Clarification des Ambiguïtés : Sur le Rôle des Types d'Ambiguïté dans les Méthodes d'Amorçage pour la Génération de Clarifications

Anfu Tang[1] Laure Soulier[1] Vincent Guigue[2]
(1) Sorbonne Université, CNRS, ISIR, F-75005 Paris, France
(2) AgroParisTech, UMR MIA-PS, Palaiseau, France
tang@isir.upmc.fr, laure.soulier@isir.upmc.fr,
vincent.guigue@agroparistech.fr

RÉSUMÉ

En recherche d'information (RI), il est essentiel de fournir des clarifications appropriées pour concevoir un système de dialogue proactif et guider l'utilisateur. Grâce au développement des grands modèles de langage (LLMs), des études récentes explorent des méthodes d'amorçage pour générer des clarifications à l'aide de chaîne de raisonnement (Chain of Thought, CoT). Cependant, l'amorçage CoT ne permet pas de distinguer les caractéristiques des différents besoins en information, impactant la résolution des ambiguïtés. Dans ce travail, nous cherchons à modéliser et intégrer les ambiguïtés liées au besoin en information dans le processus de génération de clarifications. Nous étudions l'impact des schémas d'amorçage en proposant Ambiguity Type-Chain of Thought (AT-CoT), qui impose à CoT de prédire d'abord les types d'ambiguïté, puis de générer les clarifications correspondantes. Des expériences sont menées sur divers jeux de données afin de comparer AT-CoT à plusieurs modèles de référence. Nous réalisons également des simulations utilisateur pour une évaluation extrinsèque.

ABSTRACT

Clarifying Ambiguities : On The Role of Ambiguity Types in Prompting Methods for Clarification Generation

In information retrieval, providing appropriate clarifications is crucial for building a proactive dialogue system. Due to the strong in-context learning ability of large language models (LLMs), recent studies investigate prompting methods to generate clarifications using Chain of Thought (CoT) prompts. However, vanilla CoT prompting does not distinguish the characteristics of different information needs, making it difficult to understand how LLMs resolve ambiguities. In this work, we seek to model and integrate ambiguities in the clarification process. To this end, we comprehensively study the impact of prompting schemes based on reasoning and ambiguity for clarification. We propose AMBIGUITY TYPE-CHAIN OF THOUGHT (AT-COT), which requires CoT to first predict ambiguity types, then generate clarifications correspondingly. Experiments are conducted on various datasets to compare AT-COT with multiple baselines. We also perform user simulations to implicitly measure the quality of generated clarifications.

MOTS-CLÉS : question de clarification, système de RI conversationnel, type d'ambiguïté.

KEYWORDS: clarifying question, dialogue search system, ambiguity type.

ARTICLE : **Accepté à** SIGIR.

1 Introduction

L'ambiguïté en recherche d'information (RI) est un facteur bien identifié, pouvant compromettre la qualité des documents récupérés. En effet, les utilisateurs formulent souvent des requêtes ambiguës ou vagues pour initier une recherche (Belkin *et al.*, 2003). Les raisons de cette ambiguïté peuvent varier (Belkin *et al.*, 2003; Morris *et al.*, 2008) : par exemple, préférence vers des requêtes courtes, incertitude quant aux besoins informationnels, ou encore à cause du phénomène du "mot sur le bout de la langue" (Arguello *et al.*, 2021). Afin de mieux comprendre les ambiguïtés sous-jacentes aux requêtes des utilisateurs, des études précédentes ont exploré les types d'ambiguïté (TAs) et proposé différentes taxonomies (Min *et al.*, 2020; Guo *et al.*, 2021; Clarke *et al.*, 2009) pour les classifier.

Pour guider les utilisateurs face à l'ambiguïté, une lignée de travaux s'intéresse à faciliter l'expression de leurs besoins sans compromettre leur expérience de recherche. C'est le cas notamment avec les travaux axés sur le développement des moteurs de recherche conversationnels proactifs (Radlinski & Craswell, 2017). Leur objectif est de prendre l'initiative de la discussion en fournissant des informations ou des suggestions pour améliorer la qualité des résultats de recherche. Plutôt que de recevoir passivement une liste de documents, les utilisateurs peuvent participer activement à la recherche en interagissant avec le système de dialogue proactif. Les premières études sur la clarification se sont concentrées sur la reformulation de requêtes (Carpineto & Romano, 2012; Kuzi *et al.*, 2016), en cherchant à fournir des suggestions utiles pouvant répondre aux besoins des utilisateurs, sans pour autant exploiter explicitement leur intention. Des travaux plus récents s'intéressent davantage aux questions de clarification (Aliannejadi *et al.*, 2019; Rahmani *et al.*, 2023), qui consistent à poser une question de clarification et à permettre à l'utilisateur d'y répondre librement. Les méthodes de génération de clarifications ont évolué avec le développement des grands modèles de langage (LLMs), passant des approches supervisées basées sur des données annotées manuellement (Guo *et al.*, 2021; Amplayo *et al.*, 2023), aux méthodes d'amorçage des LLM (Wang *et al.*, 2023; Zhang *et al.*, 2024). Parmi ces méthodes, la génération de chaînes de raisonnement (*Chain of Thought*, CoT)(Wei *et al.*, 2022) a démontré sa capacité à générer de meilleures questions de clarification (Deng *et al.*, 2023; Zhang *et al.*, 2024) par rapport aux amorçage sans raisonnement explicite. Cependant, les travaux existants utilisent principalement l'amorçage CoT pour générer librement un raisonnement, sans demander explicitement aux LLMs de distinguer les différents besoins informationnels.

Dans ce travail, nous pensons que comprendre et intégrer les ambiguïtés dans le raisonnement est essentiel au processus de clarification, car les humains peuvent avoir tendance à catégoriser d'abord le type d'ambiguïté avant de décider de la meilleure façon de clarifier une requête. Pour simuler la manière dont les humains traitent les requêtes ambiguës, nous analysons en premier lieu le concept d'ambiguïté sous l'angle des types d'ambiguïté, puis nous proposons de les intégrer comme contraintes de chaîne de raisonnement au sein d'un LLM pour la génération de clarifications. Pour ce faire, nous combinons les types d'ambiguïté avec l'amorçage CoT afin de construire le Ambiguity Type-Chain of Thought (AT-CoT), une méthode qui incite les LLMs à prédire d'abord les TAs sous-jacents, avant de générer les clarifications correspondantes. Afin d'assurer le bon fonctionnement de AT-CoT, nous dérivons une taxonomie des ATs fondées sur des actions à partir des études existantes. Chaque TA de notre taxonomie sert non seulement à aider les LLMs à comprendre les causes de l'ambiguïté, mais peut également être interprété comme une action permettant aux LLMs de générer des clarifications. La Figure 1 illustre AT-CoT. Pour une requête telle que *effet miroir dans une petite pièce*, notre méthode prédit d'abord deux types d'ambiguïté, puis exécute les actions correspondantes afin de générer des questions de clarification (QCs -ou CQ dans la figure 1 illustrant un exemple en

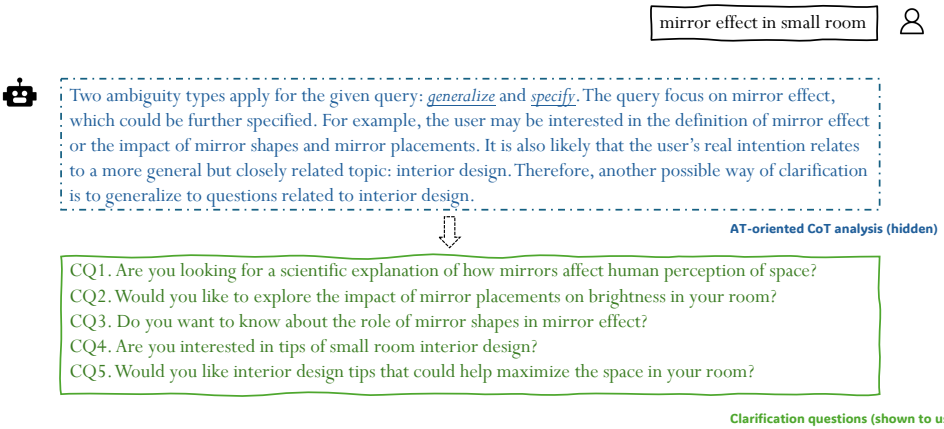

FIGURE 1 – Illustration de AT-CoT : Contrairement au CoT classique, le raisonnement généré par le LLM est limité à la prédiction des types d'ambiguïté possibles.

anglais). Par exemple, QC4 et QC5 généralisent la requête, tandis que QC1, QC2 et QC3 la spécifient.

Pour valider l'efficacité de notre méthode, des expériences sont menées sur des tâches à la fois intrinsèques et extrinsèques (respectivement la génération de clarifications et la recherche d'information) sur plusieurs jeux de données, notamment Qulac (Aliannejadi *et al.*, 2019), ClariQ (Aliannejadi *et al.*, 2021) pour la génération de clarification, ainsi que les collections de recherche d'information de TREC (Clarke *et al.*, 2009, 2010, 2011, 2012). Pour la tâche de recherche d'information, nous nous appuyons sur les travaux précédents (Aliannejadi *et al.*, 2019; Zou *et al.*, 2023; Erbacher *et al.*, 2024) et réalisons une simulation utilisateur pour générer la conversation de recherche menant à désambiguïser le besoin en information initial. Nous comparons différentes méthodes d'interaction pour la clarification, telles que la proposition de reformulations de requêtes que les utilisateurs peuvent sélectionner (*sélectionner*), ainsi qu'une interaction basée sur une unique question de clarification à laquelle l'utilisateur peut répondre en langage naturel (*répondre*).

En résumé, nos principales contributions sont les suivantes :

— Nous analysons les ambiguïtés à travers les types d'ambiguïté et examinons en profondeur l'effet de leur intégration avec le raisonnement dans les méthodes d'amorçage des LLMs pour la génération de clarification.
— Nous validons l'efficacité de notre méthode sur la génération de clarifications et les tâches de recherche d'information.

2 Travaux Connexes

Ambiguïté des requêtes utilisateur. Bien qu'il n'existe pas une unique taxonomie des ambiguïtés dans la communauté de la recherche d'information, les requêtes ambiguës ont été étudiées dans de nombreux travaux (Clarke *et al.*, 2009; Zamani *et al.*, 2020; Zhang *et al.*, 2024). L'état de l'art sur les types d'ambiguïté (TAs) peut être classé en trois catégories. Le premier groupe d'études établit une taxonomie des TAs en analysant les requêtes issues de jeux de données spécifiques (Clarke

et al., 2009; Guo *et al.*, 2021; Amplayo *et al.*, 2023; Min *et al.*, 2020). Cependant, les taxonomies proposées dans ces études ont avant tout une visée analytique et incluent des ATs très spécifiques. (par exemple, *EntityReferences* (Min *et al.*, 2020) et *CoreferenceResolution* (Guo *et al.*, 2021) qui correspondent tous deux à un type précis d'ambiguïté sémantique). Contrairement à ces approches, nous cherchons à formuler une taxonomie contenant des types d'ambiguïté mutuellement exclusifs. L'objectif est d'aider les LLMs à mieux générer des clarifications, plutôt que d'analyser en détail les causes d'ambiguïté des requêtes. Un deuxième groupe d'études se concentre sur les relations entre les requêtes en explorant des journaux de requêtes (Zamani *et al.*, 2020; Jansen *et al.*, 2009; Boldi *et al.*, 2011; Lau & Horvitz, 1999). Bien que ces travaux ne soient pas directement liés à la génération de clarifications, leurs résultats fournissent des indications précieuses sur les schémas de clarification. Par exemple, deux types courants de reformulation de requêtes observés (Jansen *et al.*, 2009; Boldi *et al.*, 2011) sont la généralisation et la spécialisation. Alors que la spécialisation est largement prise en compte dans les études précédentes, le besoin de généralisation est quant à lui moins étudié. Nous pensons que la généralisation peut également contribuer à mieux préciser les besoins informationnels des utilisateurs dans certains scénarios. Elle représente une dimension essentielle de l'ambiguïté, illustrant la possibilité qu'une requête utilisateur ne reflète pas fidèlement son intention. Le dernier groupe de travaux concerne la période post-LLM. Nous notons une étude récente (Zhang *et al.*, 2024) qui propose une taxonomie bien structurée mettant particulièrement l'accent sur les ambiguïtés propres aux LLMs, telles que les interprétations mal alignées des requêtes entre LLMs et humains. Notre approche diffère de la leur : leur taxonomie est principalement utilisée comme un outil d'évaluation, tandis que notre travail vise à intégrer les types d'ambiguïté dans le raisonnement pour la génération de clarifications.

En résumé, les travaux antérieurs exploitent majoritairement les types d'ambiguïté à des fins analytiques, sans chercher à améliorer les capacités de raisonnement des LLMs via l'intégration des TAs dans l'amorçage. Afin de mieux situer notre contribution par rapport aux études existantes, nous organisons les TAs issus des taxonomies précédentes et les présentons dans la Figure 2.

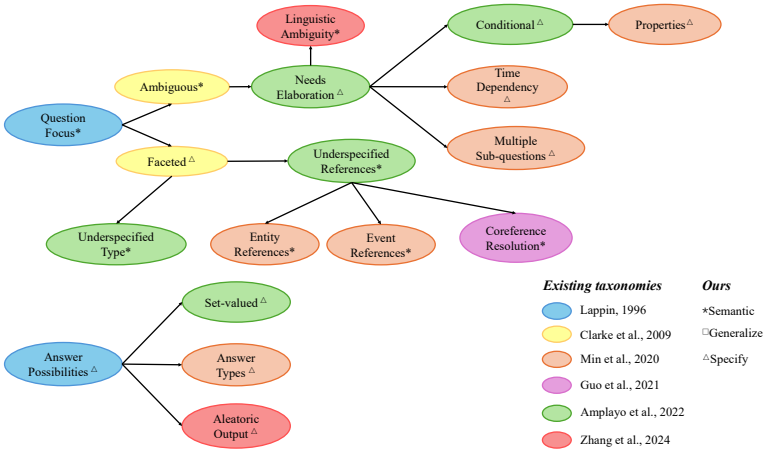

FIGURE 2 – Comparaison entre les taxonomies existantes à des fins analytiques et notre taxonomie basée sur l'action.

Clarification dans la recherche d'information La clarification a pour but d'expliciter le besoin informationnel de l'utilisateur (Zamani *et al.*, 2020). Les premières approches de génération de clarifications se sont concentrées sur l'expansion automatique des requêtes (Carpineto & Romano, 2012), où les requêtes initiales des utilisateurs sont réécrites ou enrichies (Chirita *et al.*, 2007; Cucerzan & White, 2007; Cao *et al.*, 2008; Mei *et al.*, 2008). Des travaux plus récents se concentrent davantage sur la génération de questions de clarification (Rao & Daumé III, 2018; Rahmani *et al.*, 2023; Aliannejadi *et al.*, 2019; Xu *et al.*, 2019; Lee *et al.*, 2023). Avant l'ère des LLMs, les méthodes de génération de clarifications reposaient principalement sur l'entraînement de modèles neuronaux séquence-à-séquence (e.g., Seq2Seq (Sutskever *et al.*, 2014)) sur des données annotées. Dans l'ère post-LLM, les études récentes se sont de plus en plus concentrées sur les méthodes d'amorçage des LLMs, notamment via l'amorçage en chaîne de raisonnement (CoT) (Deng *et al.*, 2023; Zhang *et al.*, 2024). Notre travail étend les études existantes sur l'amorçage des LLMs, en intégrant les types d'ambiguïté dans le raisonnement CoT.

Simulation de conversation dans la recherche d'information En RI conversationnelle, la simulation utilisateur consiste à créer des conversations artificielles entre le système de dialogue et l'utilisateur pour résoudre le besoin en information (Erbacher *et al.*, 2022; Câmara *et al.*, 2022; Maxwell *et al.*, 2015). La simulation est souvent utilisée pour tester automatiquement l'impact des systèmes de dialogue sur les performances de recherche, sans recourir à des tests utilisateurs réels (Eckert *et al.*, 1997; Dupret & Piwowarski, 2008; Chuklin *et al.*, 2013). Des hypothèses sur le comportement des utilisateurs sont formulées afin de contrôler les réponses des agents utilisateurs, en fonction de l'objectif de la simulation. Par exemple, pour évaluer la qualité de la reformulation dans les systèmes de recherche d'information, Erbacher *et al.* (2022) supposent que l'agent utilisateur est entièrement coopératif, choisissant ainsi toujours les reformulations les plus proches de son intention. Dans une autre étude (Erbacher *et al.*, 2024), les agents utilisateurs ne peuvent répondre que par "oui" ou "non", afin d'augmenter les jeux de données de RI avec des conversations multi-tours. Dans notre travail, en nous appuyant sur les recherches existantes, nous créons notre agent utilisateur à l'aide de LLMs et simulons les réponses des utilisateurs via une amorce basé sur des exemples.

3 Méthodologie

Nous présentons dans cette section les grandes lignes de notre méthodologie. Nous nous concentrons sur les questions de recherche suivantes :

QR1. Quelle est la taxonomie appropriée des ambiguïtés pour la génération de clarification (GC), compatible avec les méthodes d'amorçage des LLMs ?

QR2. Comment intégrer l'ambiguïté et le raisonnement dans les méthodes d'amorçage des LLMs pour la clarification ?

3.1 Taxonomie des types d'ambiguïté

Pour répondre à QR1, notre objectif est d'établir une taxonomie des types des ambiguïtés afin d'améliorer les capacités de raisonnement des LLMs dans le traitement des requêtes ambiguës. Des travaux précédents (Min *et al.*, 2020; Guo *et al.*, 2021; Amplayo *et al.*, 2023) ont proposé diverses taxonomies des types d'ambiguïté à des fins analytiques. Toutefois, dans l'optique d'aider les LLMs

à mieux comprendre les ambiguïtés et à générer des clarifications plus pertinentes, nous constatons que les taxonomies existantes sont redondantes et inadaptées aux méthodes d'amorçage des LLMs. Premièrement, comme l'a mis en évidence Zhang *et al.* (2024), la plupart des taxonomies existantes ont été proposées avant l'ère des LLMs ; certaines catégories d'ambiguïtés (TAs) manquent de définitions claires et ne sont pas mutuellement exclusives. Deuxièmement, les TAs des taxonomies existantes peuvent être réduits à deux types d'actions qu'un LLM peut effectuer : *Déterminer l'interprétation de la requête* ou *Préciser davantage la requête utilisateur*. En nous inspirant de Deng *et al.* (2023), qui ont proposé une approche d'amorçage proactif (où les LLMs prennent l'initiative de choisir les actions à effectuer plutôt que de simplement générer selon les amorces), nous proposons une taxonomie fondée sur les actions des LLMs, structurée en trois dimensions. Chacune d'entre elles correspond à un schéma de clarification identifié dans des travaux antérieurs (Clarke *et al.*, 2009; Zamani *et al.*, 2020; Jansen *et al.*, 2009) :

— *Sémantique* : traite l'ambiguïté liée à l'interprétation des requêtes.
— *Généraliser* : traite l'ambiguïté liée aux besoins informationnels lorsque les utilisateurs recherchent une information pertinente mais plus générale. Cette situation se produit lorsque les requêtes ne décrivent pas précisément l'intention réelle des utilisateurs.
— *Spécifier* : traite l'ambiguïté liée aux besoins informationnels lorsque les utilisateurs recherchent une information plus spécifique. Cette situation se produit lorsque les requêtes manquent de détails et couvrent un champ de recherche trop large.

La Table 1 présente une explication détaillée de notre taxonomie. Afin de faciliter la compréhension des liens entre notre taxonomie et les travaux précédents, nous regroupons chaque TA des taxonomies existantes dans notre cadre théorique, comme illustré dans la Figure 2.

TABLE 1 – Type d'ambiguïté basé sur l'action proposé.

Type d'ambiguïté	**Définition**	**TAs issus des études précédentes**
Sémantique	La requête est sémantiquement ambiguë pour plusieurs raisons courantes : elle peut contenir des homonymes ; un mot de la requête peut désigner une entité spécifique tout en ayant une signification générique ; ou encore, une entité mentionnée dans la requête peut renvoyer à plusieurs entités distinctes.	*Question Focus* (Lappin, 1996) *Linguistic Ambiguity* (Zhang *et al.*, 2024)
Généraliser	La requête se concentre sur une information spécifique ; cependant, une requête plus large mais étroitement liée pourrait mieux refléter le véritable besoin informationnel de l'utilisateur.	*Generalisation* (Jansen *et al.*, 2009; Boldi *et al.*, 2011)
Spécifier	La requête a un objectif clair, mais son champ de recherche peut être trop large. Il est possible de restreindre davantage ce champ en fournissant des informations plus spécifiques en lien avec la requête.	*Faceted* (Clarke *et al.*, 2009) *Time Dependency* (Min *et al.*, 2020) *Underspecified References* (Amplayo *et al.*, 2023)

3.2 Formulation des amorces pour la génération de clarifications

Cette section vise à répondre à QR2, c'est-à-dire comment intégrer les ambiguïtés, abstraites par la taxonomie des types d'ambiguïté présentée en Section 3.1, dans le raisonnement des méthodes

d'amorçage des LLMs pour la génération de clarifications. Nous cherchons à atteindre cet objectif en contraignant le raisonnement dans l'amorçage CoT. Pour cela, nous imposons aux LLMs de prédire les types d'ambiguïté (TAs) de notre taxonomie afin d'intégrer explicitement les ambiguïtés dans le raisonnement, ce qui leur permet de raisonner sur la base du besoin utilisateur en intégrant un processus de désambiguïsation et d'adopter des actions explicables. Nous formulons l'hypothèse que le raisonnement orienté vers l'ambiguïté est plus efficace que le raisonnement librement généré par les LLMs. Ainsi, nous proposons AMBIGUITY TYPE-CHAIN OF THOUGHT (AT-COT), une extension de l'amorçage CoT. Afin d'évaluer efficacement l'impact de l'intégration des ambiguïtés dans le raisonnement des LLMs, nous utilisons deux autres schémas d'amorçage comme points de comparaison : (1) L'amorçage STANDARD qui demande simplement aux LLMs de générer des clarifications sans étapes intermédiaires. (2) L'amorçage AT-STANDARD, qui intègre les définitions des TAs de notre taxonomie dans les amorces. Nous utilisons AT-STANDARD pour évaluer l'impact du simple fait d'informer les LLMs des types d'ambiguïté possibles, sans leur demander de générer un raisonnement explicite.

La Table 2 présente les instructions détaillées des prompts pour chaque schéma d'amorçage. D'un point de vue mathématique, chaque méthode d'amorçage est formulée comme suit :

— STANDARD (Deng *et al.*, 2023) : L'objectif de l'amorçage standard est de maximiser :

$$p(c|D, C, q) \tag{1}$$

où c désigne la clarification générée, q représente une requête ambiguë, C correspond à l'historique de conversation, et D désigne la description de la tâche.

— AT-STANDARD : La seule différence entre AT-STANDARD et STANDARD est que les définitions des TAs sont incluses dans l'amorçage :

$$p(c|D, A, C, q) \tag{2}$$

où A fait référence aux définitions des ATs de la Table 1.

— COT (*Chain of Thought*) (Wei *et al.*, 2022) : L'amorçage CoT exige que les LLMs génèrent un raisonnement textuel avant de produire des clarifications (raisonnement sans contraintes) :

$$p(a, c|D, C, q) \tag{3}$$

où a fait référence au raisonnement textuel généré.

— AT-COT : L'amorçage AT-CoT impose aux LLMs de prédire d'abord les TAs de notre taxonomie, puis de générer les clarifications correspondantes. L'objectif de l'amorçage AT-CoT est de maximiser :

$$p(a, c|D, A, C, q) \tag{4}$$

4 Protocole expérimental

Afin d'évaluer l'efficacité de AT-COT, nous réalisons des expériences sur trois types de tâches : (1) Génération de clarifications (GC) : nous utilisons des jeux de données contenant des questions de clarification (QCs) annotées manuellement et évaluons la performance en mesurant la similarité sémantique entre les QCs générées et celles annotées. (2) Recherche d'information (RI) : nous simulons des conversations multi-tours et transformons ces conversations en requêtes reformulées

TABLE 2 – Quatre schémas d'amorçage : standard, AT-standard, CoT et AT-CoT. <AT definitions> est un espace réservé pour les définitions des ATs présentées dans la Table 1.

Type d'amorçage	Instruction système
STANDARD	`Given a query in an information-seeking system, generate a clarifying question that you think is most appropriate to gain a better understanding of the user's intent.` `<query>`

AT-STANDARD	`Given a query in an information-seeking system, generate a clarifying question that you think is most appropriate to gain a better understanding of the user's intent.`
	`The ambiguity of a query can be multifaceted, and there are multiple possible ambiguity types:`
	`<AT definitions>`
	`Consider the above ambiguity types when generating.`
	`<query>`

COT	`Given a query in an information-seeking system, generate a clarifying question that you think is most appropriate to gain a better understanding of the user's intent.`
	`Before generating the clarifying question, provide a textual explanation of your reasoning about why the original query is ambiguous and how you plan to clarify it.`
	`<query>`

AT-COT	`Given a query in an information-seeking system, generate a clarifying question that you think is most appropriate to gain a better understanding of the user's intent.`
	`The ambiguity of a query can be multifaceted, and there are multiple possible ambiguity types:`
	`<AT definitions>`
	`Before generating the clarifying question, provide a textual explanation of your reasoning about which types of ambiguity apply to the given query. Based on these ambiguity types, describe how you plan to clarify the original query.`
	`<query>`

pour récupérer des documents pertinents. (3) GC+RI : nous alignons la performance en GC et en RI afin d'examiner la corrélation entre la qualité des clarifications générées et l'efficacité en RI, c'est-à-dire si de meilleures clarifications permettent d'améliorer les performances en IR.

4.1 Jeux de données

Dans cette section, nous présentons les jeux de données utilisés dans nos expériences. La Table 3 résume les statistiques des différents jeux de données.

4.1.1 Jeux de données pour GC

Qulac (Aliannejadi *et al.*, 2019) : Qulac utilise des requêtes issues du TREC Web Track 2009-2012. Les annotateurs identifient d'abord les facettes associées aux requêtes en analysant des extraits de résultats de recherche obtenus via un moteur de recherche, puis génèrent des QCs pour traiter ces facettes.

TABLE 3 – Statistiques des jeux de données utilisés dans nos expériences pour les trois tâches : GC, RI, GC+RI.

Jeu de données	# requêtes	# QCs	# intentions
Tâche 1 : GC			
Qulac	198	2575	-
ClariQ	298	3991	-
RaoCQ	500	2248	-
Tâche 2 : RI			
TREC Web Track 2009-2012	198	-	717
TREC Web Track 2013-2014	100	-	315
TREC DL Hard	50	-	350
Tâche 3 : GC+RI			
Qulac-TREC Web Track 2009-2012	198	2575	717

ClariQ (Aliannejadi *et al.*, 2021) : Similairement à Qulac, ClariQ est construit par crowdsourcing en annotant des QCs pour des requêtes fournies. Des niveaux d'ambiguïté allant de 1 à 4 sont également attribués dans ClariQ, où 4 représente des requêtes extrêmement ambiguës.

RaoCQ (Rao & Daumé III, 2018) : Un jeu de données spécifique à un domaine. Les annotateurs identifient les QCs pertinentes dans des paires (question, questions de clarification), où chaque question provient d'un post sur StackExchange, et les questions de clarification sont échantillonnées à partir des commentaires du même post. Comme dans (Rao & Daumé III, 2018), nous évaluons nos méthodes sur le sous-ensemble annoté manuellement.

4.1.2 Jeux de données pour RI

TREC Web Track 2009-2012 (Clarke *et al.*, 2009, 2010, 2011, 2012) : Un jeu de données de RI centré sur les requêtes de recherche web. Nous utilisons ClueWeb09 [1] (Catégorie B) comme collection de documents, qui contient 50 millions de pages web en anglais. Étant donné que les jugements de pertinence spécifiques aux facettes sont fournis dans les tâches de diversité du TREC Web Track, nous utilisons les facettes comme intentions utilisateur dans la simulation utilisateur.

TREC Web Track 2013-2014 (Collins-Thompson *et al.*, 2013, 2014) : Similairement au TREC Web Track 2009-2012, ce jeu de données contient des requêtes de recherche web multifacettes, mais avec des sujets plus ciblés, ce qui permet d'évaluer des requêtes plus complexes. Nous utilisons ClueWeb12 [2] comme collection de documents.

TREC DL Hard (Mackie *et al.*, 2021) : Un benchmark contenant des requêtes complexes issues de TREC DL 2019 & 2020 (Craswell *et al.*, 2019, 2020), qui peut nécessiter une clarification multi-tours pour lever des ambiguïtés implicites. Les requêtes de TREC DL Hard sont échantillonnées depuis MS Marco (Campos *et al.*, 2016). Nous utilisons la collection de documents de la tâche de classement de passages de MS Marco.

0. https://lucene.apache.org/
1. https://lemurproject.org/clueweb09.php/
2. https://lemurproject.org/clueweb12/

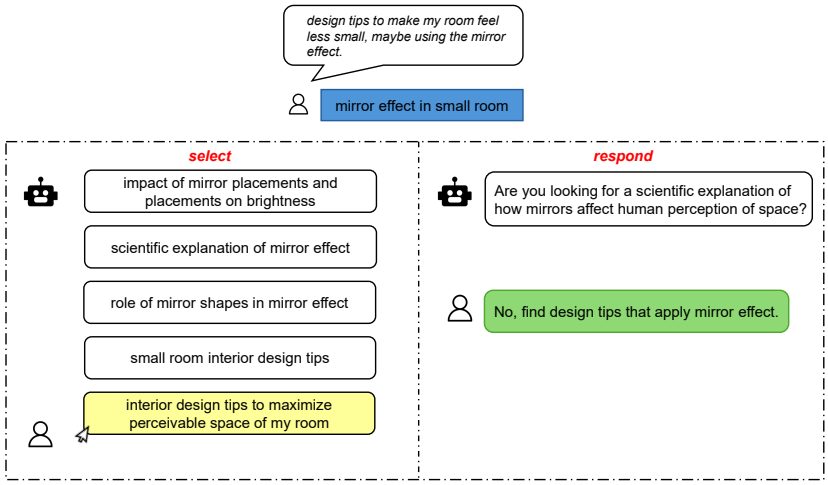

FIGURE 3 – Illustration des méthodes d'interaction pour la clarification (IC) dans notre simulation utilisateur : *sélectionner* et *répondre*.

4.1.3 Jeux de données pour GC+RI

Étant donné que Qulac est basé sur des requêtes du TREC Web Track 2009-2012, nous alignons les requêtes de Qulac avec les jugements de pertinence documentaire fournis par le TREC Web Track 2009-2012. Nous faisons référence à ce jeu de données sous le nom de **Qulac-TREC Web Track 2009-2012**, qui contient les questions de clarification annotées manuellement de Qulac ainsi que les jugements de pertinence documentaire du TREC Web Track 2009-2012.

4.2 Protocole d'évaluation

Génération de clarifications (GC). Pour chaque requête, nous générons plusieurs questions de clarification (QCs) afin d'évaluer de manière équitable les performances des différentes méthodes d'amorçage sur la tâche de génération de clarification. Plusieurs facteurs motivent cette décision : (1) Dans les jeux de données GC, chaque requête est associée à de nombreuses QCs annotées par des humains, couvrant différentes possibilités de clarification. Cependant, comme les métriques automatiques telles que BERTScore (Zhang* *et al.*, 2020) évaluent la similarité sémantique, il est possible qu'une QC de haute qualité obtienne un faible BERTScore simplement parce qu'elle ne correspond pas exactement aux QCs de référence. Pour atténuer ce problème, nous générons plusieurs QCs diversifiées, ce qui réduit la probabilité qu'aucune des QCs générées ne soit sémantiquement similaire à celles annotées. (2) Dans AT-CoT, une requête peut être associée à plusieurs TAs. Il est donc naturel de générer plusieurs QCs pour prendre en compte ces TAs distincts.

Simulation utilisateur pour RI Deux scénarios de clarification sont testés pour la simulation utilisateur (Figure 3) :

— *sélectionner* : À chaque tour, 5 requêtes reformulées sont générées. Nous avons adopté une température modérée de 0,6 afin de trouver un équilibre entre la diversité et la cohérence des

questions de recherche générées, en veillant à ce qu'elles soient variées sans faire preuve d'une créativité excessive. L'agent utilisateur sélectionne la requête correspondant le mieux à son intention, et la conversation se poursuit en fonction de la requête sélectionnée.

— *répondre* : À chaque tour, une QC est générée. L'agent utilisateur y répond en fonction de son intention.

Nous avons choisi ces deux scénarios d'interaction pour les raisons suivantes : (1) *sélectionner* a été largement étudié dans les travaux précédents (Cucerzan & White, 2007; Cao *et al.*, 2008) et est utilisé dans certains cas réels, tels que les suggestions des moteurs de recherche via la reformulation de requêtes. (2) *répondre* correspond à un cadre plus naturel pour la modélisation des interactions en recherche conversationnelle en langage naturel (Aliannejadi *et al.*, 2019; Lee *et al.*, 2023; Sekulić *et al.*, 2021).

Nous considérons également un modèle de référence sans clarification, où les requêtes originales sont utilisées sans clarification pour récupérer des documents. Pour simuler des conversations multi-tours, trois prompts sont enchaînés : *génération*, *réponse* et *reformulation*. L'amorce *génération* existe sous quatre variantes, correspondant aux méthodes d'amorçage décrites en Section 3.2. L'amorce *réponse* simule les réponses des utilisateurs et comporte deux variantes, chacune correspondant à un scénario d'interaction. Pour chaque conversation simulée, l'amorce *reformulation* prend la conversation en entrée et synthétise celle-ci sous forme de requête reformulée. Table 4 présente les amorces détaillées pour la *réponse* et la *reformulation*. L'objectif de l'amorçage par reformulation de requête est de faciliter l'évaluation des tâches de RI, car la plupart des modèles d'ordonnancement récupèrent des documents via des requêtes, et non via des conversations. Pour simplifier la simulation, nous supposons que les utilisateurs sont toujours coopératifs et que leurs intentions ne changent pas au cours de la conversation. Chaque conversation est initialisée par une requête utilisateur et simulée sur trois tours, sans règle d'arrêt intermédiaire. Les agents utilisateurs disposent toujours de descriptions complètes des intentions de l'utilisateur (par exemple, les facettes dans TREC Web Track).

4.3 Métrique d'évaluation

Conformément à Zhang *et al.* (2024), nous utilisons le BERTScore (Zhang* *et al.*, 2020) pour les tâches de GC, car les métriques basées sur la correspondance N-gram comme BLEU ou ROUGE ne permettent pas d'évaluer les capacités de clarification (Guo *et al.*, 2021).

Pour la tâche de RI, nous utilisons différentes métriques standards, conformément aux travaux précédents (Clarke *et al.*, 2009; Zhou *et al.*, 2024; Lin *et al.*, 2021) : nous utilisons nDCG@10 (Normalized Discounted Cumulative Gain) (Wang *et al.*, 2013) pour le TREC Web Track 2009-2014 ; MRR@10 (Mean Reciprocal Rank) (Radev *et al.*, 2002) pour TREC DL Hard.

4.4 Détails de l'implémentation

Modèle LLM. Nous utilisons Llama-3-8B (Dubey *et al.*, 2024) comme modèle de base, en chargeant les poids pré-entraînés depuis Huggingface. Les hyperparamètres du LLM sont fixés comme suit : $k = 10$ pour l'échantillonnage top-k, et température $t = 0.6$. Nous quantifions Llama-3 en NF4 (4-bit NormalFloat) et exécutons nos expériences sur un GPU TITAN Xp 12G.

TABLE 4 – L'amorçage de la *réponse* et *reformulation*. Il n'y pas besoin d'amorce *reformulation* pour la méthode d'interaction *sélectionner*.

Type d'amorçage	Instruction système
réponse (*sélectionner*)	`Imagine that you are a user seeking information with the help of a conversational assistant. At each turn of the conversation, the assistant provides you several reformulated queries to better understand your intention. Given a conversation history and a paragraph describing the user intent, choose the reformulation that most accurately reflects the provided user intent.` `<chat history>` `<user intent>`
réponse (*répondre*)	`Imagine that you are a user seeking information with the help of a conversational assistant. At each turn of the conversation, the assistant asks a clarification question to better understand your intention. Given a conversation history and a paragraph describing the user intent, respond to the clarification question based on the provided user intent.` `<chat history>` `<user intent>`
reformulation (*répondre*)	`Given a conversation history, summarize the conversation as a reformulated query. The chat history includes the initial query and several clarification turns between the user and a virtual assistant.` `<chat history>`

Analyse des sorties LLM. Nous demandons aux LLMs de produire des sorties structurées au format JSON. Les sorties des LLMs sont analysées via le parseur Pydantic de LangChain [3]. En cas d'erreur d'analyse, nous demandons aux LLMs de régénérer la sortie, avec un nombre maximal de tentatives fixé à 10. Dans de rares cas, nous analysons manuellement les sorties des LLMs pour traiter les erreurs de parsing persistantes.

BERTScore. Nous utilisons le troisième modèle pré-entraîné classé de BERTScore [4], basé sur les résultats expérimentaux [5] obtenus sur la tâche de traduction automatique WMT16 (Bojar *et al.*, 2016).

Pipeline RI. Conformément aux approches précédentes (Nogueira & Cho, 2019; Karpukhin *et al.*, 2020; Nogueira *et al.*, 2020), nous adoptons un ordonnancement en deux étapes (*"retriever-reranker"*) : Les documents les plus pertinents sont d'abord extraits d'une collection de documents à grande échelle en utilisant BM25. Ces documents sont ensuite réordonnancés avec MonoT5. Pour le BM25, nous utilisons l'implémentation Lucene de Pyserini sans ajustement des paramètres, avec $k = 100$. Pour MonoT5, nous utilisons un MonoT5 pré-entraîné (Nogueira *et al.*, 2020).

3. https://python.langchain.com/v0.1/docs/modules/model_io/output_parsers/types/pydantic/
4. https://huggingface.co/microsoft/deberta-large-mnli
5. https://github.com/Tiiiger/bert_score

TABLE 5 – Évaluation globale sur les jeux de données GC. *, †, $^\Delta$ indiquent des améliorations statistiquement significatives par rapport aux méthodes STANDARD, AT-STANDARD et CoT, respectivement, avec $p < 0.01$ selon un t-test.

Amorce	Qulac	ClariQ	RaoCQ
STANDARD	$77,9$	$79,3$	$60,0$
AT-STANDARD	$77,0$	$78,8$	$59,9$
CoT	$79,2^*$	$80,0$	$60,5$
AT-CoT	$\mathbf{80,6}^{*\dagger\Delta}$	$\mathbf{82,0}^{*\dagger\Delta}$	$\mathbf{62,4}^{*\dagger\Delta}$

5 Tâche 1 : Génération de clarification (GC)

Cette section vise à évaluer l'impact de l'intégration des ambiguïtés dans le raisonnement des LLMs sur la performance de la tâche de génération de clarifications (GC). La Table 5 présente une comparaison globale entre les différentes méthodes d'amorçage. Les résultats montrent que AT-CoT surpasse systématiquement les trois méthodes de référence avec des marges significatives sur tous les jeux de données. Par exemple, sur ClariQ, AT-CoT atteint un BERTScore de 82 contre des scores allant de $78,8$ à 80 pour les autres modèles de référence. Cela suggère que le raisonnement orienté vers l'ambiguïté permet de générer des questions de clarification plus pertinentes. Cette amélioration est cohérente à la fois sur des jeux de données spécifiques à un domaine (RaoCQ) et sur des jeux génériques (Qulac, ClariQ) ; ce qui montre que notre méthode généralise bien sur différents types de requêtes. En outre, en comparant AT-STANDARD et l'amorçage standard, nous constatons que seulement informer les LLMs des types d'ambiguïté n'est pas utile et peut même dégrader la performance (ex. 77 vs $77,9$ pour AT-STANDARD vs. STANDARD sur Qulac). Cela démontre que l'intégration des types d'ambiguïté est efficace uniquement lorsqu'elle est incorporée dans le raisonnement. Nos observations sur l'amorçage CoT sont cohérentes avec les travaux précédents (Deng *et al.*, 2023; Zhang *et al.*, 2024) : CoT est plus performant que l'amorçage standard. Nous allons encore plus loin en affirmant que le raisonnement orienté vers l'ambiguïté est encore plus efficace.

Stratification par niveau d'ambiguïté Nous évaluons plus en détail la performance des tâches GC en fonction des différents niveaux d'ambiguïté. Nous utilisons les étiquettes fournies dans ClariQ et présentons les résultats dans la Table 6. De manière générale, CoT et AT-CoT surpassent l'amorçage standard et AT-standard sur les trois premiers niveaux d'ambiguïté, ce qui démontre l'utilité du raisonnement librement généré par les LLM ainsi que du raisonnement orienté vers l'ambiguïté lorsque les requêtes ne sont pas extrêmement ambiguës. Cependant, en cas d'ambiguïté extrême (niveau 4), la performance de CoT diminue et devient inférieure à celle de l'amorçage standard (BERTScore de 78 vs. 78.5 et 79.7 pour respectivement standard et AT-standard). En revanche, AT-CoT maintient une amélioration constante, avec un BERTScore de 82.4. Cela suggère que les types d'ambiguïté jouent un rôle particulièrement important dans le traitement des requêtes ambiguës.

Distribution des types d'ambiguïté Afin d'approfondir l'analyse de AT-CoT, nous étudions la distribution des types d'ambiguïté prédits par AT-CoT. Nous commençons par examiner la fréquence des TAs prédits (*Sémantique*, *Généraliser* et *Spécifier*) avant d'évaluer l'impact de la prédiction de chaque TA sur la performance de la tâche GC. Les TAs prédits sont extraits du raisonnement

TABLE 6 – Résultats de GC sur ClariQ stratifiés par niveaux d'ambiguïté. *, †, $^{\Delta}$ indiquent des améliorations statistiquement significatives par rapport aux STANDARD, AT-STANDARD et COT, respectivement.

	level-1	level-2	level-3	level-4
STANDARD	$78,7$	$80,0$	$78,9$	$78,5$
AT-STANDARD	$77,6$	$79,2$	$78,4$	$79,7$
COT	$78,6$	$80,5$	$80,7^{\dagger}$	$78,0$
AT-COT	$\mathbf{80},\mathbf{9}^{*\dagger}$	$\mathbf{82},\mathbf{0}^{*\dagger\Delta}$	$\mathbf{82},\mathbf{1}^{*\dagger\Delta}$	$\mathbf{82},\mathbf{4}^{\dagger\Delta}$

TABLE 7 – Distribution des TAs prédits par AT-CoT. Entre parenthèses, nous indiquons les différences de performance en CG entre AT-CoT et CoT selon le BERTScore.

	Qulac	ClariQ	RaoCQ
Sémantique	$44,6$ ($\uparrow 1,3$)	$45,9$ ($\uparrow 1,8$)	$42,4$ ($\uparrow 2,0$)
Généraliser	$1,7$ ($\downarrow 0,6$)	$1,9$ ($\uparrow 1,4$)	$12,3$ ($\uparrow 2,0$)
Spécifier	$53,7$ ($\uparrow 1,4$)	$52,2$ ($\uparrow 2,0$)	$45,3$ ($\uparrow 1,9$)

généré par AT-CoT. La Table 7 présente des statistiques pour chaque groupe et recense la fréquence des requêtes identifiées comme appartenant à un type d'ambiguïté spécifique. La différence de performance en termes de BERTScore entre AT-CoT et CoT est indiquée entre parenthèses. Nos principales conclusions sont les suivantes : *Sémantique* et *Spécifier* sont les types d'ambiguïté les plus fréquents dans tous les jeux de données CG, avec une légère prédominance de *Spécifier* (jusqu'à 45.9% vs. jusqu'à 53.7% pour *Spécifier*). Cette observation est cohérente avec le fait que la majorité des ATs des taxonomies existantes peuvent être classés sous *Sémantique* ou *Spécifier*. Cependant, bien que moins fréquent, l'importance de *Généraliser* ne doit pas être sous-estimée. Le type *Généraliser* est plus marginal, mais le fait que 12% des requêtes de RaoCQ soient prédites comme généralisables justifie notre choix d'inclure *Généraliser* dans notre taxonomie. L'observation selon laquelle les requêtes dans RaoCQ nécessitent plus souvent une généralisation suggère que les prédictions des TAs par AT-CoT capturent effectivement les besoins de clarification des requêtes et sont moins susceptibles d'être aléatoires. Étant donné que les requêtes RaoCQ sont extraites de publications utilisateur sur StackExchange, elles sont généralement plus longues que celles de Qulac et ClariQ. Il est donc très probable qu'une requête RaoCQ ne décrive pas précisément l'intention utilisateur, nécessitant ainsi une généralisation. À l'inverse, les requêtes dans Qulac sont souvent courtes et utilisées pour la recherche web, ce qui les rend moins susceptibles de nécessiter une généralisation. L'écart observé entre la fréquence de prédiction de *Généraliser* et les améliorations de performance obtenues grâce à *Généraliser* montre que AT-CoT s'adapte bien aux jeux de données aux caractéristiques différentes.

6 Tâche 2 : Recherche d'information (RI)

Cette section vise à étudier l'impact de l'intégration des ambiguïtés dans le raisonnement des LLMs sur les performances en RI. La Table 8 présente les résultats en RI pour les deux modes d'interaction (*sélectionner* et *répondre*) ainsi que pour la méthode de référence sans clarification. Nous détaillons les résultats sur trois tours d'interaction successifs. De manière générale, nous observons la tendance suivante : AT-CoT $>$ CoT $>$ AT-standard $\approx$ standard. Par exemple, les clarifications obtenues avec AT-CoT permettent d'atteindre les meilleures performances RI pour

TABLE 8 – Résultats sur les ensembles de données IR basés sur la simulation utilisateur. Les scores sont exprimés en nDCG@10 pour TREC Web Track 2009-2012 et TREC Web Track 2013-2014, et en MRR@10 pour TREC DL Hard. *, †, $^\Delta$ indiquent des améliorations statistiquement significatives par rapport aux méthodes standard, AT-standard et CoT, respectivement, avec $p < 0.01$ selon un t-test.

	TREC Web Track 2009-2012		*TREC Web Track 2013-2014*		*TREC DL Hard*	
	sélectionner	*sélectionner*	*sélectionner*	*répondre*	*répondre*	*répondre*
sans clarification	$0,123$	$0,123$	$0,277$	$0,277$	$0,084$	$0,084$
Tour d'interaction-1						
STANDARD	$0,161$	$0,232$	$0,336$	$0,387$	$0,060$	$0,120$
AT-STANDARD	$0,165$	$0,230$	$0,337$	$0,383$	$0,066$	$0,113$
COT	$0,174^{*\dagger}$	$0,238$	$0,341$	$0,392^{\dagger}$	$0,063$	$0,123^{\dagger}$
AT-COT	$\mathbf{0,188}^{*\dagger\Delta}$	$\mathbf{0,244}^{*\dagger}$	$\mathbf{0,347}^{*\dagger}$	$\mathbf{0,397}^{\dagger}$	$\mathbf{0,074}^{*\Delta}$	$\mathbf{0,125}^{\dagger}$
Tour d'interaction-2						
STANDARD	$0,152$	$0,223$	$0,307$	$0,379$	$0,054$	$0,127$
AT-STANDARD	$0,149$	$0,228$	$0,291$	$0,376$	$0,052$	$0,151^{*}$
COT	$0,160^{*\dagger}$	$0,226$	$0,310^{\dagger}$	$0,384$	$0,062^{\dagger}$	$0,174^{*\dagger}$
AT-COT	$\mathbf{0,176}^{*\dagger\Delta}$	$\mathbf{0,233}$	$\mathbf{0,320}^{*\dagger\Delta}$	$\mathbf{0,391}^{*\dagger}$	$\mathbf{0,071}^{*\dagger\Delta}$	$\mathbf{0,184}^{*\dagger\Delta}$
Tour d'interaction-3						
STANDARD	$0,141$	$0,212$	$0,295$	$0,371$	$\mathbf{0,056}$	$0,141$
AT-STANDARD	$0,149$	$0,213$	$0,276$	$0,367$	$0,051$	$0,154$
COT	$0,148$	$\mathbf{0,216}$	$0,300^{\dagger}$	$0,373$	$0,054$	$0,184^{*\dagger}$
AT-COT	$\mathbf{0,152}$	$0,213$	$\mathbf{0,305}^{\dagger}$	$\mathbf{0,381}$	$0,052$	$\mathbf{0,188}^{*\dagger}$

TREC Web Track 2013-2014 sur tous les tours ($0,397$, $0,391$ et $0,381$ respectivement, contre $0,392$, $0,384$ et $0,373$ au mieux pour les méthodes de référence). Nous notons également que les performances RI sont systématiquement meilleures pour le mode d'interaction *répondre*, qui correspond à la génération de questions de clarification (contrairement au mode *sélectionner*, basé sur la reformulation de requêtes). Ces résultats mettent en évidence deux conclusions principales. 1) Ils confirment nos observations sur la performance GC, en montrant les bénéfices du raisonnement orienté vers l'ambiguïté pour la génération de clarification, tant d'un point de vue intrinsèque qu'extrinsèque. 2) Ils renforcent notre hypothèse selon laquelle les interactions de clarification basées sur l'ambiguïté et le raisonnement sont essentielles en RI. Nos résultats démontrent également la robustesse de notre méthodologie dans différents scénarios d'interaction. Pour les deux scénarios d'interaction, AT-CoT fournit systématiquement les meilleures performances RI. Cela implique que notre méthode est adaptable à divers scénarios réels.

Performance RI par tour de conversation. La Table 8 met en exergue un même schéma d'évolution des performances RI à travers plusieurs tours de conversation, quel que soit le schéma d'amorçage. Sous *sélectionner*, la performance atteint son maximum au premier tour, puis diminue progressivement. Pour TREC DL Hard sous *répondre*, la performance augmente régulièrement à mesure que la conversation progresse. Cette évolution des performances est cohérente avec la complexité des requêtes : TREC DL Hard contient des requêtes complexes issues des datasets TREC DL 2019/2020 (Craswell *et al.*, 2019, 2020). Ces requêtes sont relativement longues et plus difficiles à désambiguïser, nécessitant donc des conversations multi-tours. Cela se traduit par une hausse des scores au fil des tours de conversation. À l'inverse, dans les datasets TREC Web Track, la performance est optimale dès le premier tour. Les requêtes y sont moins ambiguës, ce qui rend la clarification

moins dépendante du nombre de tours de conversation. Cependant, quel que soit le tour spécifique, AT-CoT surpasse systématiquement les autres méthodes d'amorçage. Cela montre qu'il n'est pas nécessaire d'augmenter le nombre de tours pour observer les améliorations d'AT-CoT. Quelle que soit la durée d'interaction souhaitée par l'utilisateur, AT-CoT fournit de meilleures clarifications que les autres méthodes d'amorçage.

7 Tâche 3 : Alignement entre la génération de clarifications & la recherche d'information

Afin d'atténuer les biais potentiels introduits par la simulation utilisateur, nous cherchons à aligner davantage les performances de CG et IR en utilisant Qulac-TREC Web Track 2009-2012. Étant donné que les questions de clarification de référence dans Qulac sont uniquement fournies pour les requêtes initiales, nous utilisons les résultats CG et IR du premier tour sous le mode d'interaction *répondre*. Nous calculons le coefficient de corrélation de Pearson entre les résultats GC et RI, et obtenons : $r = 0.92$, $p = 0.08$, ce qui indique une forte corrélation positive. Nous supposons que l'absence de significativité statistique de cette corrélation pourrait être due à la complexité de la collection documentaire, qui ne permet pas de différencier suffisamment la qualité des clarifications. Une requête peut être raffinée par des QCs de haute qualité via la simulation utilisateur. Cependant, s'il manque des documents pertinents dans la collection pour prendre en compte ce raffinement, l'impact réel sur la performance de RI peut ne pas être reflété. Néanmoins, étant donné que nous obtenons un coefficient de corrélation supérieur à 0.9, cette observation ne remet pas en cause notre hypothèse : l'amélioration des performances RI apportée par AT-CoT est bien due à de meilleures clarifications.

8 Conclusion

Dans ce travail, nous avons exploré l'intégration des ambiguïtés et du raisonnement dans les méthodes d'amorçage des LLMs pour la clarification. Nous avons proposé une nouvelle taxonomie des types d'ambiguïté fondée sur l'action, et un nouveau schéma d'amorçage, AT-CoT. Les expériences menées sur les jeux de données de génération de clarification et de RI démontrent l'efficacité de notre méthodologie. De plus, nos analyses approfondies montrent que notre approche est robuste dans divers scénarios d'interaction de clarification et qu'elle s'adapte aux besoins de clarification de jeux de données aux caractéristiques variées. Cependant, notre étude présente certaines limitations. Nous avons défini une taxonomie des ATs comprenant trois types généraux, sans avoir étudié l'impact de l'intégration de types d'ambiguïté plus spécifiques ni si un raisonnement basé sur une taxonomie plus structurée serait avantageux. Nos expériences ont été réalisées avec Llama-3-8B, sans tester des modèles de différentes tailles. Une piste intéressante serait d'examiner comment les capacités de raisonnement orienté vers l'ambiguïté varient selon l'échelle du modèle.

Malgré ces limites, nous pensons que notre travail constitue une base solide pour mieux comprendre le rôle des types d'ambiguïté dans les méthodes d'amorçage des LLMs pour la clarification. Nous espérons que nos conclusions pourront fournir des pistes précieuses pour les recherches futures.

9 Remerciements

Ce travail a bénéficié du soutien de l'Agence nationale de la recherche (projet GUIDANCE, ANR-23-IAS1-0003) et du SCAI (Sorbonne Center for Artificial Intelligence).

Références

ALIANNEJADI M., KISELEVA J., CHUKLIN A., DALTON J. & BURTSEV M. (2021). Building and evaluating open-domain dialogue corpora with clarifying questions. In *EMNLP*.

ALIANNEJADI M., ZAMANI H., CRESTANI F. & CROFT W. B. (2019). Asking clarifying questions in open-domain information-seeking conversations. In *International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, SIGIR '19.

AMPLAYO R. K., WEBSTER K., COLLINS M., DAS D. & NARAYAN S. (2023). Query refinement prompts for closed-book long-form qa. In *Annual Meeting of the Association for Computational Linguistics*.

ARGUELLO J., FERGUSON A., FINE E., MITRA B., ZAMANI H. & DIAZ F. (2021). Tip of the tongue known-item retrieval : A case study in movie identification. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, p. 5–14.

BELKIN N. J., KELLY D., KIM G., KIM J.-Y., LEE H.-J., MURESAN G., TANG M.-C., YUAN X.-J. & COOL C. (2003). Query length in interactive information retrieval. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, p. 205–212, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/860435.860474.

BOJAR O., CHATTERJEE R., FEDERMANN C., GRAHAM Y., HADDOW B., HUCK M., YEPES A. J., KOEHN P., LOGACHEVA V., MONZ C. *et al.* (2016). Findings of the 2016 conference on machine translation (wmt16). In *First conference on machine translation*, p. 131–198 : Association for Computational Linguistics.

BOLDI P., BONCHI F., CASTILLO C. & VIGNA S. (2011). Query reformulation mining : models, patterns, and applications. *Inf. Retr.*, **14**(3), 257–289. DOI : 10.1007/s10791-010-9155-3.

CÂMARA A., MAXWELL D. & HAUFF C. (2022). Searching, learning, and subtopic ordering : A simulation-based analysis. In *Advances in Information Retrieval : 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I*, p. 142–156, Berlin, Heidelberg : Springer-Verlag. DOI : 10.1007/978-3-030-99736-6_10.

CAMPOS D. F., NGUYEN T., ROSENBERG M., SONG X., GAO J., TIWARY S., MAJUMDER R., DENG L. & MITRA B. (2016). Ms marco : A human generated machine reading comprehension dataset. *ArXiv*, **abs/1611.09268**.

CAO H., JIANG D., PEI J., HE Q., LIAO Z., CHEN E. & LI H. (2008). Context-aware query suggestion by mining click-through and session data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, p. 875–883, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/1401890.1401995.

CARPINETO C. & ROMANO G. (2012). A survey of automatic query expansion in information retrieval. *ACM Comput. Surv.*, **44**(1). DOI : 10.1145/2071389.2071390.

CHIRITA P. A., FIRAN C. S. & NEJDL W. (2007). Personalized query expansion for the web. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, p. 7–14, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/1277741.1277746.

CHUKLIN A., SERDYUKOV P. & DE RIJKE M. (2013). Modeling clicks beyond the first result page. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, p. 1217–1220.

CLARKE C. L., CRASWELL N. & SOBOROFF I. (2009). Overview of the trec 2009 web track. In *Trec*, volume 9, p. 20–29.

CLARKE C. L., CRASWELL N., SOBOROFF I. & VOORHEES E. M. (2011). Overview of the trec 2011 web track. In *TREC*.

CLARKE C. L. A., CRASWELL N., SOBOROFF I. & CORMACK G. V. (2010). Overview of the trec 2010 web track. In *Text Retrieval Conference*.

CLARKE C. L. A., CRASWELL N. & VOORHEES E. M. (2012). Overview of the trec 2012 web track. In *Text Retrieval Conference*.

COLLINS-THOMPSON K., BENNETT P. N., DIAZ F., CLARKE C. L. A. & VOORHEES E. M. (2013). Trec 2013 web track overview. In *Text Retrieval Conference*.

COLLINS-THOMPSON K., MACDONALD C., BENNETT P. N., DIAZ F. & VOORHEES E. M. (2014). Trec 2014 web track overview. In *TREC*, volume 13, p. 1–15.

CRASWELL N., MITRA B., YILMAZ E. & CAMPOS D. (2020). Overview of the trec 2020 deep learning track. In *TREC*.

CRASWELL N., MITRA B., YILMAZ E., CAMPOS D. & VOORHEES E. (2019). Overview of the trec 2019 deep learning track. In *TREC 2019*.

CUCERZAN S. & WHITE R. W. (2007). Query suggestion based on user landing pages. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, p. 875–876.

DENG Y., LIAO L., CHEN L., WANG H., LEI W. & CHUA T.-S. (2023). Prompting and evaluating large language models for proactive dialogues : Clarification, target-guided, and non-collaboration. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2023*, p. 10602–10621, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-emnlp.711.

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*.

DUPRET G. E. & PIWOWARSKI B. (2008). A user browsing model to predict search engine click data from past observations. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, p. 331–338.

ECKERT W., LEVIN E. & PIERACCINI R. (1997). User modeling for spoken dialogue system evaluation. In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, p. 80–87 : IEEE.

ERBACHER P., DENOYER L. & SOULIER L. (2022). Interactive query clarification and refinement via user simulation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 2420–2425.

ERBACHER P., NIE J.-Y., PREUX P. & SOULIER L. (2024). Augmenting ad-hoc IR dataset for interactive conversational search. *Transactions on Machine Learning Research*.

GUO M., ZHANG M., REDDY S. & ALIKHANI M. (2021). Abg-coQA : Clarifying ambiguity in conversational question answering. In *3rd Conference on Automated Knowledge Base Construction*.

JANSEN B. J., BOOTH D. L. & SPINK A. (2009). Patterns of query reformulation during web searching. *J. Am. Soc. Inf. Sci. Technol.*, **60**(7), 1358–1371.

KARPUKHIN V., OGUZ B., MIN S., LEWIS P., WU L., EDUNOV S., CHEN D. & YIH W.-T. (2020). Dense passage retrieval for open-domain question answering. In B. WEBBER, T. COHN, Y. HE & Y. LIU, Éds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 6769–6781, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.550.

KUZI S., SHTOK A. & KURLAND O. (2016). Query expansion using word embeddings. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, p. 1929–1932, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/2983323.2983876.

LAPPIN S., Éd. (1996). *The Handbook of Contemporary Semantic Theory*. Cambridge, Mass., USA : Blackwell Reference.

LAU T. & HORVITZ E. (1999). Patterns of search : analyzing and modeling web query refinement. In *Proceedings of the Seventh International Conference on User Modeling*, UM '99, p. 119–128, Berlin, Heidelberg : Springer-Verlag.

LEE D., KIM S., LEE M., LEE H., PARK J., LEE S.-W. & JUNG K. (2023). Asking clarification questions to handle ambiguity in open-domain QA. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2023*, p. 11526–11544, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-emnlp.772.

LIN J., MA X., LIN S.-C., YANG J.-H., PRADEEP R. & NOGUEIRA R. (2021). Pyserini : A python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 2356–2362.

MACKIE I., DALTON J. & YATES A. (2021). How deep is your learning : The dl-hard annotated deep learning dataset. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 2335–2341.

MAXWELL D., AZZOPARDI L., JÄRVELIN K. & KESKUSTALO H. (2015). Searching and stopping : An analysis of stopping rules and strategies. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, p. 313–322.

MEI Q., ZHOU D. & CHURCH K. (2008). Query suggestion using hitting time. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, p. 469–478, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/1458082.1458145.

MIN S., MICHAEL J., HAJISHIRZI H. & ZETTLEMOYER L. (2020). AmbigQA : Answering ambiguous open-domain questions. In B. WEBBER, T. COHN, Y. HE & Y. LIU, Éds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 5783–5797, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.466.

MORRIS D., RINGEL MORRIS M. & VENOLIA G. (2008). Searchbar : a search-centric web history for task resumption and information re-finding. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, p. 1207–1216, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/1357054.1357242.

NOGUEIRA R. & CHO K. (2019). Passage re-ranking with bert. *arXiv preprint arXiv :1901.04085*.

NOGUEIRA R., JIANG Z., PRADEEP R. & LIN J. (2020). Document ranking with a pretrained sequence-to-sequence model. In T. COHN, Y. HE & Y. LIU, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2020*, p. 708–718, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.findings-emnlp.63.

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).

RADLINSKI F. & CRASWELL N. (2017). A theoretical framework for conversational search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*,

CHIIR '17, p. 117–126, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3020165.3020183.

RAHMANI H. A., WANG X., FENG Y., ZHANG Q., YILMAZ E. & LIPANI A. (2023). A survey on asking clarification questions datasets in conversational systems. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 2698–2716, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.152.

RAO S. & DAUMÉ III H. (2018). Learning to ask good questions : Ranking clarification questions using neural expected value of perfect information. In I. GUREVYCH & Y. MIYAO, Éds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 2737–2746, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/P18-1255.

SEKULIĆ I., ALIANNEJADI M. & CRESTANI F. (2021). Towards facet-driven generation of clarifying questions for conversational search. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '21, p. 167–175, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3471158.3472257.

SUTSKEVER I., VINYALS O. & LE Q. V. (2014). Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, p. 3104–3112, Cambridge, MA, USA : MIT Press.

WANG Y., WANG L., LI Y., HE D. & LIU T.-Y. (2013). A theoretical analysis of ndcg type ranking measures. In *Annual Conference Computational Learning Theory*.

WANG Z., TU Y., ROSSET C., CRASWELL N., WU M. & AI Q. (2023). Zero-shot clarifying question generation for conversational search. In *Proceedings of the ACM Web Conference 2023*, WWW '23, p. 3288–3298, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3543507.3583420.

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 : Curran Associates, Inc.

XU J., WANG Y., TANG D., DUAN N., YANG P., ZENG Q., ZHOU M. & SUN X. (2019). Asking clarification questions in knowledge-based question answering. In K. INUI, J. JIANG, V. NG & X. WAN, Éds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 1618–1629, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1172.

ZAMANI H., DUMAIS S., CRASWELL N., BENNETT P. & LUECK G. (2020). Generating clarifying questions for information retrieval. In *Proceedings of the web conference 2020*, p. 418–428.

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

ZHANG T., QIN P., DENG Y., HUANG C., LEI W., LIU J., JIN D., LIANG H. & CHUA T.-S. (2024). CLAMBER : A benchmark of identifying and clarifying ambiguous information needs in large language models. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 10746–10766, Bangkok, Thailand : Association for Computational Linguistics. DOI : 10.18653/v1/2024.acl-long.578.

ZHOU Y., YAO J., DOU Z., TU Y., WU L., CHUA T.-S. & WEN J.-R. (2024). Roger : Ranking-oriented generative retrieval. *ACM Trans. Inf. Syst.* Just Accepted, DOI : 10.1145/3603167.

ZOU J., ALIANNEJADI M., KANOULAS E., PERA M. S. & LIU Y. (2023). Users meet clarifying questions : Toward a better understanding of user interactions for search clarification. *ACM Trans. Inf. Syst.*, **41**(1). DOI : 10.1145/3524110.