



City Research Online

City St George's, University of London

Citation: Saadaoui, S. & Alonso, E. (2025). Coordinated LLM Multi-Agent Systems for Collaborative Question-Answer Generation. Knowledge-Based Systems, 330(Part B), 114627. doi: 10.1016/j.knosys.2025.114627

This is the accepted version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: <https://openaccess.city.ac.uk/id/eprint/35956/>

Link to published version: <https://doi.org/10.1016/j.knosys.2025.114627>

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to [City Research Online policy](#).

Knowledge-Based Systems

Coordinated LLM Multi-Agent Systems for Collaborative Question-Answer Generation

--Manuscript Draft--

| | |
|------------------------------|--|
| Manuscript Number: | KNOSYS-D-25-06122 |
| Article Type: | Full Length Article |
| Keywords: | Question-answer generation; Data augmentation; Large language models; Multi-agent coordination; Multi-perspective analysis |
| Corresponding Author: | Sami Saadaoui City St George's University of London London, UNITED KINGDOM |
| First Author: | Sami Saadaoui |
| Order of Authors: | Sami Saadaoui Eduardo Alonso, Professor |
| Abstract: | <p>Large Language Models (LLMs) excel at generating coherent and human-like questions and answers (QAs) across various topics, which can be utilized in various applications. However, their performance may be limited in domain-specific knowledge outside their training data, potentially resulting in low context recall or factual inconsistencies. This is particularly true in highly technical or specialized domains that require deep comprehension and reasoning beyond surface-level content.</p> <p>To address this, we propose Collective Intentional Reading through Reflection and Refinement (CIR3), a novel multi-agent framework that leverages collective intelligence for high-quality Question-Answer Generation (QAG) from domain-specific documents. CIR3 employs a transactive reasoning mechanism to facilitate efficient communication and information flow among agents. This enables an in-depth document analysis and the generation of comprehensive and faithful QAs. Additionally, multi-perspective assessment ensures that QAs are evaluated from various viewpoints, enhancing their quality and relevance. A balanced collective convergence process is employed to ensure that the agents reach a consensus on the generated QAs, preventing inconsistencies and improving overall coherence.</p> <p>Our experiments indicate a substantial level of alignment between the CIR3-generated QAs and corresponding documents, while improving comprehensiveness by 21% and faithfulness by 17% compared to strong baseline approaches. Code and data are available at https://github.com/anonym-nlp-ai/cirrr.</p> |

Coordinated LLM Multi-Agent Systems for Collaborative Question-Answer Generation

Sami Saadaoui^a, Eduardo Alonso^a

^a*Artificial Intelligence Research Centre (CitAI), City St George's, University of London, Northampton Square, London, EC1V 0HB, UK*

Abstract

Large Language Models (LLMs) excel at generating coherent and human-like questions and answers (QAs) across various topics, which can be utilized in various applications. However, their performance may be limited in domain-specific knowledge outside their training data, potentially resulting in low context recall or factual inconsistencies. This is particularly true in highly technical or specialized domains that require deep comprehension and reasoning beyond surface-level content. To address this, we propose **Collective Intentional Reading through Reflection and Refinement (CIR3)**, a novel multi-agent framework that leverages collective intelligence for high-quality Question-Answer Generation (QAG) from domain-specific documents. CIR3 employs a transactive reasoning mechanism to facilitate efficient communication and information flow among agents. This enables an in-depth document analysis and the generation of comprehensive and faithful QAs. Additionally, multi-perspective assessment ensures that QAs are evaluated from various viewpoints, enhancing their quality and relevance. A balanced collective convergence process is employed to ensure that the agents reach a consensus on the generated QAs, preventing inconsistencies and improving overall coherence. Our experiments indicate a substantial level of alignment between the CIR3-generated QAs and corresponding documents, while improving comprehensiveness by 21% and faithfulness by 17% compared to strong baseline approaches. Code and data are available at <https://github.com/anonym-nlp-ai/cirrr>.

Keywords: Question-answer generation, Data augmentation, Large language models, Multi-agent coordination, Multi-perspective analysis, Domain-specific

1. Introduction

Question-Answer Generation (QAG) is a data augmentation task that consists of generating a set of QA pairs given a context. QAG has a variety of applications, from information retrieval [1, 2, 3] to healthcare [4, 5], and education [6, 7, 8]. Although Question Generation (QG) has been extensively researched in the context of language models [9, 10], QAG presents a more challenging task, as it requires generating both the question and the answer, rather than assuming that the answer is already provided in the input, as illustrated in Example 1. While QG models offer a more direct and focused approach, they primarily focus on surface-level features of the context, such as facts and keywords. This is due to the limited amount of explicit information that is conditioned on the input answer. Furthermore, despite the proposal of various methods, generating comprehensive and semantically distinct questions from the same context remains under-explored as highlighted in [11, 12]. The latter attributes this limitation to the lack of multi-reference training datasets that exhaustively cover all possible questions for each context. This inability is even more evident in highly technical or specialized domains, where documents are often rich in information.

Illustrative Example: QG vs. QAG

Context: *"A defined contribution pension plan is one where the final benefit depends on the contributions made and the performance of the selected investments."*

QG Output:

- How does a defined contribution plan work?
- What determines the final benefit in a defined contribution plan?

QAG Output:

- **Q:** How does a defined contribution plan work?
A: It depends on the contributions and investment performance.
- **Q:** What determines the final benefit in a defined contribution plan?
A: The final benefit depends on contributions and investment returns.

In-Context Learning (ICL) [13] is an emerging paradigm that enables LLMs to learn new tasks without the need for extensive fine-tuning on specific data. By providing a description of the task, along with a few or even zero demonstrations as part of the input context, LLMs can be conditioned to perform well in various domains. This approach has shown promising results,

surpassing state-of-the-art models in some tasks, and offers a potential solution to the challenge of limited data availability [14, 15]. Despite impressive results on popular NLP benchmarks, we find that using ICL for QAG, given a relatively complex document, often lacks robust inference mechanisms to deduce implicit relationships between the different key points inherent in the context. If the generation depends on comprehending the underlying connections that are not explicitly stated in the context, the model may fail to generate faithful QAs that accurately reflect this complexity. This is particularly problematic for information-dense contexts, which are common in highly domain-specific corpora, such as finance and health.

Recent advancements in LLM-based Multi-Agent¹ (LLM-MA) systems have shown significant improvements in problem-solving abilities through planning, collaboration, and autonomous task execution [18, 19]. These systems break down complex tasks into simpler subtasks to enhance complex task solving. Compared to standard LLMs and single-agent setups, LLM-MA systems offer advanced capabilities by leveraging collective intelligence and specialized skills [20]. Motivated by the potential of these capabilities, we augment the QAG task with collective reasoning through the adoption of LLM-MA settings.

In order to address the aforementioned limitations in relation to generating comprehensive and faithful QAs from highly domain-specific documents, we derive a list of research questions around the adoption of LLM agents for QAG tasks:

- R.1** : Can an LLM-MA workflow uncover deeper and perhaps implicit key concepts from a complex and information-dense document?
- R.2** : How can LLM-MA effectively emphasize deep engagement, with a text, from different viewpoints to enable comprehensive and consistent generation and mitigate blind spots?

¹LLM-based agents are autonomous systems that leverage LLMs as their core reasoning and decision-making engine. These agents can perceive their environment through natural language, process information, generate plans, and take actions to achieve specific goals. Unlike traditional AI systems with static functionalities, LLM-based agents exhibit a degree of general intelligence, enabling them to handle a wider range of tasks and adapt to novel situations based on their extensive knowledge and language understanding capabilities [16, 17].

R.3 : (a) How can we incentivize multiple agents to seek consensus? (b) How can we control the process of convergence to reach common QAG, while avoiding premature collapse to incomprehensive and/or unfaithful generation?

To address these research questions, we design Collective Intentional Reading through Reflection and Refinement (**CIR3**) based on three corresponding hypotheses:

H.1 : Transactive reasoning² allows the deduction of QAs that uncover the implicit relationships between key concepts within the text.

H.2 : Multi-perspective group debate leads to an in-depth analysis of the document.

H.3 : Collective convergence, the process of a group of agents moving towards a shared output, requires disruptive signals to ensure diversity is maintained and collapse is avoided.

To build upon these hypotheses, CIR3 first utilizes an optimized topology of information within the agents to maximize the effectiveness of collaborative problem-solving and ensure an in-depth analysis of the input document. Second, CIR3 gains effectiveness by dynamically allocating specialized *writer* agents, each with a distinct perspective, based on the topic categories identified within the input context. Third, to reach a shared understanding of the document, despite the diverse perspectives and reasoning capabilities of the writers, CIR3 employs a *curmudgeon* agent as a mechanism for introducing variation. The curmudgeon, coupled with an external evaluation tool, incites the writers towards a balanced collective convergence on the key concepts within the text while maintaining diversity in the generated QAs.

While lexical matching is a standard evaluation method for QA tasks, its limitations become apparent when dealing with generative models, which often produce plausible answers not found in the predefined gold standard. This issue is further compounded by LLMs generating increasingly complex

²In this paper, we mimic the concept of transactive reasoning [21, 22], a cognitive process that occurs through social interaction, where individuals build upon each other’s ideas to create new knowledge or solve problems. It involves a dynamic exchange of thoughts, critiques, and elaborations, leading to a deeper understanding of a topic.

and lengthy answers, making lexical matching even less effective [23]. To ensure a comprehensive and accurate evaluation of CIR3, we employ diverse automatic metrics, in addition to human evaluation.

To summarize, our main contributions include: (1) We shift our focus from the typical task of QG to the more challenging QAG, which is inherently more difficult due to the limited search space and increased risk of producing duplicate QA pairs; (2) To the best of our knowledge, CIR3 is the first proposed QAG approach using multi-agent LLMs; (3) Our research demonstrates that incorporating an external signal significantly improves both the convergence rate and diversity maintenance within a group of agents. Specifically, in the context of QAG, this translates to optimizing the number of iterations needed for the agents to reach collective agreement on the identified QAs pairs that comprehensively and faithfully cover the key concepts of the input text; (4) To improve alignment with human evaluation, we develop a custom metric approach leveraging Encoder and LLM-based scores over both individual and concatenated QA pairs; (5) We demonstrate that CIR3 achieves improvements over strong baselines.

2. Related Work

In this section, we briefly review relevant work in the areas of Question-Answer Generation and LLM-based Multi-Agent Systems.

2.1. Question-Answer Generation

Both rule-based [24, 25, 26] and neural [14, 27, 28] models have been extensively used for QG from text documents. Similarly, machine reading comprehension [29, 30, 31] has been employed for answer extraction (AE) from text given a question. However, traditional QG and AE methods produce either the question or the answer, unlike QAG which outputs both.

Several studies have leveraged pre-trained language models for QAG. These include fine-tuning BERT [32] for AE and QG [33], fine-tuning autoregressive LMs for QG [34] using BART [35] and RoBERTa [36] for AE, jointly fine-tuning LMs for AE and QG [37], and using QAG models to generate adversarial examples [38]. Recent advancements have also focused on dynamically identifying question-worthy context words before using them to condition subsequent question generation [12]. [39] improved QAG by designing three distinct approaches: Pipeline, Multitask, and End2end. [40]

proposed combining entity linkage with a QA system, while [41] enriched QA extraction by augmenting it with entity-level metadata.

Despite these advancements, current research predominantly focuses on specific question types, such as *Wh*-questions, rather than addressing open-ended questions. Furthermore, the focus tends to be on extracting short answers from existing text rather than generating comprehensive and detailed responses. Additionally, evaluating comprehensiveness of QAG remains an under-explored area. We tackle these challenges using CIR3.

2.2. LLM-based Multi-Agent Systems

Recent research has focused on LLM-based multi-agent systems to improve the quality of complex reasoning tasks. Studies like [42, 43, 44] have shown that collaboration and task division among multiple agents can reduce hallucinations and generate more reliable outputs. Other works, such as [45, 46], highlight the benefits of continuous debate among agents to correct misconceptions, analyze problems from diverse perspectives, and ultimately achieve higher-quality results.

Furthermore, prior research [47] has examined the issue of inter-consistency through inter-agent negotiation. Similarly, drawing inspiration from robotics, [48] investigated consensus-seeking in multi-robot collaboration by analyzing the effects of agent number, personality, and network topology. However, their work specifically focused on agent behavior within a 1D-space. Conversely, [49] explored the concept of *flocking* where agents maintain proximity while avoiding collisions and preserving formations. In this work, we improve QAG task through a balanced collective convergence process.

3. Method

Given a context c consisting of a text passage, the task of QAG aims to produce a set of QA pairs, denoted as $\mathcal{G} = \{(q_i, a_i)\}_{i=1}^N$, that satisfies two crucial properties:

1. **Comprehensiveness:** The set \mathcal{G} should cover all the key points and essential information present in the context c . In other words, for every significant aspect or piece of information $x \in c$, there exists at least one QA pair $(q_i, a_i) \in \mathcal{G}$ such that q_i elicits and a_i provides information relevant to x .

2. **Faithfulness:** Each answer a_i in \mathcal{G} must be grounded in and supported by the factual content of the context c . This constraint ensures that the generated answers are not fabricated or hallucinatory, but rather reflect accurate information derived from the given text.

Formally, the QAG task can be formulated as an optimization problem, where the objective is to find the set \mathcal{G} that maximizes both comprehensiveness and faithfulness with respect to the context c . This can be expressed as: $\mathcal{G}^* = \arg \max_{\mathcal{G}} [\text{Comp}(\mathcal{G}, c) + \text{Faith}(\mathcal{G}, c)]$, where $\text{Comp}(\mathcal{G}, c)$ and $\text{Faith}(\mathcal{G}, c)$ are scoring functions that assess the extent to which the set \mathcal{G} covers the key points of c and adheres to the factual content of c , respectively. These scoring functions are defined in terms of diversity measures as follows:

$$\mathcal{G}^* = \arg \max_{\mathcal{G}} \left[\underbrace{\frac{\alpha_{q,a}}{2} \cdot (D_q + D_a)}_{\text{Comp}(\mathcal{G},c)} + \underbrace{\alpha_{a,c} \cdot (1 - D_{a,c})}_{\text{Faith}(\mathcal{G},c)} \right] \quad (1)$$

where \mathcal{D}_q and \mathcal{D}_a denote diversity scores computed over the sets of generated questions $\{q_i\}_{i=1}^N$ and answers $\{a_i\}_{i=1}^N$, respectively, and $\mathcal{D}_{a,c}$ denotes a dissimilarity measure between the concatenated answers $a_1 \oplus \dots \oplus a_N$ and the context c . $\mathcal{D} \in [1, 2] \subset \mathbb{R}$, where $\mathcal{D} = 1$ denotes perfect similarity. The coefficients³ $\alpha_{q,a}$ and $\alpha_{a,c}$, where $\alpha_{q,a} + \alpha_{a,c} = 1$, control the relative weighting of question and answer diversity (Comprehensiveness) and the alignment of answers with the context (Faithfulness) in the overall score.

In what follows, we describe CIR3 to generate the optimal solution \mathcal{G}^* given c . This is achieved by building upon the aforementioned hypotheses to ensure that QAG is based on an in-depth analysis of the input text through an efficient flow of information and adoption of multiple views approach (3.1, 3.2), while maintaining QAG diversity and optimizing the convergence rate of agents (3.3). The pseudo-code of the algorithm serving as the conceptual foundation of our approach is outlined in Algorithm 1.

³In this study, the coefficients $\alpha_{q,a}$ and $\alpha_{a,c}$ are empirically assigned equal weights (0.5). Although this choice effectively demonstrates our framework’s capabilities, future research will explore dynamic estimation of α -values, potentially leveraging neural networks or other adaptive techniques, to further optimize Comprehensiveness and Faithfulness.

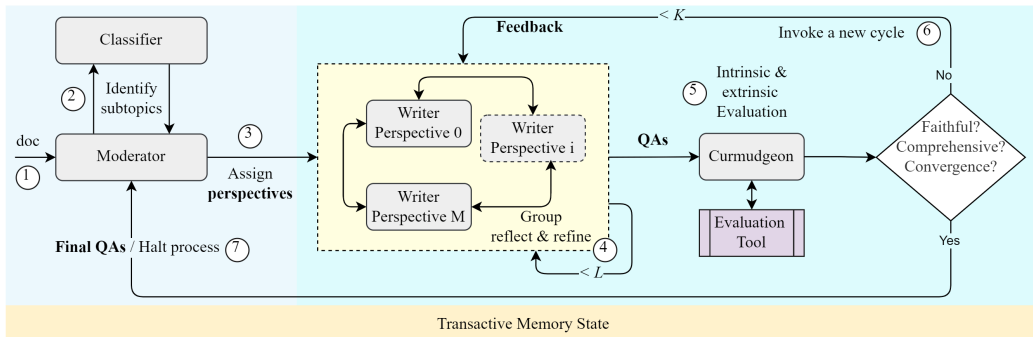


Figure 1: CIR3 takes an input document (1), identifies subtopics (2), and prompts writer agents to generate QA pairs based on their assigned perspectives (subtopics) (3). The QAs undergo iterative refinement by the writers (4), followed by an outer refinement where the curmudgeon, using its intrinsic knowledge and/or the evaluation tool, analyses the QAs and provides feedback for the next cycle (5, 6). The process halts when the curmudgeon is satisfied, and CIR3 returns the final QAs (7). The transactive memory serves as a central knowledge repository.

3.1. Multi-Perspective Analysis

Incorporating multi-perspective or various viewpoints is crucial for analyzing complex documents as it enhances the depth and breadth of understanding. Existing research highlights that a single perspective may introduce bias or overlook crucial aspects [50]. For instance, STORM [46] emphasizes the value of multiple perspectives in writing Wikipedia-like articles, by guiding participants to ask more in-depth questions in the pre-writing stage. Similarly, [51] showcased how addressing various perspectives improved document clarity and readability in document revision task.

While STORM efficiently identifies different perspectives by surveying existing articles from similar topics using a search engine, CIR3 challenge is to discover diverse perspectives from a contained and limited context without retrieving external information. Given the input context c , CIR3 leverages LLM’s language understanding capabilities to identify different subtopics within the input document c . To this end, we, first, utilize few-shot prompting, with a limited set of demonstrations, to guide a *classifier* agent to classify the context into M specific categories $P = \{p_1, \dots, p_M\}$ (Figure 1 ①-②)⁴.

⁴For example, given a finance-related document, CIR3 is prompted to discover the different M subtopics present in the context, such as *pensions*, *insurance*, and *savings*.

Next, the *moderator* agent dynamically assigns each identified perspective p_j to a different writer W_{p_j} , while prompting the agents to analyze the input context and generate a set of QA pairs, \mathcal{G}^{p_j} , based on their respective perspectives (Figure 1 ③). Subsequently, as per 3.2 and 3.3, the list of \mathcal{G}^{p_j}

Algorithm 1: CIR3

Input : Max inner-refinement cycles L ;
Max outer-refinement cycles K ;
Max perspective M , Context c

Output: QA pairs \mathcal{G}^*

```

1  $\mathcal{M} \leftarrow [ ]$  // Writer's short memory state.
2  $\mathcal{H} \leftarrow [ ]$  // Long short-term memory state.
3 // Identify and assign unique perspectives  $\mathcal{P}$ .
4  $W_{P_0} = \text{"default in-domain writer"}$ 
5  $W \leftarrow [W_{P_0}]$  // List of Writers.
6  $P \leftarrow \text{classify\_subtopics}(c, M)$  // List of subtopics  $\leq M$ .
7 foreach subtopic in  $P$  do
8   |  $W.append(\text{get\_perspective\_writer}(\text{subtopic}))$ 
9 end
10 // Outer-refinement cycles: evaluate and critic
11  $k \leftarrow 0$ 
12  $\mathcal{F}_{k+1} \leftarrow \emptyset$ 
13 do
14   | // Inner-refinement cycles: generate, then refine.
15   |  $l \leftarrow 0$ 
16   |  $\mathcal{F}_{l+1} \leftarrow \emptyset$ 
17   | do
18   |   |  $\mathcal{G}^+_l \leftarrow \text{generateQAs}(c, \mathcal{M}[-1], \mathcal{H}[-1])$ 
19   |   |  $\mathcal{F}_{l+1} \leftarrow \text{refineQAs}((\mathcal{G}^+_l, \mathcal{M}))$ 
20   |   |  $\mathcal{M}.append((\mathcal{G}^+_l, \mathcal{F}_{l+1}))$ 
21   |   |  $l++$ 
22   | while  $l < L \wedge \mathcal{F}_{l+1} \neq \emptyset$ ;
23   |  $\mathcal{G}^-_k \leftarrow \mathcal{G}^+_{l-1}$ 
24   |  $\mathcal{F}_{k+1} \leftarrow \text{curmudgeonQAs}((\mathcal{G}^-_k, \mathcal{H}))$ 
25   |  $\mathcal{H}.append((\mathcal{G}^-_k, \mathcal{F}_{k+1}))$ 
26   |  $k++$ 
27 while  $k < K \wedge \mathcal{F}_{k+1} \neq \emptyset$ ;
28  $\mathcal{G}^* \leftarrow \mathcal{G}^-_{k-1}$ 
29 return  $\mathcal{G}^*$ 

```

are aggregated into $\mathcal{Q}^+ = \{\mathcal{G}^{p_j}\}_{j=1}^M$, then subjected to iterative refinement and evaluation, ultimately resulting in \mathcal{G}^* . For better coverage of the overall information and the relationships between the key concepts within the context, CIR3 introduces W_{p_0} based on the corpus domain. Additionally, this approach guarantees at least one agent will be available even if no subtopics are identified.

3.2. Transactive Reasoning

[52] explores how the group structure, the pattern of connections between individuals, can significantly influence collective cognition and shared knowledge within the group. This suggests that the structure of a network plays a crucial role in how memories are shared and aligned within a group. For instance, centralized networks, where information flows through a few key individuals, can lead to faster memory alignment but may also result in the loss of some details. In contrast, decentralized networks, with more diverse connections, may preserve a wider range of memories but take longer to reach consensus.

Drawing upon these insights, CIR3 employs a hybrid topology that consists of decentralized network of writer agents within a centralized network of two more agents, *moderator* and *curmudgeon*. To encourage communication and interaction, the group of writers form a fully-connected graph, where they operate at the same hierarchical level. To facilitate transactive reasoning, CIR3 adopts a reflection process, which benefits from the iterative exchange of critiques and refinements among the writers (Figure 1 ④). At iteration l , CIR3 gathers and aggregates feedback from all writers into $\mathcal{F}_l = \{\mathcal{F}_l^j\}_{j=1}^M$, links it to the previous QAs, \mathcal{G}^+_{l-1} , and then appends this updated information to the transactive memory. This creates a sequential memory state that evolves with each iteration:

$$\mathcal{M} = \{(\mathcal{G}^+_0, \mathcal{F}_1), \dots, (\mathcal{G}^+_{l-1}, \mathcal{F}_l)\}_{l=0}^L \quad (2)$$

The reflection prompt is specifically designed to encourage the participants to build upon each other’s analysis, while maintaining comprehensive and faithful output. To incentivize the agents to seek an optimal consensus, CIR3 builds upon the group’s decentralized graph to (1) capitalize on the strengths inherent in centralized networks, and (2) incite the group towards a shared and optimal solution (3.3).

3.3. Guiding Collective Cognitive Convergence

In addressing **R.3**, we take inspiration from the phenomenon of *Collective Cognitive Convergence (C3)* [53, 54] and from *How social network topology can shape collective cognition* [52]. C3 highlights that while convergence facilitates mutual understanding and coordination, if left unchecked, it can lead to cognitive collapse, by reducing the diversity of concepts to which the group is exposed, hence limiting the group’s ability to explore other viewpoints and generate new ideas.

In order to generate the optimal solution \mathcal{G}^* , CIR3 capitalizes on: **(1)** The strengths of combining decentralized and centralized networks, where (a) the information flow in the group of decentralized writers facilitates the preservation of a wider range of \mathcal{G}^+ , which is amplified by the multi-perspective analysis, and (b) the rate of convergence in the broader centralized network (between (a) and the curmudgeon agent) facilitates a faster memory alignment of \mathcal{G}^+ ; **(2)** The curmudgeon agent as a mechanism for introducing variation. Coupled with external evaluation tools (Figure 1 ⑤), the curmudgeon guides the writers towards a balanced collective convergence on the key concepts within the document, while maintaining diversity in the output.

Combined with the benefits of CIR3’s hybrid topology, the cyclic process of reflection and refinement, between the writers and the curmudgeon (Figure 1 ⑥), amplifies the collective intelligence, and enables collaborative knowledge construction by sharing, discussing, and building upon each other’s analysis, leading to a deeper understanding of the document. Additionally, this approach offers a solution to mitigate the disadvantages inherent in both centralized (potential loss of information) and decentralized (potential slow convergence) networks.

Once the inner-refinement cycle reaches either an agreement or the pre-defined maximum number of iterations, L , CIR3 is prompted to create a separate record of the latest refined QA pairs, $\mathcal{G}^- = \mathcal{G}^+_l$. This state is then passed to the outer-refinement cycle k , where the curmudgeon appends its feedback $\mathcal{F}l_k$ along \mathcal{G}^-_{k-1} to the transactive memory, creating a central memory state that evolves with each outer-iteration of refinement:

$$\mathcal{H} = \{(\mathcal{G}^-_0, \mathcal{F}l_1), \dots, (\mathcal{G}^-_{k-2}, \mathcal{F}l_{k-1}), (\mathcal{G}^*, \emptyset)\}_{k=1}^K \quad (3)$$

where \emptyset denotes a satisfactory alignment between the curmudgeon and the writers, which then routes the subsequent operation to the termination phase, through the moderator, yielding the final output \mathcal{G}^* and halting the generation process (Figure 1 ⑦).

The curmudgeon is equipped with an evaluation tool to help quantifying the diversity of (a) generated questions, (b) generated answers, and (c) concatenated answers and input context. A lower diversity score in (c), combined with higher diversity scores in (a) and (b), would indicate high faithfulness and better coverage of the input context. To achieve this, we use *Vendi Score*⁵ [56] as an evaluation tool for diversity, where the objective is to minimize diversity in (c), while maximizing it in (a) and (b).

At each iteration k , the curmudgeon initially analyses \mathcal{G}^- to reason about the next action, whether to use the evaluation tool or to rely on its intrinsic knowledge to provide feedback. The result is then added to \mathcal{H} (Equation 3), which invokes another cycle of inner-refinements among the writers.

4. Experiments

This section presents an empirical evaluation of CIR3’s performance. We begin by describing the datasets employed, followed by an overview of the baselines used for comparison. Next, we detail the implementation of CIR3, and finally, we discuss the evaluation metrics, which include statistical, encoder-based, and LLM-based approaches.

4.1. Datasets

While widely used QA datasets like MS MARCO [57] and Natural Questions [58] offer valuable resources, they fall short for our purposes due to the lack of both in-domain and specialized QA datasets, as well as an insufficient coverage of comprehensive QA pairs per document. As a result, we conduct our experiments exclusively on passages from two specialized finance datasets: (1) **FiQA** [59]. This dataset⁶ was used in the Financial Opinion Mining and Question Answering challenge at the 2018 International World Wide Web Conference. FiQA comprises 6,648 questions and 57,640 answer

⁵**Vendi Score** is a similarity-based diversity metric derived from quantum statistical mechanics. It quantifies diversity while accounting for item similarity and not requiring prevalence information. For CIR3, we employ SimCSE models [55] from Princeton as foundational encoders for the Vendi score. Our implementation extends this setup to include various embedders. Empirically, SimCSE produces scores in the range of 1 to 2, with 1 indicating perfect similarity, typically observed between a given context and its corresponding concatenated answers.

⁶<https://huggingface.co/datasets/BeIR/fiqa>

passages. It was curated from financial posts on platforms such as Stackexchange⁷, Reddit⁷, and StockTwits⁷ between 2009 and 2017, with the objective of developing QA systems that can address financial queries by leveraging information from various sources such as micro-blogs, reports, and news articles; **(2) InsuranceQA** [60] (InsurQA). This corpus⁸ was sourced from the Insurance Library⁹ website, consists of 16,889 real-world user questions and 27,413 corresponding answers written by professionals with extensive domain knowledge in the insurance industry. For each dataset, a subset of 2000 passages is randomly chosen for our main experiments.

4.2. Baselines

Prior research in this area has used varied experimental setups and has not focused on generating comprehensive sets of QA pairs from individual documents. As a result, direct comparisons between these works are challenging. Therefore, we establish the following baselines for our study:

- **LLM-DP:** This baseline directly prompts META-LLAMA-3-70B-INST¹⁰ to generate QAs without explicit reasoning or tool utilization. It serves as a measure of the LLM’s ground performance.
- **qGen-aGen:** In this pipeline, we employ QUERY-GEN-MS- MARCO-T5-LARGE-V1 from the Benchmarking IR BEIR [61] to generate questions, which are then fed into META-LLAMA-3-70B to produce corresponding answers. This baseline assesses the LLM’s performance when guided by an external query generation model.

4.3. CIR3 Implementation

We develop CIR3 using the LangGraph¹¹ library. In selecting models, we opt for the INSTRUCT versions of META-LLAMA-3- $\{70B,8B\}$ due to their outstanding performance. We set a temperature of 0.1 and a nucleus sampling of 0.5. To streamline the inference process, we utilize Groq API¹², which provides seamless integration. We also limit the number of generated

⁷<https://stackexchange.com>; <https://stocktwits.com>; <https://reddit.com>

⁸<https://github.com/shuzi/insuranceQA>

⁹<https://www.insurancelibrary.com/>

¹⁰<https://ai.meta.com/blog/meta-llama-3>

¹¹<https://langchain-ai.github.io/langgraph>

¹²<https://groq.com>

QAs to 10 pairs per context, and the number of refinement iterations K , L to 6, 12, respectively.

In the current stage of our research, we focus on working with models from the same Llama-3 family. However, we recognize the potential benefits of exploring a heterogeneous setup in future research, as this could provide valuable observations into the collaborative workflow in multi-agent settings.

4.4. Evaluation Metrics

Automatic evaluation of generated text remains a challenge as traditional metrics fail to align with human assessments. To address this limitation and provide a more comprehensive and refined evaluation of CIR3, we augment standard metrics with LLM-based scores tailored to our specific use case.

Statistical Scorers. We first use ROUGE-L [62], METEOR [63], and Jaccard Index [64] to calculate the scores between (1) the generated questions \mathcal{Q} and the context c as reference, (2) the generated answers \mathcal{A} and c , and (3) \mathcal{Q} and \mathcal{A} . Then, we calculate the mean score over (1), (2) and (3), before calculating the average scores over each evaluation dataset.

Encoder-based Scorers. Beyond the token overlap, we also use embedding-based similarity metrics, such as BERTScore [65] and BAAI/BGE-LARGE-EN-V1.5 (denoted with BGE score in this study). We measure the mean semantic scores between (1) c and \mathcal{Q} , (2) c and \mathcal{A} , and (3) \mathcal{Q} and \mathcal{A} . To assess the quality of QAs when considered collectively, we also include BGE scores between (4) the concatenated questions $\mathcal{Q}_\oplus = \oplus_{i=1}^N q_i$ and c , (5) the concatenated answers $\mathcal{A}_\oplus = \oplus_{i=1}^N a_i$ and c , and (6) \mathcal{Q}_\oplus and \mathcal{A}_\oplus .

LLM-based Scorers. To further quantify the comprehensiveness and faithfulness of the generated QA pairs, we adapt the G-EVAL [66] framework by merging the task definition and evaluation criteria prompt with a Chain-of-Thoughts (CoT) prompt [67] to specify detailed evaluation steps. This modification provides greater control over the assessment process compared to the original G-EVAL, where the LLM generates the CoT automatically. We evaluate the comprehensiveness of \mathcal{G}^* based on *coverage*, *depth*, *accuracy* and *coherence*. Similarly, we evaluate the faithfulness based on *accuracy*, *exaggeration*, *consistency*, *justification*, *plausibility*, and *misrepresentation*. Additionally, we retain the G-EVAL scoring function, which normalizes scores using a weighted sum of token probabilities in LLM output. We also used GPT-4 with the temperature set to 0 to ensure reproducibility.

Further details on the metrics and score calculations employed in this work are provided in [Appendix A: Automatic Metrics](#).

5. Results and Observations

This section presents our experimental findings, covering key results, human evaluations, and ablation studies to assess the effect of multi-perspective reasoning, and the impact of introducing variation.

5.1. Main Results

In all tables, the best-performing model is highlighted in **bold**, with the second-best underlined.

Table 1: Evaluation results using standard metrics. † denotes significant differences ($p < 0.05$) from a paired t -test between **CIR3** and the best baseline **LLM-DP**.

| Dataset | Model | METEOR | | | | ROUGE-L (F1 Scores) | | | | Jaccard Index | | | |
|---------|---------------|---------------------|---------------------|-------------------------------|-----------------|---------------------|---------------------|-------------------------------|---------------|---------------------|---------------------|-------------------------------|-----------------|
| | | $s(c, \mathcal{Q})$ | $s(c, \mathcal{A})$ | $s(\mathcal{Q}, \mathcal{A})$ | Avg. | $s(c, \mathcal{Q})$ | $s(c, \mathcal{A})$ | $s(\mathcal{Q}, \mathcal{A})$ | Avg. | $s(c, \mathcal{Q})$ | $s(c, \mathcal{A})$ | $s(\mathcal{Q}, \mathcal{A})$ | Avg. |
| FiQA | LLM-DP | 0.1571 | 0.3068 | 0.2119 | <u>0.2252</u> | 0.1951 | 0.3189 | 0.2781 | 0.2640 | 0.4377 | 0.5286 | 0.4881 | <u>0.4847</u> |
| | QGEN-AGEN | 0.1288 | 0.3383 | 0.1613 | 0.2094 | 0.1771 | 0.4003 | 0.2690 | <u>0.2821</u> | 0.4161 | 0.5391 | 0.4703 | 0.4751 |
| | CIR3 | 0.1935 | 0.3791 | 0.2767 | 0.2830 † | 0.2153 | 0.3771 | 0.2893 | 0.2939 | 0.5511 | 0.6112 | 0.5983 | 0.5868 † |
| INSURQA | LLM-DP | 0.2422 | 0.3972 | 0.2717 | <u>0.3037</u> | 0.2877 | 0.4984 | 0.3447 | <u>0.3769</u> | 0.4784 | 0.5920 | 0.4987 | <u>0.5230</u> |
| | QGEN-AGEN | 0.1433 | 0.3134 | 0.1283 | 0.1949 | 0.1898 | 0.4903 | 0.2463 | 0.3088 | 0.3885 | 0.5749 | 0.4729 | 0.4787 |
| | CIR3 | 0.3197 | 0.4391 | 0.3632 | 0.3739 † | 0.2950 | 0.4891 | 0.3972 | 0.3937 | 0.5261 | 0.6716 | 0.6104 | 0.6027 † |

From Table 1, we observe that our approach outperforms both baselines on both datasets in terms of lexical metrics. Notably, CIR3 exhibits a relative improvement of 6.39%, 2.33%, 9.08% on METEOR, ROUGE-L and Jaccard Index, respectively, compared to the second best results. Although the observed overlap might not suggest a high degree of similarity, it is important to consider the limitations of lexical metrics, which are inherently less effective when evaluating generative tasks.

Table 2: Evaluation results using embedding-based metrics. † denotes significant differences ($p < 0.05$) from a paired t -test between **CIR3** and the best baseline **LLM-DP**.

| Dataset | Model | BERTScore (F1 Scores) | | | | BGE Semantic Similarity | | | | | | | |
|---------|---------------|-----------------------|---------------------|-------------------------------|-----------------|-------------------------|---------------------|-------------------------------|---------------|------------------------------|------------------------------|---|-----------------|
| | | $s(c, \mathcal{Q})$ | $s(c, \mathcal{A})$ | $s(\mathcal{Q}, \mathcal{A})$ | Avg. | $s(c, \mathcal{Q})$ | $s(c, \mathcal{A})$ | $s(\mathcal{Q}, \mathcal{A})$ | Avg. | $s(c, \mathcal{Q}_{\oplus})$ | $s(c, \mathcal{A}_{\oplus})$ | $s(\mathcal{Q}_{\oplus}, \mathcal{A}_{\oplus})$ | Avg. |
| FiQA | LLM-DP | 0.8415 | 0.8597 | 0.8701 | <u>0.8571</u> | 0.6858 | 0.6847 | 0.7872 | <u>0.7192</u> | 0.7548 | 0.8078 | 0.8488 | <u>0.8038</u> |
| | QGEN-AGEN | 0.8339 | 0.8617 | 0.8472 | 0.8475 | 0.6932 | 0.7051 | 0.7358 | 0.7113 | 0.7462 | 0.8087 | 0.8183 | 0.7910 |
| | CIR3 | 0.8702 | 0.9171 | 0.9088 | 0.8987 † | 0.8312 | 0.8542 | 0.8115 | 0.8323 | 0.8291 | 0.9118 | 0.9264 | 0.8891 † |
| INSURQA | LLM-DP | 0.8511 | 0.8810 | 0.8779 | <u>0.8700</u> | 0.7388 | 0.7540 | 0.8097 | <u>0.7675</u> | 0.8173 | 0.8948 | 0.8675 | <u>0.8598</u> |
| | QGEN-AGEN | 0.8282 | 0.8757 | 0.8472 | 0.8503 | 0.7231 | 0.7404 | 0.7344 | 0.7326 | 0.7708 | 0.8539 | 0.7487 | 0.7911 |
| | CIR3 | 0.8972 | 0.9298 | 0.9175 | 0.9148 † | 0.7591 | 0.7736 | 0.8616 | 0.7980 | 0.8450 | 0.9395 | 0.9072 | 0.8972 † |

Further analysis in Table 2 shows that CIR3 consistently surpasses other models in semantic similarity metrics. CIR3 achieves an average improvement of 4.30% on BERTScore and 7.18% on BGE compared to the second-best model. This trend extends to contextual semantic similarity between the

context and concatenated answers, suggesting that CIR3’s generated answers are more faithful to the input text, potentially indicating lower hallucination and improved comprehensiveness.

Table 3: LLM-based evaluation results for *comprehensiveness* and *faithfulness*.

| Dataset | Model | Comprehensive | Faithful | Avg. |
|---------|-------------|---------------|----------|---------------|
| FIQA | LLM-DP | 0.7169 | 0.8030 | <u>0.7599</u> |
| | QGEN-AGEN | 0.5290 | 0.8414 | 0.6852 |
| | CIR3 | 0.9312 | 0.9762 | 0.9537 |
| INSURQA | LLM-DP | 0.7317 | 0.8175 | <u>0.7746</u> |
| | QGEN-AGEN | 0.5560 | 0.8763 | 0.7161 |
| | CIR3 | 0.9389 | 0.9879 | 0.9634 |

These findings are further reinforced in Table 3 (LLM-based evaluation results) where CIR3 improves comprehensiveness and faithfulness with average scores of over 21% and 17%, respectively, compared to the second best model (LLM-DP). These results provide additional validation for Method 3.3 and Equation 1, wherein the curmudgeon, utilizing a diversity-based evaluation tool, directs the generation of diverse QAs (Comprehensiveness) while ensuring the alignment of the answers with the context (Faithfulness).

Interestingly, LLM-DP demonstrates superior performance compared to QGEN-AGEN in all tests. This implies that the added query generator may not be beneficial, possibly due to the limitations of the T5 [68] model in uncovering deeper key concepts in financial documents.

Our analysis also reveals, in Tables 1 and 2, that CIR3’s questions are significantly more aligned with the context compared to both baselines. This indicates that the CIR3’s deep engagement with the input document helps bridging the gaps in machine reading comprehension, which results in more comprehensive and relevant question generation.

The results presented in Tables 1, 2, 3, provide compelling evidence of the effectiveness of CIR3.

5.2. Human Evaluation

We further conduct human evaluation on 80 samples from the InsurQA corpus and the corresponding generated QA pairs by CIR3 and LLM-DP. We ask 8 experts in finance¹³ to assess 10 sets of QA pairs each, focusing on comprehensiveness and faithfulness. Comprehensiveness is evaluated based

¹³Volunteers have 2 to 6 years of experience in the finance domain, all based in Europe

on three aspects: *coverage*, *depth*, and *coherence*. Similarly, faithfulness is assessed based on: *accuracy*, *representation*, and *diversification*. Each aspect is scored on a scale from 1 (worst) to 5 (best).

Table 4: Human evaluation results on 80 sets of QA pairs generated by CIR3 and LLM-DP. The ratings (1 to 5) are normalised between 0 and 1. The scores are analysed using a paired *t*-test (*p*-values are presented).

| | Aspect | LLM-DP | CIR3 | <i>p</i> -value |
|-------------------|-----------------|--------|---------------|-----------------|
| Comprehensiveness | Coverage | 0.7875 | 0.9375 | 0.0033 |
| | Depth | 0.7750 | 0.9125 | 0.0038 |
| | Coherence | 0.7625 | 0.9250 | 0.0023 |
| | Avg. | 0.7750 | 0.9250 | |
| Faithfulness | Accuracy | 0.7500 | 0.9125 | 0.0020 |
| | Representation | 0.7875 | 0.9125 | 0.0042 |
| | Diversification | 0.8250 | 0.8875 | 0.0104 |
| | Avg. | 0.7875 | 0.9041 | |

Table 4 shows the average scores and paired *t*-test results, aligning with the findings in Table 3. CIR3 demonstrates significant improvement over the baseline LLM-DP, with an increase of 15% on comprehensiveness and 11.66% on faithfulness.

5.3. Ablation Studies

To provide additional support for our hypotheses in H.2 and H.3, we conduct an ablation study with two variations of CIR3:

- (1) **CIR3 w/o perspectives.** Following [46], in this variation, we aim to assess the impact of multi-perspective reasoning. We modify the moderator’s prompt by removing the section that assigns diverse perspectives to the writer agents. To ensure a fair comparison, we maintain the same number of writers as in the original model (determined by the number of identified subtopics);
- (2) **CIR3 w/o Curmudgeon.** In this variation, we disable the curmudgeon agent to evaluate the effect of introducing external variation to the writer’s sub-network.

For this study, we randomly select 200 samples, equally split between both datasets, and capped the refinement cycles between writers at 12 for each input. The results in Table 5 demonstrate that CIR3 surpasses the two alternative variations. Nonetheless, both variations outperform the baseline LLM-DP, providing some support for our hypotheses.

Effect of multi-perspective reasoning. Table 5 shows that *CIR3 w/o perspectives* yields inferior results compared to CIR3, suggesting that multi-

Table 5: Effect of multi-perspective reasoning and Curmudgeon on *comprehensiveness* and *faithfulness*

| Model | Comprehensiveness | Faithfulness | Avg. |
|-----------------------|-------------------|---------------|---------------|
| LLM-DP | 0.7399 | 0.8221 | 0.7810 |
| CIR3 | 0.9451 | 0.9895 | 0.9673 |
| CIR3 w/o perspectives | 0.9115 | 0.9653 | 0.9384 |
| CIR3 w/o Curmudgeon | 0.8370 | 0.9046 | 0.8708 |

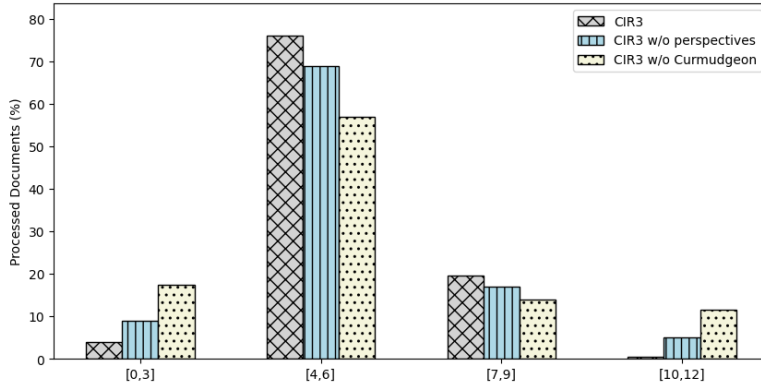


Figure 2: Number of inner-refinement cycles (x -axis), given as intervals, required to process the input documents (y -axis), given as percentage.

perspective group debate contributes to a comprehensive and faithful output, as proposed in [H.2](#).

Effect of variation. Removing the disruptive signal, in *CIR3 w/o Curmudgeon*, significantly impairs performance, reducing faithfulness by 8.49% and comprehensiveness by 10.81%. This can be explained by examining the number of refinement cycles (given as intervals) required to process the input documents, as in [Figure 2](#). Compared to CIR3, and *CIR3 w/o perspectives*, *CIR3 w/o Curmudgeon* shows a significant increase in the number of contexts falling within the refinement cycle ranges $[0, 3]$ and $[10, 12]$. For the interval $[0, 3]$, *CIR3 w/o Curmudgeon* exhibits a 13.5% increase compared to CIR3 and an 8.5% increase compared to *CIR3 w/o perspectives*. Similarly, for the interval $[10, 12]$, *CIR3 w/o Curmudgeon* shows an 11% increase over CIR3 and a 6.5% increase over *CIR3 w/o perspectives*. This aligns with [H.3](#), where the absence of variation can result in either (1) a potential immature collective convergence (collapse) and loss of information, characterized by a small number of iterations and potentially low comprehensiveness scores, or (2) a

potential slow convergence, characterized by a large number of iterations and a high likelihood of low faithfulness.

6. Conclusion and Future Work

This paper presented CIR3, a novel system for comprehensive and faithful QAG from information-dense documents. A key contribution lies in addressing the more challenging QAG task compared to traditional QG, effectively navigating a constrained search space for unique and relevant QA pairs. Notably, to the best of our knowledge, CIR3 is the first proposed QAG approach employing multi-agent LLMs, orchestrating information flow via transactive reasoning, multi-perspective assessment, and balanced collective convergence.

Our research demonstrates that integrating an external signal significantly enhances convergence and diversity within the agent group, enabling efficient agreement on comprehensive and faithful QA pairs representing core text concepts, a crucial aspect of CIR3’s design. To improve alignment with human evaluation, we developed a custom metric leveraging encoder and LLM-based scores on individual and concatenated QA pairs, providing a refined quality assessment. Empirical results confirm CIR3’s significant performance gains over strong baselines.

Future work will explore CIR3’s modularity by integrating alternative classification methods, such as fine-tuned encoder-based classifiers, to potentially refine perspective selection. Further research will investigate heterogeneous multi-agent frameworks and expand CIR3’s applicability to diverse domains and tasks, including summarization, information retrieval, and multi-modal applications.

Appendix A. Metrics

Appendix A.1. Automatic Metrics

We provide a brief description of the metrics used in this study:

ROUGE-L [62] assesses recall by evaluating the overlap between reference and generated sentences using Longest Common Subsequence statistics. We use the implementation from GOOGLE¹⁴. In this paper, we report the F1 score, the harmonic mean of precision and recall.

¹⁴<https://pypi.org/project/rouge-score>

METEOR [63] is a recall-oriented metric that measures the similarity between generated and reference text, incorporating synonyms, stemming, and paraphrasing. We use the implementation from NLTK¹⁵.

Jaccard Index¹⁶[64] is a measure of similarity between two sets. It is calculated as the size of their intersection (elements they share) divided by the size of their union (total unique elements). Values range from 0 (no similarity) to 1 (identical sets). We adopt SCIKIT-LEARN’s implementation¹⁷.

BERTScore¹⁸ [65] uses contextual embeddings to assess word-level similarity via cosine similarity, correlating with human judgment in sentence and system evaluation, and providing precision, recall, and F1 metrics.

BAAI/bge-large¹⁹ is a high-performance sentence embedding model, designed for semantic similarity tasks. It encodes text into dense vectors, allowing similarity to be measured via cosine similarity between embeddings.

Appendix A.2. Score Calculations

We denote $s(c, \mathcal{Q})$, $s(c, \mathcal{A})$, $s(\mathcal{Q}, \mathcal{A})$, $s(c, \mathcal{Q}_\oplus)$, $s(c, \mathcal{A}_\oplus)$, and $s(\mathcal{Q}_\oplus, \mathcal{A}_\oplus)$ the scores between (context and questions), (context and answers), (questions and answers), (context and concatenated questions), (context and concatenated answers), and (concatenated questions and concatenated answers), respectively. The scores are calculated as follows:

$$s(c, \mathcal{Q}) = \frac{1}{N} \sum_{i=1}^N s(c, q_i) \quad (\text{A.1}) \quad s(c, \mathcal{Q}_\oplus) = s(c, \oplus_{i=1}^N q_i) \quad (\text{A.4})$$

$$s(c, \mathcal{A}) = \frac{1}{N} \sum_{i=1}^N s(c, a_i) \quad (\text{A.2}) \quad s(c, \mathcal{A}_\oplus) = s(c, \oplus_{i=1}^N a_i) \quad (\text{A.5})$$

$$s(\mathcal{Q}, \mathcal{A}) = \frac{1}{N} \sum_{i=1}^N s(q_i, a_i) \quad (\text{A.3}) \quad s(\mathcal{Q}_\oplus, \mathcal{A}_\oplus) = s(\oplus_{i=1}^N q_i, \oplus_{i=1}^N a_i) \quad (\text{A.6})$$

where s is the scoring function and \oplus is the concatenation function.

¹⁵<https://www.nltk.org>

¹⁶https://en.wikipedia.org/wiki/Jaccard_index

¹⁷<https://scikit-learn.org>

¹⁸https://github.com/Tiiiger/bert_score

¹⁹<https://github.com/FlagOpen/FlagEmbedding>

References

- [1] Y. Zhu, H. Yuan, S. Wang, J. Liu, W. Liu, C. Deng, H. Chen, Z. Dou, J.-R. Wen, Large Language Models for Information Retrieval: A Survey (Jan. 2024). [arXiv:2308.07107](https://arxiv.org/abs/2308.07107).
- [2] L. Silva, L. Barbosa, [Improving dense retrieval models with LLM augmented data for dataset search](#), Knowledge-Based Systems 294 (2024) 111740. [doi:10.1016/j.knosys.2024.111740](https://doi.org/10.1016/j.knosys.2024.111740).
URL <https://linkinghub.elsevier.com/retrieve/pii/S0950705124003757>
- [3] R. Li, Y. Wang, Z. Wen, M. Cui, Q. Miao, [Different paths to the same destination: Diversifying LLMs generation for multi-hop open-domain question answering](#), Knowledge-Based Systems 309 (2025) 112789. [doi:10.1016/j.knosys.2024.112789](https://doi.org/10.1016/j.knosys.2024.112789).
URL <https://linkinghub.elsevier.com/retrieve/pii/S0950705124014230>
- [4] S. Shen, Y. Li, N. Du, X. Wu, Y. Xie, S. Ge, T. Yang, K. Wang, X. Liang, W. Fan, On the Generation of Medical Question-Answer Pairs, Proceedings of the AAAI Conference on Artificial Intelligence 34 (05) (2020) 8822–8829. [doi:10.1609/aaai.v34i05.6410](https://doi.org/10.1609/aaai.v34i05.6410).
- [5] Z. Zeng, Q. Cheng, X. Hu, Y. Zhuang, X. Liu, K. He, Z. Liu, [KoSEL: Knowledge subgraph enhanced large language model for medical question answering](#), Knowledge-Based Systems 309 (2025) 112837. [doi:10.1016/j.knosys.2024.112837](https://doi.org/10.1016/j.knosys.2024.112837).
URL <https://linkinghub.elsevier.com/retrieve/pii/S0950705124014710>
- [6] D. Lindberg, F. Popowich, J. Nesbit, P. Winne, Generating Natural Language Questions to Support Learning On-Line, in: A. Gatt, H. Saggion (Eds.), Proceedings of the 14th European Workshop on Natural Language Generation, Association for Computational Linguistics, Sofia, Bulgaria, 2013, pp. 105–114.
- [7] N.-T. Le, T. Kojiri, N. Pinkwart, Automatic Question Generation for Educational Applications – The State of Art, in: T. van Do, H. A. L. Thi, N. T. Nguyen (Eds.), Advanced Computational Methods for Knowledge

Engineering, Springer International Publishing, Cham, 2014, pp. 325–338. [doi:10.1007/978-3-319-06569-4_24](https://doi.org/10.1007/978-3-319-06569-4_24).

- [8] G. Kumar, R. Banchs, L. F. D’Haro, RevUP: Automatic Gap-Fill Question Generation from Educational Texts, in: J. Tetreault, J. Burstein, C. Leacock (Eds.), Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, Denver, Colorado, 2015, pp. 154–161. [doi:10.3115/v1/W15-0618](https://doi.org/10.3115/v1/W15-0618).
- [9] M. Uto, Y. Tomikawa, A. Suzuki, Difficulty-Controllable Neural Question Generation for Reading Comprehension using Item Response Theory, in: E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, T. Zesch (Eds.), Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 119–129. [doi:10.18653/v1/2023.bea-1.10](https://doi.org/10.18653/v1/2023.bea-1.10).
- [10] Y. Meng, L. Pan, Y. Cao, M.-Y. Kan, FollowupQG: Towards information-seeking follow-up question generation, in: J. C. Park, Y. Arase, B. Hu, W. Lu, D. Wijaya, A. Purwarianti, A. A. Krisnadi (Eds.), Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Nusa Dua, Bali, 2023, pp. 252–271. [doi:10.18653/v1/2023.ijcnlp-main.17](https://doi.org/10.18653/v1/2023.ijcnlp-main.17).
- [11] R. Zhang, J. Guo, L. Chen, Y. Fan, X. Cheng, A Review on Question Generation from Natural Language Text, ACM Transactions on Information Systems 40 (1) (2022) 1–43. [doi:10.1145/3468889](https://doi.org/10.1145/3468889).
- [12] S. Vakulenko, B. Byrne, A. de Gispert, Uniform Training and Marginal Decoding for Multi-Reference Question-Answer Generation, in: ECAI 2023, IOS Press, 2023, pp. 2378–2385. [doi:10.3233/FAIA230539](https://doi.org/10.3233/FAIA230539).
- [13] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray,

- B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language Models are Few-Shot Learners (Jul. 2020). [arXiv:2005.14165](https://arxiv.org/abs/2005.14165), [doi:10.48550/arXiv.2005.14165](https://doi.org/10.48550/arXiv.2005.14165).
- [14] M. Alaofi, L. Gallagher, M. Sanderson, F. Scholer, P. Thomas, Can Generative LLMs Create Query Variants for Test Collections? An Exploratory Study, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 1869–1873. [doi:10.1145/3539618.3591960](https://doi.org/10.1145/3539618.3591960).
- [15] X. Yuan, T. Wang, Y.-H. Wang, E. Fine, R. Abdelghani, H. Sauz on, P.-Y. Oudeyer, Selecting Better Samples from Pre-trained LLMs: A Case Study on Question Generation, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 12952–12965. [doi:10.18653/v1/2023.findings-acl.820](https://doi.org/10.18653/v1/2023.findings-acl.820).
- [16] Y. Cheng, C. Zhang, Z. Zhang, X. Meng, S. Hong, W. Li, Z. Wang, Z. Wang, F. Yin, J. Zhao, X. He, Exploring Large Language Model based Intelligent Agents: Definitions, Methods, and Prospects (Jan. 2024). [arXiv:2401.03428](https://arxiv.org/abs/2401.03428), [doi:10.48550/arXiv.2401.03428](https://doi.org/10.48550/arXiv.2401.03428).
- [17] X. Liu, H. Yu, H. Zhang, Y. Xu, X. Lei, H. Lai, Y. Gu, H. Ding, K. Men, K. Yang, S. Zhang, X. Deng, A. Zeng, Z. Du, C. Zhang, S. Shen, T. Zhang, Y. Su, H. Sun, M. Huang, Y. Dong, J. Tang, AgentBench: Evaluating LLMs as Agents, in: The Twelfth International Conference on Learning Representations, 2023, pp. –.
- [18] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, I. Mordatch, Improving Factuality and Reasoning in Language Models through Multiagent Debate (May 2023). [arXiv:2305.14325](https://arxiv.org/abs/2305.14325).
- [19] T. Liang, Z. He, W. Jiao, X. Wang, Y. Wang, R. Wang, Y. Yang, Z. Tu, S. Shi, Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate (May 2023). [arXiv:2305.19118](https://arxiv.org/abs/2305.19118).
- [20] T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, X. Zhang, Large Language Model based Multi-Agents: A Survey of Progress and Challenges (Apr. 2024). [arXiv:2402.01680](https://arxiv.org/abs/2402.01680).

- [21] M. L. Blanton, D. A. Stylianou, [Understanding the role of transactive reasoning in classroom discourse as students learn to construct proofs](#), *The Journal of Mathematical Behavior* 34 (2014) 76–98. doi:[10.1016/j.jmathb.2014.02.001](#).
URL <https://www.sciencedirect.com/science/article/pii/S0732312314000157>
- [22] A. W. Woolley, P. Gupta, [Understanding collective intelligence: Investigating the role of collective memory, attention, and reasoning processes](#), *Perspectives on Psychological Science* 19 (2023) 344 – 354.
URL <https://api.semanticscholar.org/CorpusID:261334778>
- [23] E. Kamaloo, N. Dziri, C. Clarke, D. Rafiei, [Evaluating Open-Domain Question Answering in the Era of Large Language Models](#), in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 5591–5606. doi:[10.18653/v1/2023.acl-long.307](#).
- [24] M. Heilman, N. A. Smith, [Good Question! Statistical Ranking for Question Generation](#), in: R. Kaplan, J. Burstein, M. Harper, G. Penn (Eds.), *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, Los Angeles, California, 2010, pp. 609–617.
- [25] J. Mostow, W. Chen, [Generating Instruction Automatically for the Reading Strategy of Self-Questioning](#), *International Conference on Artificial Intelligence in Education - (-)* (2009).
- [26] Y. Huang, L. He, [Automatic generation of short answer questions for reading comprehension assessment](#), *Natural Language Engineering* 22 (3) (2016) 457–489. doi:[10.1017/S1351324915000455](#).
- [27] X. Du, J. Shao, C. Cardie, [Learning to Ask: Neural Question Generation for Reading Comprehension](#), in: R. Barzilay, M.-Y. Kan (Eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 1342–1352. doi:[10.18653/v1/P17-1123](#).

- [28] K. D. Dhole, S. Bajaj, R. Chandradevan, E. Agichtein, QueryExplorer: An Interactive Query Generation Assistant for Search and Exploration (Mar. 2024). [arXiv:2403.15667](https://arxiv.org/abs/2403.15667).
- [29] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, SQuAD: 100,000+ Questions for Machine Comprehension of Text, in: J. Su, K. Duh, X. Carreras (Eds.), Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 2383–2392. [doi:10.18653/v1/D16-1264](https://doi.org/10.18653/v1/D16-1264).
- [30] D. Weissenborn, G. Wiese, L. Seiffe, Making Neural QA as Simple as Possible but not Simpler, in: R. Levy, L. Specia (Eds.), Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 271–280. [doi:10.18653/v1/K17-1028](https://doi.org/10.18653/v1/K17-1028).
- [31] S. Min, V. Zhong, R. Socher, C. Xiong, Efficient and Robust Question Answering from Minimal Context over Documents, in: I. Gurevych, Y. Miyao (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 1725–1735. [doi:10.18653/v1/P18-1160](https://doi.org/10.18653/v1/P18-1160).
- [32] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. [doi:10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423).
- [33] C. Alberti, D. Andor, E. Pitler, J. Devlin, M. Collins, Synthetic QA Corpora Generation with Roundtrip Consistency, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 6168–6173. [doi:10.18653/v1/P19-1620](https://doi.org/10.18653/v1/P19-1620).
- [34] R. Puri, R. Spring, M. Shoeybi, M. Patwary, B. Catanzaro, Training Question Answering Models From Synthetic Data, in: B. Webber,

- T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 5811–5826. [doi:10.18653/v1/2020.emnlp-main.468](https://doi.org/10.18653/v1/2020.emnlp-main.468).
- [35] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, in: D. Jurafsky, J. Chai, N. Schlueter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7871–7880. [doi:10.18653/v1/2020.acl-main.703](https://doi.org/10.18653/v1/2020.acl-main.703).
- [36] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach (Jul. 2019). [arXiv:1907.11692](https://arxiv.org/abs/1907.11692), [doi:10.48550/arXiv.1907.11692](https://doi.org/10.48550/arXiv.1907.11692).
- [37] L. Murakhovs'ka, C.-S. Wu, P. Laban, T. Niu, W. Liu, C. Xiong, MixQG: Neural Question Generation with Mixed Answer Types (May 2022). [arXiv:2110.08175](https://arxiv.org/abs/2110.08175).
- [38] M. Bartolo, T. Thrush, R. Jia, S. Riedel, P. Stenetorp, D. Kiela, Improving Question Answering Model Robustness with Synthetic Adversarial Data Generation, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 8830–8848. [doi:10.18653/v1/2021.emnlp-main.696](https://doi.org/10.18653/v1/2021.emnlp-main.696).
- [39] A. Ushio, F. Alva-Manchego, J. Camacho-Collados, An Empirical Comparison of LM-based Question and Answer Generation Methods, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 14262–14272. [doi:10.18653/v1/2023.findings-acl.899](https://doi.org/10.18653/v1/2023.findings-acl.899).
- [40] W. Zhang, W. Hua, K. Stratos, EntQA: Entity Linking as Question Answering (Mar. 2022). [arXiv:2110.02369](https://arxiv.org/abs/2110.02369), [doi:10.48550/arXiv.2110.02369](https://doi.org/10.48550/arXiv.2110.02369).

- [41] V. Puranik, A. Majumder, V. Chaoji, PROTEGE: Prompt-based Diverse Question Generation from Web Articles, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 5449–5463. [doi:10.18653/v1/2023.findings-emnlp.362](https://doi.org/10.18653/v1/2023.findings-emnlp.362).
- [42] S. Hong, M. Zhuge, J. Chen, X. Zheng, Y. Cheng, C. Zhang, J. Wang, Z. Wang, S. K. S. Yau, Z. Lin, L. Zhou, C. Ran, L. Xiao, C. Wu, J. Schmidhuber, MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework (Nov. 2023). [arXiv:2308.00352](https://arxiv.org/abs/2308.00352), [doi:10.48550/arXiv.2308.00352](https://doi.org/10.48550/arXiv.2308.00352).
- [43] G. Li, H. A. A. K. Hammoud, H. Itani, D. Khizbullin, B. Ghanem, CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society (Nov. 2023). [arXiv:2303.17760](https://arxiv.org/abs/2303.17760).
- [44] C. Qian, X. Cong, W. Liu, C. Yang, W. Chen, Y. Su, Y. Dang, J. Li, J. Xu, D. Li, Z. Liu, M. Sun, Communicative Agents for Software Development (Dec. 2023). [arXiv:2307.07924](https://arxiv.org/abs/2307.07924).
- [45] C.-M. Chan, W. Chen, Y. Su, J. Yu, W. Xue, S. Zhang, J. Fu, Z. Liu, ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate (Aug. 2023). [arXiv:2308.07201](https://arxiv.org/abs/2308.07201).
- [46] Y. Shao, Y. Jiang, T. A. Kanell, P. Xu, O. Khattab, M. S. Lam, Assisting in Writing Wikipedia-like Articles From Scratch with Large Language Models (Apr. 2024). [arXiv:2402.14207](https://arxiv.org/abs/2402.14207), [doi:10.48550/arXiv.2402.14207](https://doi.org/10.48550/arXiv.2402.14207).
- [47] K. Xiong, X. Ding, Y. Cao, T. Liu, B. Qin, Examining Inter-Consistency of Large Language Models Collaboration: An In-depth Analysis via Debate, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 7572–7590. [doi:10.18653/v1/2023.findings-emnlp.508](https://doi.org/10.18653/v1/2023.findings-emnlp.508).
- [48] H. Chen, W. Ji, L. Xu, S. Zhao, Multi-Agent Consensus Seeking via Large Language Models (Oct. 2023). [arXiv:2310.20151](https://arxiv.org/abs/2310.20151).

- [49] P. Li, V. Menon, B. Gudiguntla, D. Ting, L. Zhou, Challenges Faced by Large Language Models in Solving Multi-Agent Flocking (Apr. 2024). [arXiv:2404.04752](https://arxiv.org/abs/2404.04752).
- [50] D. J. Hall, R. A. Davis, Engaging multiple perspectives: A value-based decision-making model, *Decision Support Systems* 43 (4) (2007) 1588–1604. [doi:10.1016/j.dss.2006.03.004](https://doi.org/10.1016/j.dss.2006.03.004).
- [51] M. Ithori, H. Sato, T. Tanaka, R. Masumura, Multi-Perspective Document Revision, in: N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, S.-H. Na (Eds.), *Proceedings of the 29th International Conference on Computational Linguistics*, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 6128–6138.
- [52] I. Momennejad, Collective minds: Social network topology shapes collective cognition, *Philosophical Transactions of the Royal Society B: Biological Sciences* 377 (1843) (2021) 20200315. [doi:10.1098/rstb.2020.0315](https://doi.org/10.1098/rstb.2020.0315).
- [53] H. V. D. Parunak, T. C. Belding, R. Hilscher, S. A. Brueckner, *Cognitive collapse: Recognizing and addressing the hidden threat in collaborative technologies*, in: Defense Technical Information Center, ,, , 2008, pp. ,. URL <https://api.semanticscholar.org/CorpusID:17063782>
- [54] H. V. Parunak, T. C. Belding, R. Hilscher, S. Brueckner, Understanding Collective Cognitive Convergence, in: N. David, J. S. Sichman (Eds.), *Multi-Agent-Based Simulation IX*, Springer, Berlin, Heidelberg, 2009, pp. 46–59. [doi:10.1007/978-3-642-01991-3_4](https://doi.org/10.1007/978-3-642-01991-3_4).
- [55] T. Gao, X. Yao, D. Chen, SimCSE: Simple Contrastive Learning of Sentence Embeddings, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 6894–6910. [doi:10.18653/v1/2021.emnlp-main.552](https://doi.org/10.18653/v1/2021.emnlp-main.552).

- [56] D. Friedman, A. B. Dieng, The Vendi Score: A Diversity Evaluation Metric for Machine Learning (Jul. 2023). [arXiv:2210.02410](https://arxiv.org/abs/2210.02410), [doi:10.48550/arXiv.2210.02410](https://doi.org/10.48550/arXiv.2210.02410).
- [57] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, M. Rosenberg, X. Song, A. Stoica, S. Tiwary, T. Wang, MS MARCO: A Human Generated MACHine Reading COMprehension Dataset (Oct. 2018). [arXiv:1611.09268](https://arxiv.org/abs/1611.09268), [doi:10.48550/arXiv.1611.09268](https://doi.org/10.48550/arXiv.1611.09268).
- [58] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M.-W. Chang, A. M. Dai, J. Uszkoreit, Q. Le, S. Petrov, Natural Questions: A Benchmark for Question Answering Research, *Transactions of the Association for Computational Linguistics* 7 (2019) 452–466. [doi:10.1162/tacl_a_00276](https://doi.org/10.1162/tacl_a_00276).
- [59] M. Maia, S. Handschuh, A. Freitas, B. Davis, R. McDermott, M. Zarrouk, A. Balahur, WWW’18 Open Challenge: Financial Opinion Mining and Question Answering, in: *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW ’18*, ACM Press, Lyon, France, 2018, pp. 1941–1942. [doi:10.1145/3184558.3192301](https://doi.org/10.1145/3184558.3192301).
- [60] M. Feng, B. Xiang, M. R. Glass, L. Wang, B. Zhou, Applying deep learning to answer selection: A study and an open task, in: *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, IEEE, Scottsdale, AZ, USA, 2015, pp. 813–820. [doi:10.1109/ASRU.2015.7404872](https://doi.org/10.1109/ASRU.2015.7404872).
- [61] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, I. Gurevych, [BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models](#), in: *Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, ,, ,, 2021, pp. ., .
URL <https://api.semanticscholar.org/CorpusID:233296016>
- [62] C.-Y. Lin, ROUGE: A Package for Automatic Evaluation of Summaries, in: *Text Summarization Branches Out*, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81.

- [63] S. Banerjee, A. Lavie, METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, in: J. Goldstein, A. Lavie, C.-Y. Lin, C. Voss (Eds.), Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 65–72.
- [64] L. da Fontoura Costa, [Further generalizations of the jaccard index](#), ArXiv abs/2110.09619 (2021) –.
URL <https://api.semanticscholar.org/CorpusID:239024336>
- [65] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating Text Generation with BERT (Feb. 2020). [arXiv:1904.09675](#), [doi:10.48550/arXiv.1904.09675](#).
- [66] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, C. Zhu, G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment (May 2023). [arXiv:2303.16634](#), [doi:10.48550/arXiv.2303.16634](#).
- [67] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, Chain-of-Thought Prompting Elicits Reasoning in Large Language Models (Jan. 2023). [arXiv:2201.11903](#), [doi:10.48550/arXiv.2201.11903](#).
- [68] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer (Sep. 2023). [arXiv:1910.10683](#).

- We propose CIR3, a novel framework for comprehensive and faithful QA generation.
- The efficient information flow of CIR3 enables in-depth document analysis.
- We employ transactive reasoning for deeper understanding in CIR3.
- Our approach's multi-perspective assessment ensures balanced views.
- CIR3's balanced collective convergence yields robust results.
- CIR3 improves QA comprehensiveness (+21) and faithfulness (+17)

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

02 May 2025

Cover Letter

Dear Editors,

We are pleased to submit our manuscript entitled "**Coordinated LLM Multi-Agent Systems for Collaborative Question-Answer Generation**" for consideration for publication in **Knowledge-Based Systems**.

In this work, we propose Collective Intentional Reading through Reflection and Refinement (CIR3), a novel framework designed for comprehensive and faithful question-answer (QA) generation. CIR3 enables efficient information flow for in-depth document analysis, employs transactive reasoning for deeper understanding, and integrates multi-perspective assessment to ensure balanced views. Our framework achieves robust results through balanced collective convergence and significantly improves QA comprehensiveness (+21) and faithfulness (+17).

Our main contributions include: (1) shifting from conventional question generation (QG) to the more challenging task of question-answer generation (QAG); (2) introducing, to the best of our knowledge, the first QAG approach utilizing multi-agent large language models (LLMs); (3) demonstrating that incorporating external signals enhances convergence rates and diversity within agent groups; (4) proposing a custom metric leveraging encoder and LLM-based scores to better align automatic evaluation with human judgment; and (5) showing that CIR3 achieves strong improvements over existing baselines.

We believe that our work is particularly relevant to research in data augmentation and knowledge base creation, both of which are critical areas for the journal's readership.

This manuscript is original, has not been published previously, and is not under consideration for publication elsewhere.

Thank you for your time and consideration. We look forward to your feedback.

Sincerely,

Eduardo Alonso (E.Alonso@city.ac.uk)
Sami Saadaoui (Sami.Saadaoui@city.ac.uk)
Artificial Intelligence Research Centre (CitAI),
City St George's, University of London

02 May 2025

Author Agreement Statement

We the undersigned declare that this manuscript, "**Coordinated LLM Multi-Agent Systems for Collaborative Question-Answer Generation**", is original, has not been published before and is not currently being considered for publication elsewhere.

We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

We understand that the Corresponding Author is the sole contact for the Editorial process. He is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs.

Signed by all authors as follows:

Eduardo Alonso (E.Alonso@city.ac.uk)

Sami Saadaoui (Sami.Saadaoui@city.ac.uk)

Artificial Intelligence Research Centre (CitAI),
City St George's, University of London