisons tous les LLMs de l'ensemble comme évaluateurs, y compris le LLM répondant lui-même. Nous procédons ainsi en sollicitant plusieurs fois les LLMs pour le classement, tout en modifiant aléatoirement l'ordre des réponses candidates et en agrégeant ensuite les classements obtenus. Cela garantit qu'aucun LLM n'a un avantage systématique sur les autres, et il a été prouvé théoriquement et empiriquement par Tang *et al.* (2024) que cette approche converge vers le classement réel. Nous sollicitons donc chaque LLM N fois pour obtenir N classements estimés $\hat{\sigma}_{i,j,l_S,l_E,n}$ indexés par n. Le classement optimal σ_{i,j,l_S,l_E} est celui dont la somme des distances Kendall tau (Kendall, 1938) par rapport à tous les classements estimés est minimale (Tang *et al.*, 2024) :

$$\sigma_{i,j,l_S,l_E} := \arg\min_\sigma \sum_{n=1}^{N} d_\kappa(\hat{\sigma}_{i,j,l_S,l_E,n}, \sigma).$$

La distance Kendall tau $d_\kappa(.,.)$ quantifie la dissimilarité entre deux classements, représentant spécifiquement le nombre de paires *discordantes* :

$$d_\kappa(\sigma_1, \sigma_2) := \sum_{k=1}^{|L|} \text{inv}(\sigma_1^{-1} \circ \sigma_2)_k \quad , \quad \text{inv}(\sigma)_k := \#\{k' : \sigma(k') > \sigma(k), k' < k\}.$$

Après avoir obtenu le classement optimal σ_{i,j,l_S,l_E} pour chaque question j, nous calculons le Rang Réciproque Moyen (MRR) (Radev *et al.*, 2002) de chaque LLM répondant l_R sur toutes les questions. Le MRR est calculé comme suit : $\text{MRR}_{i,l_S,l_R,l_E} = \frac{1}{J_i} \sum_{j=1}^{J_i} \frac{1}{\sigma_{i,j,l_S,l_E}^{-1}(l_R)}$.

Ce MRR dépend de l'évaluateur l_E. Pour sélectionner le meilleur répondant de manière indépendante de l'évaluateur, nous calculons le score total $\text{Score}_{i,l_S,l_R} = \sum_{l_E \in L} \alpha_{l_R,l_E} \text{MRR}_{i,l_S,l_R,l_E}$ où le facteur de pondération $\alpha_{l_R,l_E} = \begin{cases} 0.8 & \text{Si } l_R = l_E \\ 1 & \text{Sinon} \end{cases}$ pénalise lorsque le même modèle sert à la fois de répondant et d'évaluateur. Enfin, pour chaque dialogue D_i et résumé S_{i,l_S}, nous sélectionnons le répondant ayant le score le plus élevé et utilisons ces réponses candidates correspondantes $\hat{A}^*_{i,j,l_S}$ comme entrées pour la deuxième étape.

Évaluation de deuxième niveau *(Sélection du meilleur résumé).* Maintenant que nous avons identifié les meilleures réponses candidates pour chaque résumé, nous procédons à l'identification du meilleur résumé pour chaque dialogue. De manière similaire à ce qui précède, nous utilisons chaque modèle $l_E \in L$ comme évaluateur. L'évaluateur reçoit une question $Q_{i,j}$, la réponse correcte $A_{i,j}$, et l'ensemble des $|L|$ meilleures réponses candidates $\{\hat{A}^*_{i,j,l_S}\}_{l_S \in L}$ obtenues à partir de tous les résumés, et il est chargé de classer ces réponses candidates. Nous promptons chaque LLM N fois et calculons le classement optimal σ_{i,j,l_E} comme précédemment. Nous calculons le MRR de chaque résumeur LLM l_S sur toutes les questions comme suit : $\text{MRR}_{i,l_S,l_E} = \frac{1}{J_i} \sum_{j=1}^{J_i} \frac{1}{\sigma_{i,j,l_E}^{-1}(l_S)}$.

Pour sélectionner le meilleur résumé indépendamment de l'évaluateur, nous calculons le score total : $\text{Score}_{i,l_S} = \sum_{l_E \in L} \alpha_{l_S,l_E} \text{MRR}_{i,l_S,l_E}$ avec un facteur de pondération similaire à celui de α_{l_R,l_E}. Enfin, pour chaque dialogue D_i, nous gardons le résumé S_i^* ayant le score le plus élevé.

3.3 Troisième étape : Affinage

Finalement, nous constituons un ensemble d'entraînement qui comprend tous les dialogues et les résumés générés sélectionnés correspondants. Une étape d'affinage est ensuite appliquée au LLM résumeur l_S^* qui a produit la majorité des résumés sélectionnés. Ce processus optimise $\max_\theta \sum_{i=1}^{I} \log P(S_i^* \mid D_i, T, \theta)$ Où le processus de génération est conditionné à la fois sur les paramètres θ du LLM et un prompt T spécifique à la tâche.

4 Paramètres expérimentaux

Jeux de données. Pour évaluer l'efficacité de *QUARTZ*, nous avons réalisé des expériences sur trois jeux de données orientés tâche, dont les statistiques sont résumées dans le Tableau 1.

① DialogSum (Chen *et al.*, 2021) est un jeu de dialogues de vie courante couvrant divers scénarios orientés tâche (négociation commerciale, discussion professionnelle, visite médicale, etc.), avec pour objectif de soutenir des applications aval pour des usages professionnels et personnels.

② MTS-Dialog (Abacha *et al.*, 2023) est un jeu de dialogues médecin-patient accompagnés de notes cliniques réelles couvrant plusieurs spécialités, dont la Médecine Générale, SOAP (Subjectif, Objectif, Évaluation, Plan), Neurologie, Orthopédie, Dermatologie et Allergie/Immunologie. Chaque dialogue est accompagné d'un en-tête (évaluation, allergie, diagnostic, examen, antécédents médicaux et chirurgicaux, etc.) qui définit l'objectif pour la génération de rapports médicaux orientés tâche.

③ SimSAMU (NUN *et al.*, 2025) est un jeu de données français constitué des transcriptions de 3 heures d'enregistrements de dialogues de régulation médicale simulés. Étant données les caractéristiques de ce jeu de données présentées dans le Tableau 1 (par exemple, dialogues longs avec plusieurs tours), les résumés orientés tâche peuvent jouer un rôle crucial en assistant la prise de note lors des appels de régulation médicale, permettant ainsi une prise de décision médicale plus rapide et plus précise.

Détails d'implémentation. Nous utilisons un pool de trois LLMs : Llama-3.1-8B-Instruct (Dubey *et al.*, 2024), Gemma-2-9b-it (Team *et al.*, 2024) et Qwen2-7B-Instruct (Yang *et al.*, 2024). Chaque modèle génère entre 10 et 15 paires de questions-réponses (QA) de référence, qui sont fusionnées en un ensemble unifié de J_i paires QA. Pour le classement des LLMs, nous avons constaté que la cohérence du classement est satisfaisante pour $N = 5$ (cf. Annexe B).

Données		# Dial.	Tokens par dialogue	Tours par dialogue
DialogSum	Entraînement	12 460	131,0	9,5
	Validation	500	129,3	9,4
	Test	1 500	134,5	9,7
MTS-Dialog	Entraînement	1 201	87,99	8,9
	Validation	100	77,46	7,7
	Test	200	87,69	9,1
SimSAMU	-	61	502,47	50,63

TABLEAU 1 – Statistiques des jeux de données. **"# Dial."** fait référence au nombre de dialogues.

Méthode	R-1 ↑	R-2 ↑	R-L ↑	BLEU ↑	BERT-Score ↑
Données ① DialogSum – Orienté tâche (dialogues réels sous scenarios multiples)					
Résumé de dialogue supervisé					
InstructDS (Wang *et al.*, 2023)	47,80	22,20	39,40	-	47,00
MoE (Tian *et al.*, 2024)	**49,82**	**24,80**	**47,34**	**18,41**	**68,48**
Résumé de dialogue non supervisé					
Llama-3.1-8B-Instruct (Dubey *et al.*, 2024)	21,35±0,79	7,36±0,41	16,46±0,60	3,09±0,19	53,70±0,39
DeepSeek-R1-Distill-Llama-8B (Guo *et al.*, 2025)	27,11±0,77	7,53±0,74	20,37±1,09	2,79±0,21	54,27±0,40
DeepSeek-R1-Distill-Qwen-14B (Guo *et al.*, 2025)	29,03±0,77	8,72±0,62	21,99±0,87	3,44±0,32	55,47±0,43
QUARTZ (Meilleur résumeur : Llama 3.1)	**39,73**±1,08	**16,02**±0,87	**32.68**±0,92	**15,73**±0,65	**68,72**±0,44
Données ② MTS-Dialog – Domaine médical (interactions docteur-patient avec notes cliniques)					
Résumé de dialogue supervisé					
BART-GS-DA (Abacha *et al.*, 2023)	**42,52**	**17,50**	**34,90**	-	**40,80**
Résumé de dialogue non supervisé					
GPT3-ICL (Suri *et al.*, 2023)	19,87	8,67	15,60	-	57,03
Llama-3.1-8B-Instruct (Dubey *et al.*, 2024)	28,29±2,37	10,19±1,53	21.14±1,79	7,57±1,28	55,82±1,54
DeepSeek-R1-Distill-Llama-8B (Guo *et al.*, 2025)	17,76±2,31	5,68±1,14	13.02±1,71	2,46±0,60	47,54±1,49
DeepSeek-R1-Distill-Qwen-14B (Guo *et al.*, 2025)	26,31±2,85	10,17±1,84	19,78±2,33	4,29±1,01	52,84±1,69
QUARTZ (Meilleur résumeur : Llama 3.1)	**34,63**±2,88	**13,98**±2,19	**27,72**±2,60	**12,75**±2,12	**61,18**±1,73

TABLEAU 2 – Performance de *QUARTZ* sur les ensembles de test de DialogSum et MTS-Dialog.

Configuration	Initial	0-Shot *QUARTZ*	10% Sup.	*QUARTZ*	*QUARTZ* + 10% Sup.	Tout Sup.
BERT-Score	53,70	65,99	67,10	68,72	73,02	73,06
Rouge-L	16,46	29,16	29,38	32,68	37,80	37,73

TABLEAU 3 – Impact des variantes de supervision sur *QUARTZ*. **"Sup."** désigne le ratio de données d'entraînement supervisées utilisées.

Protocole d'évaluation. Conformément aux travaux précédents, nous évaluons les résumés de test par les métriques ROUGE (R-1, R-2 et R-L) (Lin, 2004), BLEU (Papineni *et al.*, 2002) et BERT-Score (Zhang* *et al.*, 2020). Afin d'assurer la fiabilité des résultats, nous utilisons le rééchantillonnage Jackknife (Efron & Stein, 1981) pour estimer les intervalles de confiance à 95%.

Bases de comparaison. À notre connaissance, il existe peu d'approches non supervisées pour le résumé de dialogues orienté tâche (Section 2). Ainsi, nous comparons notre méthode avec des méthodes entièrement supervisées, à l'exception de GPT3-ICL (In-Context-Learning) (Suri *et al.*, 2023) utilisé pour le résumé de conversations médicales dans MEDIQA-Chat 2023 (Ben Abacha *et al.*, 2023). Malgré cette comparaison inéquitable, nous montrons que *QUARTZ* exploite efficacement les modèles de manière collaborative, produisant des résumés orientés tâche rivalisant avec ceux des méthodes supervisées. Tian *et al.* (2024) ont proposé une méthode MoE basée sur le routage orienté rôle pour le résumé de dialogues. Wang *et al.* (2023) ont proposé une méthode d'*instruction-tuning* pour le résumé de dialogues instructifs, tandis que Abacha *et al.* (2023) ont utilisé des données augmentées pour le pré-entraînement avant de procéder à un affinage sur des données supervisées. En revanche, Suri *et al.* (2023) ont utilisé GPT3 avec des conversations similaires et des résumés pour chaque échantillon de test. Nous étendons notre comparaison aux modèles distillés DeepSeek-R1 (Guo *et al.*, 2025), disponibles en tailles 8B et 14B, pour étudier l'impact du raisonnement Chaîne de Pensées (Wei *et al.*, 2022) sur le résumé de dialogues orienté tâche.

Résumés	R-1	R-2	R-L	BLEU	BERT-Score
Génération (Première Étape)					
Lama 3.1	40,52	16,12	32,38	17,06	69,03
Gemma 2	39,35	15,16	31,38	13,32	66,45
Qwen 2	38,29	14,76	30,29	13,59	67,84
Évaluation (Deuxième Étape)					
Sélectionnés	**45.39**	**20,09**	**36,97**	**19,38**	**71,14**

TABLEAU 4 – Évaluation des résumés générés puis sélectionnés sur DialogSum.

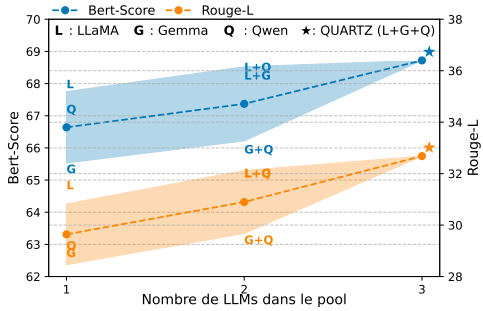

FIGURE 2 – Impact de la taille du pool de LLM sur la performance, avec moyenne ***(pointillés)*** et écart-type ***(ombré)*** (voir Annexe A.1).

5 Résultats

5.1 Analyse quantitative

Comparaison avec les méthodes de pointe. Comme première étape dans l'analyse des résultats expérimentaux, nous utilisons des métriques établies pour capturer à la fois le recouvrement de N-gramme (R-1, R-2, R-L et BLEU) et l'information abstraite (BERT-Score). Bien que nous reconnaissons que ces métriques sont insuffisantes pour évaluer pleinement la qualité des résumés, elles fournissent un aperçu utile et forment une base pour une analyse plus approfondie.

① DialogSum : Le Tableau 2 (haut) montre que *QUARTZ* surpasse les méthodes supervisées SoTA avec un BERT-Score de 68.72%±0.44. DeepSeek-R1-Distill-Qwen-14B est la meilleure méthode de référence non supervisée, surpassant légèrement les autres. Le meilleur résumeur (Llama-3.1-8B-Instruct) voit sa performance en BERT-Score s'améliorer de 28% après la troisième étape de *QUARTZ*. Le Tableau 4 présente les performances des différents résumeurs sur ce jeu de données pour la première étape (Génération) ainsi que les résultats de la deuxième étape (Évaluation). Cette dernière illustre comment *QUARTZ* améliore la performance des modèles grâce à la sélection collaborative des résumés générés.

② MTS-Dialog : Ici encore, le meilleur résumeur est Llama-3.1-8B-Instruct, dont la performance s'améliore de 31%, 9% et 22% en termes de Rouge-L, BERT-Score et BLEU, respectivement. Notamment, il dépasse BART-GS-DA (Abacha *et al.*, 2023) de 50% en BERT-Score tout en surpassant aussi d'autres méthodes SoTA non supervisées, y compris GPT-3, Llama et DeepSeek.

Impact de la supervision. Bien que *QUARTZ* puisse surpasser les méthodes supervisées (voir Tableau 2), nous examinons comment il peut tirer parti d'un volume limité de données supervisées (10% de l'ensemble d'entraînement) pour améliorer encore ses performances. Le Tableau 3 montre que l'affinage LoRA du modèle *QUARTZ* en utilisant seulement 10% des données d'entraînement (*QUARTZ + 10% Sup.*) produit des résultats presque identiques à ceux obtenus par un entraînement entièrement supervisé. En revanche, l'entraînement direct du même modèle de base (Llama-3.1-8B-Instruct) sur ces 10% de données n'atteint pas un BERT-Score supérieur à 68%, ce qui met en évidence les connaissances supplémentaires introduites par *QUARTZ*. Il est à noter que *0-Shot QUARTZ*, c'est-à-dire l'application directe de *QUARTZ* sur les exemples de test sans affinage (en

omettant la Troisième Étape), dépasse le modèle *Initial* de $+23\%$ en BERT-Score et de $+77\%$ en Rouge-L. Cela souligne la capacité de *QUARTZ* à maintenir des performances élevées même dans des contextes où les dialogues orientés tâche sont rares ou lorsque l'affinage est inapplicable.

Les avantages d'un pool de LLMs diversifié par rapport à un modèle unique. Des travaux récents (Subramaniam *et al.*, 2025) soutiennent que l'utilisation d'un pool de LLMs favorise la diversité et enrichit les interactions entre modèles. Cela est clairement démontré dans le Tableau 4, où les résumés sélectionnés à l'aide du pool de LLMs surpassent ceux générés par chaque modèle individuellement selon toutes les métriques. Plus précisément, bien que Llama-3.1-8B-Instruct ait produit 65% des meilleurs résumés sur l'ensemble des jeux de données, les contributions des autres modèles restent significatives (voir Annexe A.2, Figure 3). Dans la Figure 2, nous examinons l'impact de la taille du pool de LLMs ($|L|$) sur les performances. Pour chaque configuration de taille de pool, nous avons exploré toutes les combinaisons possibles (détails en Annexe A.1) et avons constaté qu'avec l'augmentation de la taille du pool, un ensemble plus diversifié de résumés est exploré, offrant un plus grand potentiel de trouver des résumés prometteurs, au prix d'un temps de calcul accru.

5.2 Analyse qualitative

Bien que *QUARTZ* soit conçu pour favoriser la génération de résumés factuellement cohérents en évaluant les réponses aux questions liées à la tâche, les métriques classiques (Section 5.1) ne capturent pas réellement cet aspect.

Amélioration de la pertinence et de la fiabilité factuelle. Les meilleurs résumés générés sélectionnés intègrent systématiquement plus d'informations liées à la tâche. Cela est rendu possible grâce à la génération de questions-réponses spécifiques à la tâche et à l'assurance de la cohérence factuelle tout au long du processus de classement des réponses candidates. En effet, le meilleur résumé sélectionné apporte de nouvelles informations par rapport aux autres résumés (cf. Annexe E, Figure 4).

Impact de l'affinage sur la qualité du résumé. Sur le jeu de données SimSAMU, l'expert médical impliqué dans cette étude a noté que, au-delà de la production de notes cliniques plus précises, les résumés affinés étaient plus clairs et allaient directement à l'essentiel par rapport à ceux générés avant l'affinage. Cet aspect est particulièrement apprécié par les régulateurs d'urgence dans des situations où le facteur temps est critique. Le Tableau 5 met en évidence les améliorations ainsi que les potentielles dégradations observées. De plus, les résumés affinés incluent souvent des détails pertinents sur le plan de traitement et les décisions de régulation, renforçant ainsi leur utilité (cf. Annexe E et Tableau 8).

6 LLM as a Judge

Motivés par des travaux récents montrant que les LLMs peuvent servir de substituts fiables aux annotateurs humains (Song *et al.*, 2024), nous utilisons le cadre G-Eval (Liu *et al.*, 2023) avec deux modèles open-source parmi les mieux classés sur Judge Arena (Zheng *et al.*, 2023) : *Selene-1-Mini-Llama-3.1-8B* (Alexandru *et al.*, 2025) et *Llama-3.3-70B-Instruct* (Dubey *et al.*, 2024). Contrairement à l'approche initiale basée sur GPT-4, afin d'assurer la reproductibilité. Nous évaluons les résumés selon quatre dimensions : cohérence (COH), consistance (CON), fluidité (FLU) et pertinence (PER) (voir Détails en Annexe D.4). Comme le montre le Tableau 6, les deux juges attribuent des scores souvent plus faibles aux résumés de référence, indiquant que les LLMs produisent déjà de bons candidats.

Zero-Shot *QUARTZ* (Première + deuxième étape uniquement)	*QUARTZ* (Trois étapes)	Note
Trouble de l'enfant à la suite de coups de la part du père.	**Violence conjugale sur un enfant** : le patient a appelé pour signaler que son mari a frappé leur enfant, causant des **blessures**.	Clarification et précision
Hypothèses diagnostiques : Blessure par coup de couteau auto-infligée.	Hypothèses diagnostiques : Blessure par coup de couteau auto-infligée, **intention suicidaire potentielle**.	Indicateurs d'intention suicidaire
Antécédents médicaux : Pas d'informations fournies.	Antécédents médicaux : **Intoxication volontaire l'année dernière.**	Précision médicale
Le conducteur a essayé de se relever mais n'y est pas parvenu.	Le conducteur du scooter a essayé de se relever mais n'y est pas parvenu. **Le patient est incapable d'approcher le conducteur du scooter en raison de son travail et doit partir.**	Précision contextuelle
Symptômes du patient : Pas de symptômes médicaux rapportés.	Symptômes généraux : Anxiété ou inquiétude due à la situation d'incendie.	Ajout de l'état émotionnel
La grand-mère du patient ne répond plus au téléphone **depuis au plus 48 heures.**	La grand-mère du patient ne répond plus au téléphone.	Absence de détails
Antécédents médicaux : -Diabète. -**Allergie au DOLIPRANE.**	Antécédents médicaux : -Diabète.	Omission allergie

TABLEAU 5 – Impact de la troisième étape de *QUARTZ* (Affinage) sur l'alignement du modèle. **Gauche :** *QUARTZ* en zéro-shot génère plusieurs résumés et sélectionne les meilleurs. **Droite :** Un affinage supplémentaire est appliqué aux résumés sélectionnés.

Bien que les scores absolus des juges diffèrent, les classements relatifs sont globalement cohérents. La configuration 0-Shot (*QUARTZ* : Étapes 1 + 2) permet d'identifier des résumés performants sur certaines dimensions (ex. : CON, FLU), mais il reste difficile de trouver des candidats équilibrés sur l'ensemble des critères, en raison des compromis inhérents aux résumés générés. L'étape de fine-tuning permet de corriger cela en orientant la génération vers de meilleurs compromis globaux.

Juge / Config	Selene-1-Mini-Llama-3.1-8B					Llama-3.3-70B-Instruct				
	COH	CON	FLU	PER	Moy	COH	CON	FLU	PER	Moy
Llama 3.1	3,79	4,12	3,54	3,86	3,82	4,26	4,34	3,92	4,29	4,20
Gemma	3,77	4,13	3,62	**4,04**	**3,89**	**4,41**	4,39	**4,12**	**4,52**	**4,36**
Qwen	**3,88**	4,12	3,63	3,96	**3,89**	4,29	4,37	3,98	4,41	4,26
0-Shot	3,86	**4,14**	**3,66**	3,86	3,88	4,28	**4,40**	4,00	4,36	4,26
Reference	3,47	3,91	3,31	3,52	3,55	3,89	3,92	3,60	3,86	3,81
QUARTZ	**3,89**	**4,18**	**3,66**	3,87	**3,90**	**4,47**	**4,41**	3,96	**4,53**	**4,34**

TABLEAU 6 – Évaluation LLM-as-Judge des résumés par deux modèles selon les cinq critères.

	COH	CON	FLU	PER	Moy
QUARTZ (1–5 Likert)	4,17	4,01	4,27	4,06	4,12
Cohen's kappa ([-1 ; 1])	0,12	0,17	0,14	0,09	0,13
Accord exact (%)	42	42	40	34	39,5

TABLEAU 7 – Évaluation humaine des résumés finaux de *QUARTZ*, avec les coefficients Cohen's kappa et les taux d'accord exact.

7 Évaluation humaine

Étant donné le coût élevé de l'évaluation humaine, nous adoptions un protocole en deux phases avec quatre annotateurs experts (tous informaticiens, dont un médecin afin de mieux annoter les dialogues cliniques). **Phase 1.** À l'aide de l'outil Potato (Pei *et al.*, 2022), les annotateurs comparent deux résumés anonymisés (généré par *QUARTZ* vs. référence) pour chaque dialogue, en choisissant celui de meilleure qualité. Dans 48% des cas, les résumés *QUARTZ* ont été préférés, avec un accord

inter-annotateur (Fleiss' kappa) de 0,14. Certains annotateurs ont signalé que les deux résumés étaient souvent d'une qualité similaire. **Phase 2.** Les annotateurs évaluent les résumés finaux de *QUARTZ* selon les quatre dimensions qualitatives définies en Section 6. Le taux moyen d'accord inter-annotateur était de 39,5%.

8 Limitations

Bien que nous ayons cherché à minimiser les coûts d'annotation tout en maintenant une haute qualité des résumés, quelques limitations subsistent. Premièrement, par souci de simplicité, nous n'avons pas modifié les paires question-réponse de référence générées. Bien que nous puissions explicitement intégrer des questions pertinentes pour la tâche, il serait bénéfique d'établir un certain contrôle sur les questions en fonction de leur contribution à l'évaluation des résumés, car toutes les paires question-réponse n'ont pas la même importance. Deuxièmement, bien que *QUARTZ* soit conçu pour le résumé non supervisé, la dépendance aux performances des LLMs peut entraîner des divergences par rapport aux résumés préférés par les humains, ainsi que des erreurs qu'une supervision plus fine permettrait d'éviter.

9 Conclusion et perspectives

Dans ce travail, nous avons introduit *QUARTZ*, une méthode conçue pour améliorer de manière collaborative les LLMs dans la tâche de résumé de dialogue non supervisé. *QUARTZ* exploite d'abord les LLMs pour générer un ensemble diversifié de résumés et de paires question-réponse liées à la tâche. La qualité des résumés est ensuite évaluée à travers un processus en deux niveaux, qui sélectionne les meilleures réponses, puis identifie les résumés les plus informatifs. Enfin, le meilleur LLM résumeur est affiné sur ces résumés sélectionnés. Les expériences menées sur nos ensembles de données couvrant différents domaines démontrent l'efficacité de *QUARTZ*, surpassant constamment les méthodes supervisées de l'état de l'art. Nous croyons que *QUARTZ* peut répondre à de nombreux cas d'usage réels, tels que le résumé de conversations médicales ou de réunions, en assistant les professionnels dans leurs tâches quotidiennes. Son application aux services médicaux peut aider les praticiens à structurer les informations clés, ouvrant ainsi la voie à la recherche sur l'extraction d'entités et la communication clinique. Dans nos travaux futurs, nous prévoyons d'examiner l'impact de *QUARTZ Itératif* (sélection et spécialisation multi-itérations des LLMs) sur la qualité des résumés. De plus, nous explorerons l'intégration de *QUARTZ* avec d'autres techniques d'affinage, en particulier la Group Relative Policy Optimization (Guo *et al.*, 2025), afin d'aborder la gestion du pool de LLMs sous un angle différent.

Références

ABACHA A. B., YIM W.-W., FAN Y. & LIN T. (2023). An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, p. 2291–2302.

ALEXANDRU A., CALVI A., BROOMFIELD H., GOLDEN J., DAI K., LEYS M., BURGER M., BARTOLO M., ENGELER R., PISUPATI S., DRANE T. & PARK Y. S. (2025). Atla selene mini : A general purpose evaluation model.

ALSAEDI N., BURNAP P. & RANA O. (2016). Temporal TF-IDF : A high performance approach for event summarization in Twitter. In *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, p. 515–521.

BAVARESCO A., BERNARDI R., BERTOLAZZI L., ELLIOTT D., FERNÁNDEZ R., GATT A., GHALEB E., GIULIANELLI M., HANNA M., KOLLER A. *et al.* (2024). Llms instead of human judges ? a large scale empirical study across 20 nlp evaluation tasks. *CoRR*.

BEN ABACHA A., YIM W.-W., ADAMS G., SNIDER N. & YETISGEN M. (2023). Overview of the MEDIQA-chat 2023 shared tasks on the summarization & generation of doctor-patient conversations. In T. NAUMANN, A. BEN ABACHA, S. BETHARD, K. ROBERTS & A. RUMSHISKY, Éds., *Proceedings of the 5th Clinical Natural Language Processing Workshop*, p. 503–513, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.clinicalnlp-1.52.

BROWN T., MANN B., RYDER N., SUBBIAH M., KAPLAN J. D., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A. *et al.* (2020). Language models are few-shot learners. *Advances in neural information processing systems*, **33**, 1877–1901.

BULIAN J., BUCK C., GAJEWSKI W., BÖRSCHINGER B. & SCHUSTER T. (2022). Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 291–305.

CHEN D., FISCH A., WESTON J. & BORDES A. (2017). Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)* : Association for Computational Linguistics.

CHEN Y., LIU Y., CHEN L. & ZHANG Y. (2021). Dialogsum : A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*, p. 5062–5074.

CHINTAGUNTA B., KATARIYA N., AMATRIAIN X. & KANNAN A. (2021). Medically aware GPT-3 as a data generator for medical dialogue summarization. In *Machine Learning for Healthcare Conference*, p. 354–372.

DUBEY A., JAUHRI A., PANDEY A., KADIAN A., AL-DAHLE A., LETMAN A., MATHUR A., SCHELTEN A., YANG A., FAN A. *et al.* (2024). The llama 3 herd of models. *arXiv preprint arXiv :2407.21783*.

EFRON B. & STEIN C. (1981). The jackknife estimate of variance. *The Annals of Statistics*, p. 586–596.

FEIGENBLAT G., GUNASEKARA C., SZNAJDER B., JOSHI S., KONOPNICKI D. & AHARONOV R. (2021). Tweetsumm-a dialog summarization dataset for customer service. In *Findings of the Association for Computational Linguistics : EMNLP 2021*, p. 245–260.

FENG S. Y., GANGAL V., WEI J., CHANDAR S., VOSOUGHI S., MITAMURA T. & HOVY E. (2021). A survey of data augmentation approaches for nlp. In *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*, p. 968–988.

FENG X., FENG X. & QIN B. (2022). A survey on dialogue summarization : Recent advances and new frontiers. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence*.

GUNASEKARA C., FEIGENBLAT G., SZNAJDER B., JOSHI S. & KONOPNICKI D. (2021). Summary grounded conversation generation. In *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*, p. 3748–3756.

GUO D., YANG D., ZHANG H., SONG J., ZHANG R., XU R., ZHU Q., MA S., WANG P., BI X. *et al.* (2025). Deepseek-r1 : Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv :2501.12948*.

GUO Z., WANG P., WANG Y. & YU S. (2023). Improving small language models on PubMedQA via generative data augmentation. *arXiv preprint arXiv :2305.07804*.

HU E. J., WALLIS P., ALLEN-ZHU Z., LI Y., WANG S., WANG L., CHEN W. *et al.* (2022). Lora : Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

IZACARD G. & GRAVE É. (2021). Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, p. 874–880.

JACOBS R. A., JORDAN M. I., NOWLAN S. J. & HINTON G. E. (1991). Adaptive mixtures of local experts. *Neural computation*, **3**(1), 79–87.

KAMALLOO E., UPADHYAY S. & LIN J. (2024). Towards robust QA evaluation via open LLMs. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 2811–2816.

KENDALL M. G. (1938). A new measure of rank correlation. *Biometrika*, **30**(1-2), 81–93.

LIN C.-Y. (2004). ROUGE : A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, p. 74–81, Barcelona, Spain : Association for Computational Linguistics.

LIN H. & NG V. (2019). Abstractive summarization : A survey of the state of the art. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, p. 9815–9822.

LIU Y., ITER D., XU Y., WANG S., XU R. & ZHU C. (2023). G-eval : Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 2511–2522.

LIU Y., MAYNEZ J., SIMÕES G. & NARAYAN S. (2022). Data augmentation for low-resource dialogue summarization. In *Findings of the Association for Computational Linguistics : NAACL 2022*, p. 703–710.

MIHALCEA R. & TARAU P. (2004). TextRank : Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, p. 404–411.

NENKOVA A., MCKEOWN K. *et al.* (2011). Automatic summarization. *Foundations and Trends® in Information Retrieval*, **5**(2–3), 103–233.

NUN A., OLIVIER B., GAËL G., FRÉDÉRIC L. & IVAN L. (2025). Simsamu - a french medical dispatch dialog open dataset. *Computer Methods and Programs in Biomedicine*, p. 108857. DOI : https ://doi.org/10.1016/j.cmpb.2025.108857.

OUYANG S., CHEN J., HAN J. & YANG D. (2023). Compositional data augmentation for abstractive conversation summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1471–1488.

PANICKSSERY A., BOWMAN S. R. & FENG S. (2024). LLM evaluators recognize and favor their own generations. *arXiv preprint arXiv :2404.13076*.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, p. 311–318.

PEI J., ANANTHASUBRAMANIAM A., WANG X., ZHOU N., DEDELOUDIS A., SARGENT J. & JURGENS D. (2022). Potato : The portable text annotation tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*.

PENG B., LI C., HE P., GALLEY M. & GAO J. (2023). Instruction tuning with GPT-4. *arXiv preprint arXiv :2304.03277*.

PU X., GAO M. & WAN X. (2023). Summarization is (almost) dead. *arXiv preprint arXiv :2309.09558*.

RADEV D. R., QI H., WU H. & FAN W. (2002). Evaluating web-based question answering systems. In M. GONZÁLEZ RODRÍGUEZ & C. P. SUAREZ ARAUJO, Éds., *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain : European Language Resources Association (ELRA).

RENNARD V., SHANG G., HUNTER J. & VAZIRGIANNIS M. (2023). Abstractive meeting summarization : A survey. *Transactions of the Association for Computational Linguistics*, **11**, 861–884.

RIABI A., SCIALOM T., KERARON R., SAGOT B., SEDDAH D. & STAIANO J. (2021). Synthetic data augmentation for zero-shot cross-lingual question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 7016–7030.

SONG H., SU H., SHALYMINOV I., CAI J. & MANSOUR S. (2024). Finesure : Fine-grained summarization evaluation using llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 906–922.

SUBRAMANIAM V., DU Y., TENENBAUM J. B., TORRALBA A., LI S. & MORDATCH I. (2025). Multiagent finetuning : Self improvement with diverse reasoning chains. *arXiv preprint arXiv :2501.05707*.

SURI K., SAHA S. & SINGH A. (2023). Healthmavericks@ mediqa-chat 2023 : Benchmarking different transformer based models for clinical dialogue summarization. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, p. 472–489.

TANG R., HAN X., JIANG X. & HU X. (2023). Does synthetic data generation of LLMs help clinical text mining ? *arXiv preprint arXiv :2303.04360*.

TANG R., ZHANG C., MA X., LIN J. & TÜRE F. (2024). Found in the middle : Permutation self-consistency improves listwise ranking in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (Volume 1 : Long Papers)*, p. 2327–2340.

TEAM G., RIVIERE M., PATHAK S., SESSA P. G., HARDIN C., BHUPATIRAJU S., HUSSENOT L., MESNARD T., SHAHRIARI B., RAMÉ A. *et al.* (2024). Gemma 2 : Improving open language models at a practical size. *arXiv preprint arXiv :2408.00118*.

TIAN Y., XIA F. & SONG Y. (2024). Dialogue summarization with mixture of experts based on large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 7143–7155.

TONMOY S. M. T. I., ZAMAN S. M. M., JAIN V., RANI A., RAWTE V., CHADHA A. & DAS A. (2024). A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv :2401.01313*.

TÖRNBERG P. (2023). ChatGPT-4 outperforms experts and crowd workers in annotating political Twitter messages with zero-shot learning. *arXiv preprint arXiv :2304.06588*.

VAN VEEN D., VAN UDEN C., BLANKEMEIER L., DELBROUCK J.-B., AALI A., BLUETHGEN C., PAREEK A., POLACIN M., REIS E. P., SEEHOFNEROVÁ A. *et al.* (2024). Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, **30**(4), 1134–1142.

WANG B., LIU Z. & CHEN N. (2023). Instructive dialogue summarization with query aggregations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 7630–7653.

WANG C., CHENG S., GUO Q., YUE Y., DING B., XU Z., WANG Y., HU X., ZHANG Z. & ZHANG Y. (2024). Evaluating open-qa evaluation. *Advances in Neural Information Processing Systems*, **36**.

WEI J., WANG X., SCHUURMANS D., BOSMA M., ICHTER B., XIA F., CHI E., LE Q. V. & ZHOU D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In S. KOYEJO, S. MOHAMED, A. AGARWAL, D. BELGRAVE, K. CHO & A. OH, Éds., *Advances in Neural Information Processing Systems*, volume 35, p. 24824–24837.

XIA M., KONG X., ANASTASOPOULOS A. & NEUBIG G. (2019). Generalized data augmentation for low-resource translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 5786–5796.

XU W., ZHU G., ZHAO X., PAN L., LI L. & WANG W. (2024). Pride and prejudice : LLM amplifies self-bias in self-refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 15474–15492.

YANG A., YANG B., HUI B., ZHENG B., YU B., ZHOU C., LI C., LI C., LIU D., HUANG F. *et al.* (2024). Qwen2 technical report. *CoRR*.

YANG W., XIE Y., TAN L., XIONG K., LI M. & LIN J. (2019). Data augmentation for BERT fine-tuning in open-domain question answering. *arXiv preprint arXiv :1904.06652*.

ZECHNER K. (2002). Automatic summarization of open-domain multiparty dialogues in diverse genres. *Computational Linguistics*, **28**(4), 447–485.

ZHANG* T., KISHORE* V., WU* F., WEINBERGER K. Q. & ARTZI Y. (2020). Bertscore : Evaluating text generation with bert. In *International Conference on Learning Representations*.

ZHAO L., ZHENG F., HE K., ZENG W., LEI Y., JIANG H., WU W., XU W., GUO J. & MENG F. (2021). TODSum : Task-oriented dialogue summarization with state tracking. *arXiv preprint arXiv :2110.12680*.

ZHENG L., CHIANG W.-L., SHENG Y., ZHUANG S., WU Z., ZHUANG Y., LIN Z., LI Z., LI D., XING E. *et al.* (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, **36**, 46595–46623.

ZOU Y., ZHU B., HU X., GUI T. & ZHANG Q. (2021). Low-resource dialogue summarization with domain-agnostic multi-source pretraining. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 80–91.

A Configuration du pool de LLMs et dynamiques de sélection

A.1 Détails sur les configurations du pool de LLMs

Nous avons présenté dans la Figure 2 comment le nombre $|L|$ de LLMs dans le pool impacte la qualité du résumé. Les LLMs sont choisis parmi ces 3 modèles : Llama-3.1-8B-Instruct, Gemma-2-9b-it, et Qwen2-7B-Instruct. Pour chaque taille de pool $|L|$, nous évaluons toutes les $\mathcal{C}_3^{|L|}$ combinaisons possibles de modèles. Dans la Figure 2, les lignes pointillées représentent les valeurs moyennes, tandis que les zones ombrées indiquent l'écart type. Pour $|L| = 3$, il n'y a qu'une seule combinaison possible, qui correspond à la configuration réelle de *QUARTZ* rapportée dans la dernière ligne du Tableau 2 pour le jeu de données DialogSum.

A.2 Priorisation du meilleur LLM résumeur pour l'affinage

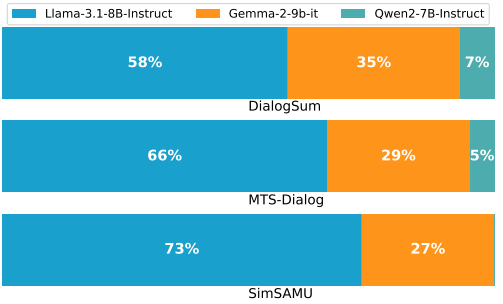

FIGURE 3 – Taux de victoire des LLMs sur les jeux de données.

La dernière étape de *QUARTZ* (voir Section 3.3) vise à affiner le meilleur LLM résumeur, c'est-à-dire le modèle ayant généré la majorité des résumés les mieux classés. La Figure 3 illustre la proportion de résumés sélectionnés comme les meilleurs à travers les différents ensembles de données, avec Llama-3.1-8B-Instruct contribuant à 65 %, Gemma-2-9b-it à 30% et Qwen2-7B-Instruct à 5%. Les résultats empiriques montrent que l'affinage du modèle le plus performant produit les meilleurs résultats de résumé, surpassant les autres choix possibles pour cette dernière étape.

B Détails de l'implémentation

Nous affinons le LLM résumeur en utilisant LoRA (Hu *et al.*, 2022) pendant 3 époques avec un rang $r_{\text{LoRA}} = 8$ et un facteur d'échelle $\alpha_{\text{LoRA}} = 16$. L'optimisation est réalisée à l'aide de l'optimiseur AdamW avec un taux d'apprentissage de 5×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$ et $\epsilon = 1 \times 10^{-8}$. Un planificateur de taux d'apprentissage linéaire est appliqué. Toutes les expériences ont été menées sur un seul GPU A100-40GB. Pour l'ensemble du pipeline *QUARTZ* (composé des trois étapes), le temps de calcul total s'élève à 12 heures GPU. Une fois le meilleur résumeur obtenu, il peut être utilisé pour de l'inférence en ligne et être optimisé davantage à l'aide de kits d'optimisation pour l'inférence.

C Consignes pour les annotateurs

Afin de garantir une évaluation humaine fiable, nous avons mis en place un processus d'annotation en deux phases impliquant quatre annotateurs experts. Lors de la **Phase 1**, les annotateurs recevaient un dialogue accompagné de deux résumés anonymisés (l'un généré par *QUARTZ*, l'autre étant le résumé de référence) et devaient sélectionner celui qu'ils jugeaient de meilleure qualité. En **Phase 2**, chaque annotateur évaluait individuellement les résumés produits par *QUARTZ* selon quatre dimensions qualitatives — *Cohérence*, *Consistance*, *Fluidité* et *Pertinence* — sur une échelle de Likert à 5 points. Les critères d'évaluation étaient les suivants :

— **Cohérence (1–5) :** Juge la fluidité logique et la structure générale du résumé.
— **Consistance (1–5) :** Évalue l'alignement factuel avec le dialogue source.
— **Fluidité (1–5) :** Apprécie la qualité grammaticale, la lisibilité et le style linguistique.
— **Pertinence (1–5) :** Mesure la capacité du résumé à capturer les éléments clés sans redondance.

D Modèles de prompt

D.1 Génération de résumé

DialogSum :

"instruction": `You will be provided with a conversational exchange that simulates a natural messaging or chat-like interaction. Your task is to produce a short, concise and clear summary that captures the most important points and key information conveyed throughout the exchange.`

"input": `[Conversation]`

MTS-Dialog :

"instruction": `You are a medical scribe tasked with writing concise yet informative medical notes based on doctor-patient interactions. Your goal is to create clear and professional note text about from the patient doctor dialogue focusing on [Header]`

"input": `[Conversation]`

SimSAMU :

L'ingénierie de *prompt* pour cet ensemble de données lié à la régulation médicale a été assistée par un expert médical afin de garantir la pertinence contextuelle des notes générées.

"instruction": `Vous êtes un médecin urgentiste expérimenté gérant une consultation téléphonique. Votre tâche est de résumer le dialogue médical suivant sous la forme d'un compte rendu clinique précis et structuré. Lors du résumé, assurez-vous de :`

- Utiliser un langage concis, clair et professionnel.
- Traduire les termes informels ou courants en terminologie médicale appropriée lorsque possible.
- Respecter la structure fournie ci-dessous, en laissant vides les sections pour lesquelles les informations ne sont pas disponibles.

Format du compte rendu clinique:
Veuillez compléter les sections suivantes dans l'ordre indiqué :
1-Motif principal de l'appel: Problème médical principal ayant motivé l'appel (ex. : douleur thoracique).
2-Contexte de l'appel: Relation entre l'appelant et le patient (ex. : patient lui-même, conjoint, témoin) et lieu de l'appel (ex. : domicile, rue).
3-Contexte du patient: Informations démographiques (âge, sexe), situation sociale (ex. : vit seul, en maison de retraite) et degré d'autonomie.
4-Traitement habituel: Médicaments ou traitements en cours pour des comorbidités connues.
5-Antécédents médicaux: Antécédents médicaux, allergiques ou chirurgicaux pertinents (ex. : diabète, chirurgies antérieures).
6-Symptômes du patient: Symptômes rapportés, classés en:
- Symptômes généraux (ex. : fièvre, fatigue).
- Symptômes spécifiques à un organe (ex. : respiratoire : dyspnée).

7-Histoire de la maladie actuelle: Récit détaillé et chronologique des événements ayant conduit à l'appel, décrivant leur enchaînement et leurs liens (texte libre).
8-Hypothèses diagnostiques: Diagnostiques possibles sur la base des informations fournies (ex. : infarctus du myocarde).
9-Plan de traitement: Recommandations incluant médicaments, thérapies, conseils de mode de vie, orientations ou examens complémentaires.
10-Décision d'orientation: Conduite à tenir proposée (ex. : maintien à domicile, envoi d'un médecin, transport autonome aux urgences, envoi d'une ambulance, envoi d'une ambulance médicalisée).

"input": [Conversation]

D.2 Génération de questions-réponses (QAs)

SimSAMU :

"instruction":Étant donné cet appel d'urgence entre un médecin et un patient, génère 10 à 15 paires de questions et réponses pertinentes pour le triage médical et qui devraient figurer dans le dossier clinique.
Ne répète pas la même question ou réponse.

Ne pose pas de questions auxquelles tu ne peux pas répondre sur la base des informations fournies dans le dialogue.
Formate les questions et réponses comme suit : Q1 : <question1> R1 : <réponse1>

"input":[Conversation]

D.3 Évaluation des réponses

DialogSum et MTS-Dialog :

"instruction": You will be provided with a ground truth answer and a list of generated answers. Your task is to rank the generated answers based on their correctness and closeness to the ground truth answer. The ground truth answer appears first, followed by the list of generated answers. The ranking should be a three-element list of integers between 1 and 3, where 1 represents the most accurate answer and 3 represents the least accurate. If an answer is NOT_INCLUDED, place it at the end of the ranking.
"input": Question: [Question]
Ground Truth Answer: [Ground Truth Answer]
Possible answers:
1) [Answer_1]
2) [Answer_2]
3) [Answer_3]

D.4 LLM as a Judge

Métrique de cohérence :

"instruction": You are a helpful assistant tasked with evaluating how coherent a summary is for a given dialogue. Your goal is to rate the summary based only on how well its sentences form a clear, logical, and well-structured presentation of the dialogue content. Assign a score from 1 to 5 based solely on *coherence*:
5: Excellent - the summary is well-organized, easy to follow, and logically structured.
4: Good - mostly coherent with only minor issues in flow or structure.
3: Fair - somewhat coherent but with noticeable issues in clarity or organization.
2: Poor - disorganized, unclear, or hard to follow.
1: Very poor - sentences feel disconnected or incoherent, severely impacting understanding.
Only reply with the number **1**, **2**, **3**, **4**, or **5**.
Do not include any explanation or extra text.
Your reply should strictly follow this format: **Score:** <1, 2,

3, 4, or 5>

"input": Dialogue: [Dialogue]
Summary: [Summary]

Métrique de consistance :

"instruction": You are a helpful assistant tasked with evaluating how factually consistent a summary is with a given dialogue. Your goal is to rate the summary based only on whether it accurately reflects the facts stated in the original dialogue without introducing unsupported or hallucinated information. Assign a score from 1 to 5 based solely on *consistency*:
5: Excellent -- all statements in the summary are fully supported by the dialogue.
4: Good -- minor inaccuracies or slight overgeneralizations, but mostly faithful to the dialogue.
3: Fair -- some factual inconsistencies or minor hallucinations are present.
2: Poor -- several statements in the summary are not supported or contradict the dialogue.
1: Very poor -- the summary contains major hallucinations or is largely inconsistent with the dialogue.
Only reply with the number **1**, **2**, **3**, **4**, or **5**.
Do not include any explanation or extra text.
Your reply should strictly follow this format: **Score:** <1, 2, 3, 4, or 5>

"input": Dialogue: [Dialogue]
Summary: [Summary]

Métrique de fluidité :

"instruction": You are a helpful assistant tasked with evaluating how fluent a summary is for a given dialogue. Your goal is to rate the summary based only on grammar, spelling, punctuation, word choice, and sentence structure. Assign a score from 1 to 5 based solely on *fluency*:
5: Excellent -- the summary is free of errors and reads very smoothly.
4: Good -- the summary has minor errors but is easy to read.
3: Fair -- the summary has some noticeable errors that slightly affect clarity or flow.
2: Poor -- the summary has many errors that affect understanding or naturalness.
1: Very Poor -- the summary contains frequent errors making it difficult to understand.
Only reply with the number **1**, **2**, **3**, **4**, or **5**.
Do not include any explanation or extra text.

Your reply should strictly follow this format: **Score:** <1, 2, 3, 4, or 5>

"input": Dialogue: [Dialogue]
Summary: [Summary]

Métrique de pertinence :

"instruction": You are a helpful assistant tasked with evaluating how relevant a summary is for a given dialogue. Your goal is to rate the summary based only on how well it includes important information from the dialogue while avoiding irrelevant or redundant content. Assign a score from 1 to 5 based solely on *relevance*:
5: Excellent -- the summary covers all key points and contains no irrelevant or redundant information.
4: Good -- the summary covers most key points with minimal irrelevant or redundant content.
3: Fair -- the summary covers some key points but includes noticeable irrelevant or redundant information.
2: Poor -- the summary misses many key points and contains significant irrelevant or redundant content.
1: Very Poor -- the summary largely fails to capture key points and is mostly irrelevant or redundant.
Only reply with the number **1**, **2**, **3**, **4**, or **5**.
Do not include any explanation or extra text.
Your reply should strictly follow this format: **Score:** <1, 2, 3, 4, or 5>

"input": Dialogue: [Dialogue]
Summary: [Summary]

E Exemples de résumés de *QUARTZ*

SimSAMU :
Nous utilisons ROUGE pour indiquer les informations omises et VERT pour les nouveaux détails ajoutés, afin de mettre en évidence les modifications apportées avant et après l'affinage.

Exemple 1 :
1-Motif principal de l'appel: Blessure auto-infligée par arme blanche.
2-Contexte de l'appel: Le cousin du patient a appelé pour son frère, qui est le patient. L'appel était depuis le domicile du patient.
3-Contexte du patient: L'âge et le sexe ne sont pas disponibles. Le patient vit à domicile avec son frère. Le degré d'autonomie est incertain. Il est mentionné que le frère du patient se trouve dans

une résidence où réside également le patient.
4-Traitement habituel: Non disponible.
5-Antécédents médicaux: Non disponibles.
6-Symptômes du patient: Symptômes généraux : agitation, refus de demander de l'aide. Symptômes spécifiques aux organes : non disponibles.
7-Histoire de la maladie actuelle: Le frère du patient a été retrouvé avec du sang sur son pantalon après que le patient l'a vu se poignarder à nouveau. Le frère a refusé de laisser le patient examiner la blessure et ne voulait pas qu'il appelle à l'aide. Le cousin du patient a appelé les secours à la place. Le patient était réticent à impliquer les autorités, craignant la réaction de son frère.
8-Hypothèses diagnostiques: Blessure auto-infligée par arme blanche, potentiel risque suicidaire.
9-Plan de traitement: Envoi des pompiers sur place pour gérer la situation. Le patient est conseillé de rester auprès de son frère et d'informer les autorités de tout nouvel élément.
10-Décision de triage: Les services d'urgence ont décidé d'envoyer les pompiers sur place pour gérer la situation.

Exemple 2 :
1-Motif principal de l'appel: Malaise, malaise vagal possible, épi sode de sensation de partir sans perte de conscience, caractérisé par une sensation de syncope imminente sans perte de conscience réelle.
2-Contexte de l'appel: La patiente est la fille de l'homme de 85 ans éprouvant le malaise, appelant depuis son domicile.
3-Contexte du patient: Le patient est un homme de 85 ans avec hypertension, vivant avec sa fille, qui travaille autant que secouriste.
4-Traitement habituel: Il prend des traitements pour l'hypertension.
5-Antécédents médicaux: Hypertension.
6-Symptômes du patient: Symptômes généraux : Malaise, Sensation de syncope imminente. Symptômes spécifiques à un organe : Aucun mentionné.
7-Histoire de la maladie actuelle : Le patient a ressenti un malaise dans la salle de bain, caractérisé par une sensation de syncope imminente. Il a appelé sa fille, qui l'a aidé et il s'est rétabli., qui est arrivée et l'a mis en position couchée avec les jambes surélevées, ce qui a amélioré son état.
8-Hypothèses diagnostiques : Malaise vagal.
9-Plan de traitement : Le patient doit être transporté aux urgences pour une évaluation complémentaire. Il doit apporter sa prescription pour ses traitements habituels.
10-Décision de triage : Le patient doit être transporté aux

urgences par sa fille, qui doit assurer son confort pendant le transport.

DialogSum :
Dans la Figure 4, nous présentons un exemple de dialogue accompagné des résumés et des paires question-réponse générés par les LLMs. En particulier, le meilleur résumé sélectionné (LLM Résumeur : Gemma-2-9b-it) intègre des faits supplémentaires par rapport aux résumés des autres modèles et le résumé de référence. Bien que cela enrichisse le résumé orienté tâche en apportant davantage d'informations, cet apport est sous-évalué par les métriques de surface.

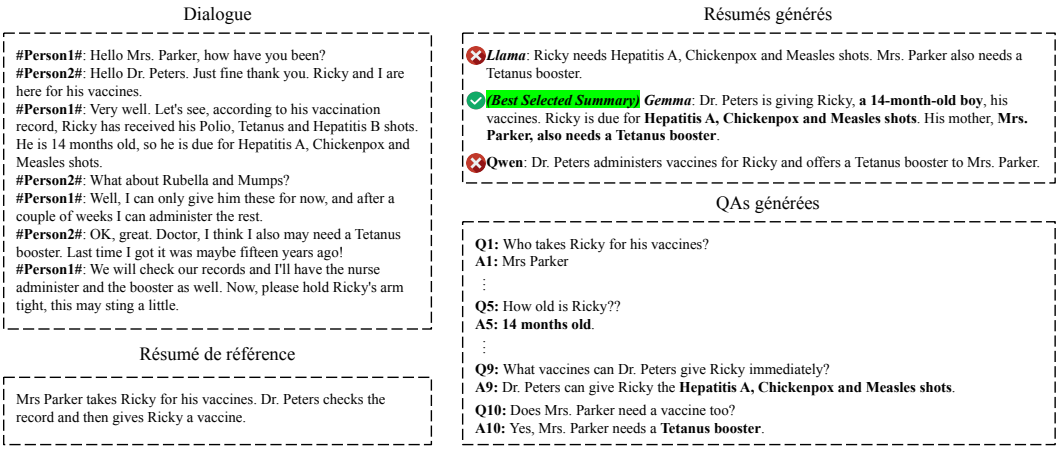

FIGURE 4 – Exemple de dialogue du jeu de données DialogSum. **Gauche :** Dialogue original et résumé de référence. **Droite :** Résumés générés par chaque LLM du pool, accompagnés des questions-réponses liées à la tâche. Les informations supplémentaires introduites par le meilleur résumé sélectionné sont mises en **gras**.

Zero-Shot *QUARTZ* (Première + deuxième étape uniquement)	***QUARTZ* (Trois étapes)**	**Note**
Trouble de l'enfant à la suite de coups de la part du père.	**Violence conjugale sur un enfant** : le patient a appelé pour signaler que son mari a frappé leur enfant, causant des **blessures**.	Clarification et précision
Hypothèses diagnostiques : Blessure par coup de couteau auto-infligée.	Hypothèses diagnostiques : Blessure par coup de couteau auto-infligée, **intention suicidaire potentielle**.	Indicateurs d'intention suicidaire
Antécédents médicaux : Pas d'informations fournies.	Antécédents médicaux : **Intoxication volontaire l'année dernière.**	Précision médicale
Le conducteur a essayé de se relever mais n'y est pas parvenu.	Le conducteur du scooter a essayé de se relever mais n'y est pas parvenu. **Le patient est incapable d'approcher le conducteur du scooter en raison de son travail et doit partir.**	Précision contextuelle
L'enfant doit être maintenu à domicile et surveillé par le père jusqu'à l'arrivée du médecin.	Le médecin a indiqué que le médecin est en route et que **le père doit continuer à appuyer sur le thorax de l'enfant jusqu'à l'arrivée du médecin.**	Clarification des instructions
Le patient éprouve un malaise et a fait une chute, avec des symptômes incluant des vertiges et une perte d'équilibre.	Malaise, **syncope (évanouissement)**, et une chute.	Comprend la terminologie médicale
Brûlure de degré inconnu, **nécessitant une évaluation médicale.**	Brûlure de degré inconnu.	Évaluation omise
La grand-mère du patient ne répond plus au téléphone **depuis au plus 48 heures.**	La grand-mère du patient ne répond plus au téléphone.	Absence de détails
Antécédents médicaux : -Diabète. -**Allergie au DOLIPRANE.**	Antécédents médicaux : -Diabète.	Omission allergie

TABLEAU 8 – Autres exemples de résumés avant et après l'affinage sur les résumés sélectionnés par *QUARTZ*. **Gauche :** *QUARTZ* en zéro-shot génère plusieurs résumés et sélectionne les meilleurs. **Droite :** Un affinage supplémentaire est appliqué aux résumés sélectionnés.