Knowledge-Based Systems 330 (2025) 114627

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Coordinated LLM multi-agent systems for collaborative question-answer generation

Sami Saadaoui *, Eduardo Alonso

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

ARTICLE INFO

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

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 **C**ollective **I**ntentional **R**eading through **R**eflection and **R**efinement (**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 for 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 23 % and faithfulness by 17 % compared to strong baseline approaches. Code and data are available at https://github.com/anonym-nlp-ai/cirrr.

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 [1–3] to healthcare [4,5], and education [6–8]. 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 Zhang et al. [11], Vakulenko et al. [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.

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 [14–16]. 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.

* Corresponding author.
E-mail address: Sami.Saadaoui@citystgeorges.ac.uk (S. Saadaoui).

https://doi.org/10.1016/j.knosys.2025.114627
Received 2 May 2025; Received in revised form 14 September 2025; Accepted 4 October 2025
Available online 13 October 2025

Box 1. **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.

Recent advancements in LLM-based Multi-Agent[1] (LLM-MA) systems have shown significant improvements in problem-solving abilities through planning, collaboration, and autonomous task execution [19,20]. 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 [21]. 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: **(R1)** Can an LLM-MA workflow uncover deeper and perhaps implicit key concepts from a complex and information-dense document? **(R2)** How can LLM-MA effectively emphasize deep engagement, with a text, from different viewpoints to enable comprehensive and consistent generation and mitigate blind spots? **R3** (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 reasoning[2] 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 and lengthy answers, making lexical matching even less effective [24]. To ensure a comprehensive and accurate evaluation of CIR3, we employ diverse automatic metrics, in addition to human evaluation.

Following similar motivations as Zeng and Zubiaga [25], Amiri-Margavi et al. [26], this work also introduces a novel cross-model semantic evaluation framework that addresses the challenge of validating subtopic identification systems without human gold standards. The framework leverages semantic consensus among multiple LLMs as a reliability baseline, using agglomerative clustering to group semantically similar outputs from sentence transformer embeddings into conceptual topic clusters. The performance of the target system is then assessed through semantic agreement metrics, including *Soft-F1* and *Bipartite matching* for partial conceptual overlap, while consensus reliability is quantified using *Krippendorff's Alpha* [27,28] applied to clustered semantic concepts rather than exact label matches. This approach provides a scalable, model-agnostic solution to evaluate classification consistency in low-resource domains where annotated datasets are unavailable.

In summary, our main contributions include:

(1) We introduce CIR3, a dual-loop multi-agent framework for QAG that formalizes Comprehensiveness and Faithfulness as a diversity-alignment objective and operationalizes transactive memory via inner/outer iterative refinement with explicit termination conditions.
(2) We design a curmudgeon-guided convergence mechanism that maintains diversity while driving consensus, with systematic ablations isolating the contributions of agent reasoning versus diversity measurement.
(3) We propose a two-stage cross-model semantic evaluation protocol for subtopic identification without gold standards, combining agglomerative clustering with pairwise semantic measures (Soft-F1, Bipartite matching) and Krippendorff's Alpha for holistic reliability.
(4) We demonstrate consistent improvements over strong LLM baselines across finance and medical domains through comprehensive automatic and human evaluations, with open-source implementation for reproducibility.

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 [29–31] and neural [14,32,33] models have been extensively used for QG from text documents. Similarly, machine reading comprehension [34–36] 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.

[1] LLM-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 [17,18].

[2] In this paper, we mimic the concept of transactive reasoning [22,23], 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.

Several studies have leveraged pre-trained language models for QAG. These include fine-tuning BERT [37] for AE and QG [38], fine-tuning auto-regressive LMs for QG [39] using BART [40] and RoBERTa [41] for AE, jointly fine-tuning LMs for AE and QG [42], and using QAG models to generate adversarial examples [43]. Recent advances have also focused on dynamically identifying question-worthy context words before using them to condition subsequent question generation [12]. Ushio et al. [44] improved QAG by designing three distinct approaches: Pipeline, Multitask, and End2end. Zhang et al. [45] proposed combining entity linkage with a QA system, while Puranik et al. [46] 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 on generating comprehensive and detailed responses. Additionally, evaluating the comprehensiveness of QAG remains an underexplored area. We address 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 such as Hong et al. [47], Li et al. [48], Qian et al. [49] have shown that collaboration and task division among multiple agents can reduce hallucinations and generate more reliable results. Other works, such as Chan et al. [50], Shao et al. [51], 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 [52] has examined the issue of inter-consistency through inter-agent negotiation. Similarly, drawing inspiration from robotics, Chen et al. [53] 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, Li et al. [54] explored the concept of *flocking* where agents maintain proximity while avoiding collisions and preserving formations. In this work, we improve the 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: $\mathcal{G}^* = \arg\max_{\mathcal{G}} \left[\mathrm{Comp}(\mathcal{G}, c) + \mathrm{Faith}(\mathcal{G}, c)\right]$, where $\mathrm{Comp}(\mathcal{G}, c)$ and $\mathrm{Faith}(\mathcal{G}, c)$ 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:

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

where $\mathcal{D}_\mathrm{q}$ and $\mathcal{D}_\mathrm{a}$ denote diversity scores computed over the sets of generated questions $\{q_i\}_{i=1}^{N}$ and answers $\{a_i\}_{i=1}^{N}$, respectively, and $\mathcal{D}_{\mathrm{a},c}$ denotes a dissimilarity measure between the concatenated answers $a_1 \oplus \cdots \oplus a_N$ and the context c. $\mathcal{D} \in [1,2] \subset \mathbb{R}$, where $\mathcal{D} = 1$ denotes perfect similarity. The coefficients[3] $\alpha_{q,a}$ 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 and 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.

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 [55]. For instance, STORM [51] 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, Ihori et al. [56] 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 $P = \{p_1, \dots, p_M\}$ (Fig. 1 ①-②)[4]. 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 (Fig. 1 ③). Subsequently, as per 3.2 and 3.3, the list of $\mathcal{G}^{p_j}$ 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] In 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.

[4] For 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*.

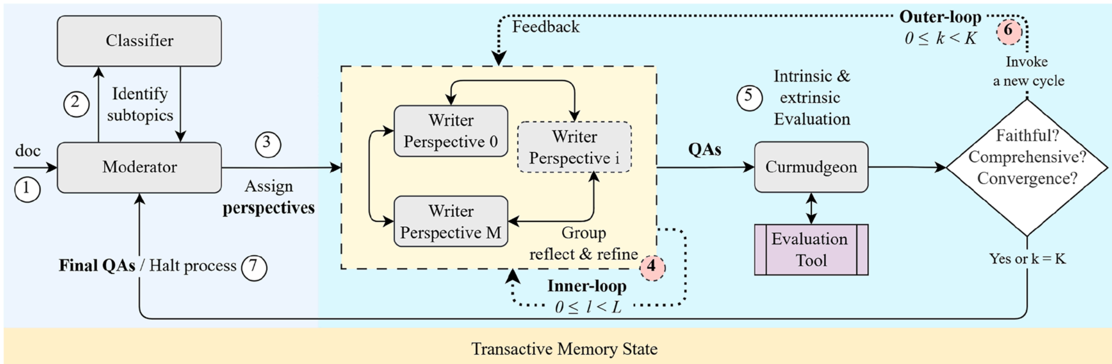

Fig. 1. CIR3 takes an input document (1), identifies subtopics (2), and prompts writer agents to generate QA pairs based on their assigned perspectives (subtopics) (3). The QAs undergo iterative refinement by the writers (4), followed by an outer refinement where the curmudgeon, using its intrinsic knowledge and 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.

Algorithm 1: Pseudo-algorithm describing the CIR3 conceptual framework. The loops are designed to terminate gracefully when either the feedback is empty (indicating satisfactory) or the number of iterations reaches the predefined threshold. Further implementation details are given in C: Algorithm Implementation Details.

Input : Max inner-refinement cycles L;
Max outer-refinement cycles K;
Max perspective M, Context c
Output: QA pairs $\mathcal{G}*$

$\mathcal{M} \leftarrow [\varepsilon\,\varepsilon]$ `// Writer's short memory state. (Eq. (2))`
$\mathcal{H} \leftarrow [\varepsilon\,\varepsilon]$ `// Long short-term memory state. (Eq. (3))`
`// Identify and assign unique perspectives` $\mathcal{P}$.
W_{P_0} = "default in-domain writer"
$W \leftarrow [W_{P_0}]$ `// List of Writers.`
$P \leftarrow$ `classify_subtopics`(c, M) `// List of subtopics` $\le M$.
foreach *subtopic in* P **do**
  W.append(`get_perspective_writer`(subtopic))
end
`// Outer-refinement cycles: evaluate and critic`
$k \leftarrow 0; \mathcal{F}'_{k+1} \leftarrow \varnothing$
do
  `// Inner-refinement cycles: generate, then refine.`
  $l \leftarrow 0$
  $\mathcal{F}_{l+1} \leftarrow \varnothing$
  do
    $\mathcal{G}^+{}_l \leftarrow$ generateQAs(c, $\mathcal{M}[-1]$, $\mathcal{H}[-1]$)
    $\mathcal{F}_{l+1} \leftarrow$ refineQAs(($\mathcal{G}^+{}_l, \mathcal{M}$))
    $\mathcal{M}$.append(($\mathcal{G}^+{}_l, \mathcal{F}_{l+1}$))
    $l++$
  while $l < L \wedge \mathcal{F}_{l+1} \neq \varnothing$;
  $\mathcal{G}^-{}_k \leftarrow \mathcal{G}^+{}_{l-1}$
  $\mathcal{F}'_{k+1} \leftarrow$ curmudgeonQAs(($\mathcal{G}^-{}_k, \mathcal{H}$))
  $\mathcal{H}$.append(($\mathcal{G}^-{}_k, \mathcal{F}'_{k+1}$))
  $k++$
while $k < K \wedge \mathcal{F}'_{k+1} \neq \varnothing$;
$\mathcal{G}^* \leftarrow \mathcal{G}^-{}_{k-1}$
return $\mathcal{G}^*$

3.2. Transactive reasoning

Momennejad [57] 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 (Fig. 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:

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

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 R3, we take inspiration from the phenomenon of *Collective Cognitive Convergence (C3)* [58,59] and from *How social network topology can shape collective cognition* [57]. 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 (Fig. 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 (Fig. 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 predefined 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 $\mathcal{F}\prime_k$ along $\mathcal{G}^-{}_{k-1}$ to the transactive memory, creating a central memory state that evolves with each outer-iteration of refinement:

$$\mathcal{H} = \{(\mathcal{G}^-{}_0, \mathcal{F}\prime_1), \dots, (\mathcal{G}^-{}_{k-2}, \mathcal{F}\prime_{k-1}), (\mathcal{G}^*, \emptyset)\}_{k=1}^{K} \tag{3}$$

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 (Fig. 1 ⑦).

The curmudgeon is equipped with an evaluation tool to help quantify 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] Friedman and Dieng [60] 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 evaluates $\mathcal{G}^-$ to determine the next action, leveraging the Vendi tool to augment its intrinsic knowledge and generate informed feedback. The result is then added to $\mathcal{H}$ (Eq. (3)), which invokes another cycle of inner-refinements among the writers.

3.3.1. Diversity metric selection

Vendi Score is a similarity-based diversity metric inspired by quantum statistical mechanics. It quantifies the diversity of a sample set by analyzing the eigenvalues of a similarity matrix, which captures the correlations and relationships among all data points. This approach enables Vendi Score to measure complex, multidimensional diversity that pairwise metrics may miss. Unlike many traditional diversity metrics, it does not require prevalence information on items, which makes it particularly useful when such data are unavailable or irrelevant. Vendi Score captures both the number and balance of distinct outputs, is less sensitive to sequence length variations, and operates independently of specific embedding model architectures. By evaluating the joint diversity across all attributes simultaneously, Vendi Score provides a comprehensive measure of a set's overall variability, going beyond simple pairwise comparisons to assess the full spectrum of diversity within the samples.

3.3.2. Convergence properties and limitations

CIR3 operates as a heuristic iterative framework without theoretical global optimality guarantees. However, several design safeguards mitigate common convergence issues: bounded iterations prevent infinite loops, the hybrid topology preserves diversity while enabling coordination, and the external variation signal (curmudgeon + Vendi tool) prevents premature consensus. Empirically, ablation studies (Section 5.5) demonstrate that removing these safeguards leads to either premature convergence with information loss or slow convergence with reduced faithfulness, validating our design choices.

3.3.3. Domain robustness

CIR3's architecture is intentionally domain-agnostic to ensure cross-domain generalizability through three key mechanisms:

1. **Subtopic identification** uses LLM semantic understanding without domain-specific ontologies, reducing dependence on specialized vocabularies.
2. **Perspective-assigned writers** leverage general reasoning capabilities based on textual evidence rather than domain-specific knowledge bases or specialized ontologies.
3. **Curmudgeon evaluation** integrates two complementary mechanisms: intrinsic reasoning capabilities for the coverage assessment and Vendi score-based diversity optimization for comprehensive exploration.

This integrated approach ensures both comprehensive document coverage and diverse solution exploration across domains. The semantic subtopic decomposition enables flexible problem partitioning, perspective-based writing generates diverse viewpoints through general reasoning, and the dual curmudgeon mechanism balances coverage optimization with diversity preservation to prevent convergence to suboptimal local solutions.

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 [61] and Natural Questions [62] 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 four specialized datasets in Finance and Medical fields:

(1) FiQA [63]. This dataset[6] was used in the Financial Opinion Mining and Question Answering challenge at the 2018 International World Wide Web Conference. FiQA comprises 6648 questions and 57,640 answer passages. It was curated from financial posts on platforms such as Stackexchange[7], Reddit[7], and StockTwits[7] between 2009 and 2017, with the objective of developing QA systems that can address financial queries by leveraging information from various sources such as microblogs, reports, and news articles;

(2) InsuranceQA [64] (InsurQA). This corpus[8] was sourced from the Insurance Library[9] 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.

(3) MedQA [65] is a free form multilingual multiple choice QA dataset[10] specifically curated for medical problem-solving, sourced from

[5] We employ SimCSE models from Princeton and BGE models as foundational encoders for the Vendi score. Our implementation extends this setup to include various embedders. Empirically, it 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
[7] https://stackexchange.com; https://stocktwits.com; https://reddit.com
[8] https://github.com/shuzi/insuranceQA
[9] https://www.insurancelibrary.com/
[10] https://huggingface.co/datasets/bigbio/med_qa

professional medical board exams. It encompasses over 61K questions, making it a valuable resource for evaluating and training models on clinical reasoning and medical knowledge in diverse contexts. The dataset is complemented by a large-scale corpus extracted from medical textbooks, supporting comprehensive reading comprehension and open-domain medical QA research.

(4) MedMCQA [66] is a large-scale, multiple-choice QA dataset[11] designed to mirror real-world medical entrance exams, notably AIIMS and NEET PG. It contains over 194K high-quality MCQs spanning 2400 healthcare topics and 21 medical subjects, with each question accompanied by detailed explanations. The dataset is notable for its topical diversity and complexity, requiring models to demonstrate advanced reasoning abilities across a broad spectrum of medical disciplines, making it a robust benchmark for open-domain medical QA systems.

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-INST[12] 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 [67] 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 implement CIR3 using the LangGraph[13] library, supporting both heterogeneous (CIR3-Het) and homogeneous (CIR3-Hom) configurations.

- **CIR3-Hom** uses the INSTRUCT variants of Meta-Llama-3-{70B,8B} for their strong performance and moderate resource requirements.
- **CIR3-Het**[14] leverages GPT-4o-mini[15], Gemma-3-27B-it[16], Meta-Llama-3-{70B,8B}, and Claude Sonnet 4[17].

Inference is conducted with a temperature of 0.1 and nucleus sampling of 0.5. We use the Groq[18] API for Llama models and self-host Gemma via vLLM[19] Kwon et al. [68], both offering seamless integration. Generation is limited to 10 QA pairs per context, with refinement iterations set to $K = 6$ and $L = 12$.

4.4. Evaluation metrics

This section delineates the metrics and evaluation framework used to assess CIR3's performance. We first discuss the primary evaluation, then address common generation errors, and finally, we discuss our approach to evaluating the classifier agent through cross-model agreement.

4.4.1. Main evaluation

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 [69], METEOR [70], and Jaccard Index [71] 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 token overlap, we also use embedding-based similarity metrics, such as BERTScore [72] 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 [73] framework by merging the task definition and evaluation criteria prompt with a Chain-of-Thoughts (CoT) prompt [74] 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 scoring calculations used in this study are provided in Appendix A: Automatic Metrics. Sample prompts designed for evaluating CIR3 can be found in Appendix E: Evaluation Prompts.

4.4.2. Evaluation of common generation errors

To further assess the robustness of our framework, we evaluate CIR3 in its ability to mitigate common generation errors: *hallucination*, *irrelevance*, *duplication*, and *over-specificity*. We conduct experiments using 400 samples, with 100 samples from each of the four datasets. The generated QA pair sets are evaluated using gpt-4o as a model-based evaluator, employing G-Eval with detailed evaluation steps for each error type. To quantify duplication, we assess the semantic similarity across all possible pairs of generated questions for a given document, and report the averaged score. We compare the performance of our approach, **CIR3-Hom** and **CIR3-Het**, against the two baselines: **LLM-DP** and **qGen-aGen**.

4.4.3. Evaluating classifier agents via cross-model agreement

This section details the quantitative evaluation of our classifier agent's accuracy and robustness in identifying subtopics within FiQA, InsurQA, MedQA and MedMCQA. Given the absence of a human-annotated gold standard, we employ a two-pronged approach for cross-model agreement: Pairwise Semantic Agreement and Holistic Semantic Agreement (Krippendorff's Alpha), leveraging consensus among multiple advanced LLMs.

(1) Pairwise Semantic Agreement. For each document, our classifier's identified subtopics are compared against those generated by a set of cross-LLM models (GPT-4o-mini, Gemma-3-27b-it, Claude Sonnet 4). We compute agreement using four distinct metrics:

[11] https://huggingface.co/datasets/openlifescienceai/medmcqa
[12] https://ai.meta.com/blog/meta-llama-3
[13] https://langchain-ai.github.io/langgraph
[14] The implementation is designed for flexibility and scalability, enabling compatibility with a wide range of LLMs through external configuration alone, without requiring code modifications, as detailed in our repository.
[15] https://platform.openai.com/docs/models/gpt-4o-mini
[16] https://deepmind.google/models/gemma/gemma-3
[17] https://www.anthropic.com/news/claude-4
[18] https://groq.com
[19] Our repository provides a scalable, containerized vLLM setup with integrated monitoring tools, enabling streamlined deployment and robust observability.

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	0.2252	0.1951	0.3189	0.2781	0.2640	0.4377	0.5286	0.4881	0.4847
	QGen-AGen	0.1288	0.3383	0.1613	0.2094	0.1771	0.4003	0.2690	0.2821	0.4161	0.5391	0.4703	0.4751
	CIR3-Hom	0.1935	0.3791	0.2767	**0.2831**†	0.2153	0.3771	0.2893	**0.2939**	0.5511	0.6112	0.5983	**0.5868**†
	CIR3-Het	0.2140	0.3905	0.2938	**0.2994**†	0.2190	0.3913	0.2952	**0.3018**	0.5633	0.6959	0.5983	**0.6191**†
InsurQA	LLM-DP	0.2422	0.3972	0.2717	0.3037	0.2877	0.4984	0.3447	0.3769	0.4784	0.5920	0.4987	0.5230
	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-Hom	0.3197	0.4391	0.3632	**0.3739**†	0.2950	0.4891	0.3972	**0.3937**	0.5261	0.6716	0.6104	**0.6027**†
	CIR3-Het	0.3278	0.4579	0.3802	**0.3887**†	0.3143	0.5002	0.4031	**0.4059**	0.5354	0.6769	0.6210	**0.6111**†
MedQA MedMCQA	LLM-DP	0.1506	0.3042	0.2077	0.2208	0.1959	0.3161	0.2715	0.2611	0.4351	0.5303	0.4814	0.4823
	QGen-AGen	0.1231	0.3353	0.1575	0.2053	0.1708	0.2982	0.2708	0.2466	0.4136	0.5390	0.4680	0.4735
	CIR3-Hom	0.1958	0.3847	0.2765	**0.2857**†	0.2204	0.3739	0.2887	**0.2943**	0.5512	0.6146	0.6019	**0.5892**†
	CIR3-Het	0.2070	0.3934	0.2766	**0.3739**†	0.2317	0.3798	0.3050	**0.3055**	0.5582	0.6220	0.6104	**0.6212**†

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	0.8571	0.6858	0.6847	0.7872	0.7192	0.7548	0.8078	0.8488	0.8038
	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-Hom	0.8702	0.9171	0.9088	**0.8987**†	0.8312	0.8542	0.8115	**0.8323**	0.8291	0.9118	0.9264	**0.8891**†
	CIR3-Het	0.9085	0.9412	0.9378	**0.9292**†	0.8555	0.8689	0.8407	**0.8551**	0.8554	0.9384	0.9587	**0.9175**†
InsurQA	LLM-DP	0.8511	0.8810	0.8779	0.8700	0.7388	0.7540	0.8097	0.7675	0.8173	0.8948	0.8675	0.8598
	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-Hom	0.8972	0.9298	0.9175	**0.9148**†	0.7591	0.7736	0.8616	**0.7980**	0.8450	0.9395	0.9072	**0.8972**†
	CIR3-Het	0.9218	0.9414	0.9352	**0.9328**†	0.7833	0.8008	0.8809	**0.8217**	0.8389	0.9530	0.9215	**0.9044**†
MedQA MedMCQA	LLM-DP	0.8358	0.8525	0.8696	0.8526	0.6850	0.6795	0.7848	0.7165	0.7486	0.8039	0.8483	0.8003
	QGen-AGen	0.7845	0.8633	0.8347	0.8275	0.6528	0.6734	0.6869	0.6710	0.7288	0.7540	0.7770	0.7533
	CIR3-Hom	0.8776	0.9256	0.9235	**0.9089**†	0.7777	0.7592	0.8511	**0.7960**	0.8559	0.9292	0.8851	**0.8901**†
	CIR3-Het	0.8971	0.9454	0.9525	**0.9316**†	0.8053	0.7835	0.8708	**0.8199**	0.8787	0.9552	0.9056	**0.9132**†

- **Jaccard Similarity**: Measures the overlap of unique subtopics.
- **Soft-F1 Score**: A semantic F1 score, measuring precision and recall based on cosine similarity between subtopic embeddings.
- **Bipartite Matching (Hungarian Algorithm-based F1)**: Employs the Hungarian algorithm to find the optimal one-to-one semantic mapping between subtopics from two lists, maximizing the sum of cosine similarities above a threshold.
- **Average Cosine Similarity (Avg-Cosine)**: The average of all pairwise cosine similarities between subtopic embeddings from two lists.

(2) Holistic Semantic Agreement (Krippendorff's Alpha). Our framework utilizes a two-stage methodology to rigorously assess the reliability of semantic consensus between all annotators and our CIR3 classifier.

In **Stage 1**, we compute Krippendorff's Alpha over semantically clustered topics, derived from subtopics identified by all consensus models. These subtopics are encoded using sentence transformer embeddings and clustered via agglomerative methods with cosine distance. We employ a range of distance thresholds to ensure optimal semantic granularity, balancing the risk of merging distinct concepts against fragmenting related subtopics. The resulting Alpha scores, averaged across thresholds, provide a robust baseline for consensus reliability.

Stage 2 involves evaluating CIR3's semantic agreement with the established consensus using pairwise semantic metrics, thereby validating the classifier's alignment with the consensus.

Further methodological details are given in Appendix B: Classifier Agent: Cross-Model Agreement.

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.

As shown in Table 1, our proposed approaches outperform both baselines across all lexical metrics on both datasets. Specifically, CIR3-Hom achieves relative improvements of 6.43 % on METEOR, 2.66 % on ROUGE-L, and 9.62 % on the Jaccard Index over the next best-performing model. Furthermore, CIR3-Het demonstrates even more substantial gains, with relative improvements of 10.41 %, 3.70 %, and 12.04 % on the same metrics, respectively. 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.

Further analysis in Table 2 shows that CIR3 consistently surpasses other models in semantic similarity metrics. CIR3 achieves an average improvement of 5.94 % on BERTScore and 8.33 % 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	0.7599
	QGEN-AGEN	0.5290	0.8414	0.6852
	CIR3-Hom	0.9312	0.9762	**0.9537**
	CIR3-Het	0.9419	0.9749	**0.9584**
INSURQA	LLM-DP	0.7317	0.8175	0.7746
	QGEN-AGEN	0.5560	0.8763	0.7161
	CIR3-Hom	0.9389	0.9879	**0.9634**
	CIR3-Het	0.9501	0.9893	**0.9697**
MEDQA MEDMCQA	LLM-DP	0.6807	0.7961	0.7384
	QGEN-AGEN	0.5052	0.8371	0.6711
	CIR3-Hom	0.9148	0.9511	**0.9329**
	CIR3-Het	0.9372	0.9629	**0.9500**

Table 4
LLM-based evaluation results for common generation errors (*semantic duplication*, *hallucinated answers*, *irrelevant QAs*, *over-specific* and *over-generalized answers*). Higher scores indicate better performance.

Dataset	Model	Duplication	Hallucination	Irrelevance	Over Specificity
FIQA INSURQA	LLM-DP	0.7846	0.8088	0.8006	0.8976
	QGEN-AGEN	0.4771	0.8519	0.8481	**0.9781**
	CIR3-Hom	**0.9515**	**0.9796**	**0.9783**	0.9549
	CIR3-Het	**0.9689**	**0.9853**	**0.9825**	0.9533
MEDQA MEDMCQA	LLM-DP	0.7012	0.7935	0.7933	0.8860
	QGEN-AGEN	0.4919	0.8466	0.8452	**0.9703**
	CIR3-Hom	**0.9317**	**0.9690**	**0.9630**	0.9561
	CIR3-Het	**0.9481**	**0.9811**	**0.9847**	0.9598

Table 3 (LLM-based evaluation results) further supports these findings, showing that CIR3-HOM achieves average improvements of 21.85 % in comprehensiveness and 16.62 % in faithfulness, while CIR3-HET attains 23.33 % and 17 % improvements, respectively, outperforming the second-best model, LLM-DP. These results provide additional validation for Method 3.3 and Eq. (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 [75] model in uncovering deeper key concepts in financial and medical 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 1–3, provide compelling evidence of the effectiveness of CIR3. These results further demonstrate that CIR3-HET consistently outperforms CIR3-HOM, a performance gap attributable to the advantages conferred by architectural and behavioral diversity among heterogeneous agents, where differing model architectures, training data, and inductive biases result in varied reasoning strategies and error profiles, enabling the system to explore a broader solution space and mitigate redundancy or shared failure modes common in homogeneous configurations. This diversity enhances robustness through inter-agent disagreement, encourages complementary specialization, and supports more effective ensemble decision-making. Recent work highlights how heterogeneity improves collaborative reasoning, debate, and problem-solving performance in LLM-based agents [76]. Such findings also align with our hypotheses in **H.2** and **H.3**, where cognitive diversity often correlates with improved group performance.

5.2. Common generation error analysis

The results presented in Table 4 demonstrate that both CIR3 variants significantly outperform the baseline methods across all error categories and domains.CIR3-Het achieves the highest performance, with scores exceeding 0.94 across all metrics for both Finance and Medical domains.

Duplication Analysis. The most substantial performance gap appears in duplication scores, where CIR3 achieves scores between 0.93 and 0.97, while LLM-DP scores range from 0.70-0.79, and qGen-aGen performs poorly with scores below 0.5. This suggests that T5-based question generation tends to produce highly similar questions, limiting diversity in the generated QAs.

Over-specificity. Interestingly, qGen-aGen scores high in over-specificity (~ 0.97). This can be attributed to the characteristics of the fine-tuned T5 question generation model (query-gen-ms-marco-t5-large-v1), which tends to produce more generic, template-driven questions. Although this approach helps prevent overfitting to specific contextual details, it comes at the steep cost of diversity, as evidenced by the low duplication scores.

Hallucination and Irrelevance Control. CIR3 variants exhibit superior hallucination prevention compared to baseline methods, with CIR3-Het achieving scores of 0.98 +, substantially higher than LLM-DP (~ 0.80) and baseline qGen-aGen (~ 0.84).

The results demonstrate that CIR3's approach to QAG effectively balances all evaluation criteria, producing high-quality, diverse, and contextually appropriate QAs while minimizing common generation errors.

5.3. Classifier agent: cross-model agreement

Table 5 presents the evaluation of our classifier agent using multiple agreement metrics. Soft-F1 scores are particularly high (0.9141) for the finance domain and 0.9157 for the medical domain, indicating strong semantic alignment between our classifier agent and the reference models, even when subtopic labels differ in wording or structure.Krippendorff's Alpha, which captures overall agreement beyond chance, also shows strong results (0.9338 for finance and 0.9207 for medical), validating the robustness of our consensus establishment process. The CIR3 classifier demonstrates excellent alignment with the consensus between models in both domains. In the finance domain, CIR3 achieves a semantic agreement score of 0.9434 ($\sigma = 0.0491$), while in the medical domain it achieves 0.9316 ($\sigma = 0.0473$). These scores indicate a strong semantic alignment between the CIR3 classifications and the consensus baseline. The slightly higher agreement in the finance domain likely reflects the more controlled and domain-specific vocabulary typical of financial texts. Conversely, Jaccard scores are lower (0.55 and 0.54), reflecting the stricter nature of this metric in requiring exact token overlap, which can penalize semantically equivalent but differently phrased outputs. This contrast highlights the importance of using soft or semantic-aware metrics when evaluating classification tasks that involve natural language.

Overall, these results validate both the effectiveness of our classifier agent and our general approach to scalable subtopic classification in scenarios where manual annotations are unavailable or impractical.

5.3.1. Classifier performance via bipartite agreement scores

To evaluate the consistency of our subtopic prediction, we analyze the distribution of Bipartite F1 scores across 400 documents sampled uniformly from the four datasets (100 each). We select the bipartite score due to its higher standard deviation (as shown in Table 5), which makes it particularly sensitive to performance variation and thus a suitable measure for evaluating classifier agreement.

Our analysis, summarized in Table 6 and Fig. 2, revealed robust semantic alignment between CIR3's classifier agent and the reference models. A significant majority of documents (74 %) achieved scores in

Table 5
Cross-model agreement metrics for subtopic identification.

Metric	FiQA/InsurQA	MedQA/MedMCQA
Pairwise Semantic Agreement (Mean ± Std Dev)		
Jaccard	0.5530 ± 0.2819	0.5403 ± 0.2938
Soft-F1	0.9161 ± 0.1060	0.9157 ± 0.0993
Bipartite	0.8388 ± 0.1596	0.8198 ± 0.1538
Avg-Cosine	0.7093 ± 0.0368	0.7074 ± 0.0332
Holistic Semantic Agreement		
Cross-Model Reliability (α)	0.9338	0.9207
CIR3 Semantic Agreement	0.9434 ± 0.0491	0.9316 ± 0.0473

Table 6
Statistics of classifier bipartite F1 scores across 400 documents.

Score Interval	Doc Count (%)	Avg Score	Median	Min	Max	CIR3 Comp	Faith
[0.50, 0.66]	4	0.600	0.600	0.600	0.600	0.880	0.923
[0.66, 0.75]	22	0.724	0.733	0.667	0.733	0.918	0.958
[0.75, 1.00]	74	0.875	0.867	0.800	1.000	0.949	0.994

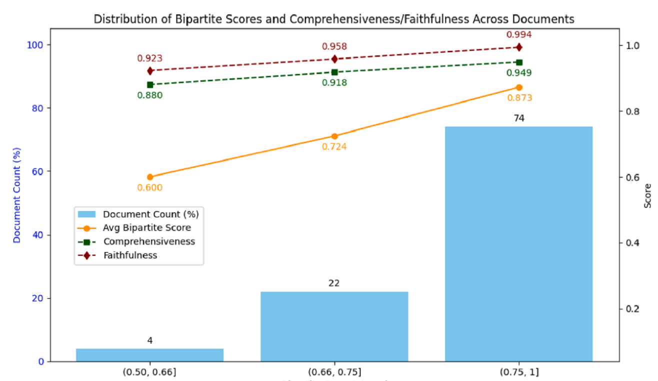

Fig. 2. A bar plot depicts the distribution of 400 documents across score intervals based on their Bipartite F1 scores.

the highest agreement interval (0.75, 1]. The middle interval (0.66, 0.75] accounted for 23 % of documents, while a small fraction (4 %) scored 0.6. Notably, no document scored below 0.6, which confirms a consistent baseline of semantic agreement and underscores the strong reliability of the CIR3 classifier.

Importantly, system-level metrics, Comprehensiveness and Faithfulness, demonstrate a monotonic improvement as bipartite F1 scores increased. In the top interval, these metrics reached 0.949 and 0.994, respectively. This direct correlation confirms that more accurate subtopic classification directly enhances overall CIR3 performance. Consequently, the bipartite F1 score proves to be a reliable indicator of meaningful agreement and its positive impact on CIR3's end-task performance.

5.4. 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 finance[20] to assess 10 sets of QA pairs each, focusing on comprehensiveness and faithfulness. Comprehensiveness is evaluated based 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).

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

Table 7
Human evaluation results on 80 sets of QA pairs generated by CIR3 and LLM-DP. The ratings (1 to 5) are normalized between 0 and 1. The scores are analyzed using a paired t-test (p-values are presented).

	Aspect	LLM-DP	CIR3	p-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**	

A partial excerpt of the evaluation guidelines is given in Appendix D: Human Evaluation Guidelines.

Table 7 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.

5.5. 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 Shao et al. [51], 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.

Table 8
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

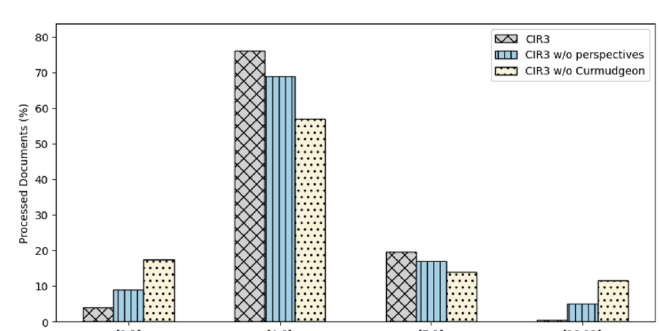

Fig. 3. Number of inner-refinement cycles (x-axis), given as intervals, required to process the input documents (y-axis), given as percentage.

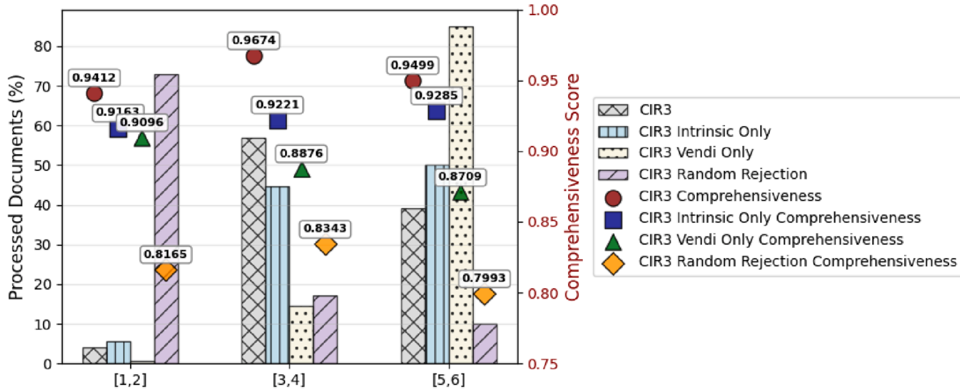

Fig. 4. Comparative analysis of four CIR3 variants, distinguished by their Curmudgeon strategies, across defined outer-refinement cycle ranges. Bars show the percentage of documents processed within each cycle range (left y-axis), and scatter points denote the average comprehensiveness score (right y-axis).

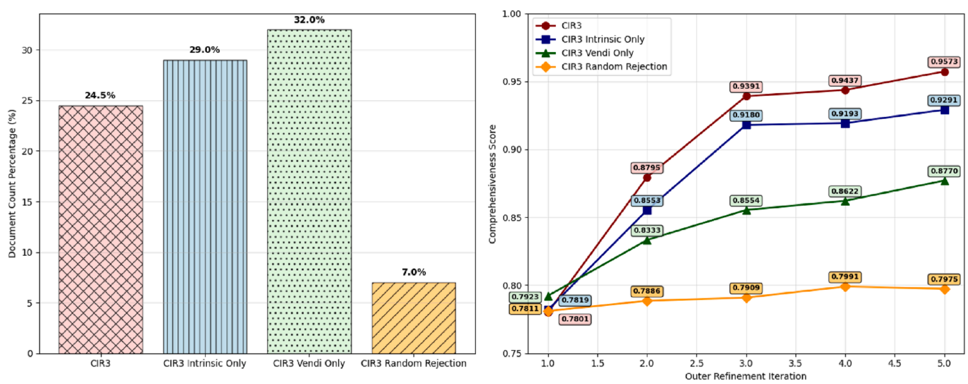

Fig. 5. To analyze the behavior of different Curmudgeon strategies, this figure breaks down performance on documents requiring exactly five outer refinement iterations, showing the percentage of such documents (left) and the corresponding evolution of their comprehensiveness score (right).

The results in Table 8 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 8 shows that *CIR3 w/o perspectives* yields inferior results compared to CIR3, suggesting that multi-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 Fig. 3. 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 $[10, 12]$, as shown in Table 9. 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.

Table 9
Effect of multi-perspective reasoning and curmudgeon on document distribution per cycle.

Model	Context Distribution Per Refinement Cycle Ranges (%) $[0, 3]$	$[4, 6]$	$[7, 9]$	$[10, 12]$
LLM-DP	100	–	–	–
CIR3	04.00	76.00	19.50	00.50
CIR3 w/o perspectives	09.00	69.00	17.00	05.00
CIR3 w/o Curmudgeon	17.50	57.00	14.00	11.50

5.5.1. Ablation studies: curmudgeon strategies

To evaluate the individual contributions of the curmudgeon agent and the Vendi diversity tool in CIR3, we conduct comprehensive ablation studies under four experimental conditions using 200 documents uniformly sampled from the four datasets (50 each):

1. **CIR3 (baseline)**: The curmudgeon agent combines its intrinsic knowledge with the Vendi diversity tool.
2. **CIR3 Intrinsic Only**: The curmudgeon agent operates solely using its intrinsic knowledge without access to the Vendi diversity tool.
3. **CIR3 Vendi Only**: A simplified configuration using only the Vendi tool for diversity measurement (binary feedback[21]), with no curmudgeon agent providing qualitative feedback.
4. **CIR3 Random Rejection**: A control condition employing random feedback with an acceptance probability of 0.35, eliminating both agent reasoning and diversity measurement.

All configurations utilizing LLM agents employed GPT-4o-mini for its superior reasoning capabilities and cost-efficiency. Ablation studies reveal distinct behavioral patterns across the four configurations, measured by document distribution across outer refinement iteration intervals and corresponding comprehensiveness scores.

[21] "*QA pair meets or not diversity and / or alignment criteria.*"

Table 10
Ablation study of Curmudgeon feedback strategies.

Strategy	Avg. Comp↑	Avg. Faith↑	Avg. cycles↓	% docs processed in: [1,2]	[3,4]	[5,6]
CIR3	0.9528	0.9895	2.35	4.0	57.0	39.0
CIR3 Intrinsic Only	0.9223	0.9696	2.45	5.5	44.5	50.0
CIR3 Vendi Only	0.8893	0.9331	2.85	0.5	14.5	85.0
CIR3 Random Rejection	0.8167	0.8969	1.37	73.0	17.0	10.0

Table 11
Document count percentage and comprehensiveness score progression for documents requiring exactly 5 outer refinement iterations.

Model Variant	Doc %	Iter 1	Iter 2	Iter 3	Iter 4	Iter 5	Improvement
CIR3	24.5 %	0.7801	0.8795	0.9391	0.9437	0.9573	+ **0.1772**
CIR3 Intrinsic Only	29.0 %	0.7819	0.8553	0.9180	0.9193	0.9291	+ 0.1472
CIR3 Vendi Only	32.0 %	0.7923	0.8333	0.8554	0.8622	0.8770	+ 0.0847
CIR3 Random Rejection	7.0 %	0.7811	0.7886	0.7909	0.7991	0.7975	+ 0.0164

Comprehensiveness Performance. As shown in Table 10, the *CIR3* baseline achieved the highest average comprehensiveness score at 0.9528, demonstrating the effectiveness of combining the reasoning of the curmudgeon agent with the diversity measurement tool. *CIR3 Intrinsic Only* performed moderately well (0.9223) but showed reduced quality without diversity guidance. *CIR3 Vendi Only* achieved lower comprehensiveness (0.8893), suggesting that diversity measurement alone is insufficient for an effective quality assessment. The *random rejection* baseline predictably performed the lowest (0.8167), reinforcing the need for intelligent feedback mechanisms.

Document Distribution Patterns. As illustrated in Fig. 4, the distribution of documents across refinement intervals [1,2], [3,4], and [5,6] revealed distinct processing patterns. Random rejection processed 73 % of documents within the first two cycles, indicating premature convergence (Comprehensiveness 0.8165). In contrast, *CIR3 Vendi Only* required extensive refinement, with 85 % of the documents needing 5–6 cycles, indicating slow convergence (Comprehensiveness 0.8709). The original *CIR3* demonstrated a more balanced convergence, with document distributions of 4 %, 57 % and 39 % across the [1,2], [3,4], and [5,6] intervals, achieving respective comprehensiveness scores of 0.9412, 0.9674, and 0.9499.

Complex Document Analysis. For the most challenging documents[22], which required exactly five refinement iterations ($K - 1$), *CIR3* again showed superior performance, with a significant improvement from 0.7801 to 0.9573 (+ 17.72 % - Table 11). In contrast, *CIR3 Vendi Only* exhibited the smallest improvement (0.7923 → 0.8770), as detailed in Fig. 5. This highlights the critical role of intelligent agent feedback in difficult refinement tasks.

Key Finding. The ablation results demonstrate that:

(1) the intrinsic reasoning capabilities of the curmudgeon agent are the main driver of quality improvements,
(2) the Vendi diversity tool provides measurable enhancement when combined with the reasoning of the agent,
(3) neither component alone achieves the performance of the integrated system, and
(4) the substantial performance gap between random rejection and all other conditions validates the importance of intelligent feedback in iterative refinement processes.

These findings confirm the synergistic benefit of integrating both components in CIR3. This combination outperforms either element in isolation while achieving balanced convergence, which is reflected in the high scores.

[22] These were among the most challenging documents, processed in five iterations, just below the maximum limit ($K = 6$) used to halt refinement.

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. This work also presents a cross-model semantic agreement framework that evaluates LLM subtopic identification through multi-model consensus rather than human annotation. Using Soft-F1, bipartite matching, and Krippendorff's Alpha on semantically clustered topics, it provides scalable evaluation for classifiers in low-resource domains. Empirical results confirm CIR3's significant performance gains over strong baselines.

Future research aims to broaden the applicability of CIR3 across a diverse range of tasks, such as summarization, information retrieval, and multi-modal applications.

CRediT authorship contribution statement

Sami Saadaoui: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization; **Eduardo Alonso:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Data availability

The link is provided in the paper.

Declaration of competing interest

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.

Appendix A. Metrics

A.1. Automatic metrics

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

ROUGE-L [69] assesses recall by evaluating the overlap between reference and generated sentences using Longest Common Subsequence statistics. We use the implementation from GOOGLE[23]. In this paper, we report the F1 score, the harmonic mean of precision and recall.

METEOR [70] is a recall-oriented metric that measures the similarity between generated and reference text, incorporating synonyms, stemming, and paraphrasing. We use the implementation from NLTK[24].

Jaccard Index[25] [71] 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 implementation[26].

BERTScore[27] [72] 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-large[28] 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.

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:

$$s(c, \mathcal{Q}) = \frac{1}{N} \sum_{i=1}^{N} s(c, q_i) \tag{A.1}$$

$$s(c, \mathcal{A}) = \frac{1}{N} \sum_{i=1}^{N} s(c, a_i) \tag{A.2}$$

$$s(\mathcal{Q}, \mathcal{A}) = \frac{1}{N} \sum_{i=1}^{N} s(q_i, a_i) \tag{A.3}$$

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

$$s(c, \mathcal{A}_\oplus) = s(c, \oplus_{i=1}^{N} a_i) \tag{A.5}$$

$$s(\mathcal{Q}_\oplus, \mathcal{A}_\oplus) = s(\oplus_{i=1}^{N} q_i, \oplus_{i=1}^{N} a_i) \tag{A.6}$$

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

Appendix B. Classifier agent: cross-model agreement

This section provides further methodological details for the two distinct evaluation approaches used to assess the performance of the CIR3 classifier: (1) direct pairwise evaluation and (2) holistic consensus-based evaluation.

[23] https://pypi.org/project/rouge-score
[24] https://www.nltk.org
[25] https://en.wikipedia.org/wiki/Jaccard_index
[26] https://scikit-learn.org
[27] https://github.com/Tiiiger/bert_score
[28] https://github.com/FlagOpen/FlagEmbedding

B.1. Subtopic representation: semantic embeddings

Each subtopic s is represented as a dense vector embedding using the BAAI/bge-large-en-v1.5 sentence transformer model:

$$E(s) = \text{SentenceTransformer}_{\text{BAAI/bge-large-en-v1.5}}(s) \in \mathbb{R}^{1024} \tag{B.1}$$

The cosine similarity between embeddings is computed as:

$$\text{sim}(s_i, s_j) = \frac{E(s_i) \cdot E(s_j)}{\|E(s_i)\|_2 \|E(s_j)\|_2} \tag{B.2}$$

B.2. Approach 1: direct pairwise evaluation

This approach directly compares the CIR3 subtopics with each consensus model using four complementary metrics. Given CIR3 subtopics $L_{\text{CIR3}} = \{s_{1,i}\}_{i=1}^{m}$ and consensus model subtopics $L_{\text{consensus}} = \{s_{2,j}\}_{j=1}^{n}$:

A. Jaccard Similarity (exact string matching baseline):

$$J(L_{\text{CIR3}}, L_{\text{consensus}}) = \frac{|L_{\text{CIR3}} \cap L_{\text{consensus}}|}{|L_{\text{CIR3}} \cup L_{\text{consensus}}|} \tag{B.3}$$

B. Soft-F1 Score with semantic similarity threshold θ:

$$P_{\text{soft}} = \frac{1}{m} \sum_{i=1}^{m} \mathbb{I}\left(\max_j \text{sim}(s_{1,i}, s_{2,j}) \geq \theta \right) \tag{B.4}$$

$$R_{\text{soft}} = \frac{1}{n} \sum_{j=1}^{n} \mathbb{I}\left(\max_i \text{sim}(s_{1,i}, s_{2,j}) \geq \theta \right) \tag{B.5}$$

$$F1_{\text{soft}} = \begin{cases} \dfrac{2 \cdot P_{\text{soft}} \cdot R_{\text{soft}}}{P_{\text{soft}} + R_{\text{soft}}} & \text{if } P_{\text{soft}} + R_{\text{soft}} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{B.6}$$

C. Bipartite Matching (Hungarian Algorithm) using cost matrix:

$$C_{ij} = 1 - \text{sim}(s_{1,i}, s_{2,j}) \tag{B.7}$$

Optimal assignment π^* minimizes total cost:

$$\pi^* = \arg\min_{\pi} \sum_{i=1}^{\min(m,n)} C_{i,\pi^*(i)} \tag{B.8}$$

Valid matches $VM = \sum_{i=1}^{\min(m,n)} \mathbb{I}[\text{sim}(s_{1,i}, s_{2,\pi^*(i)}) \geq \theta]$ yield:

$$P_{\text{bip}} = \frac{VM}{m}, \quad R_{\text{bip}} = \frac{VM}{n}, \quad F1_{\text{bip}} = \frac{2 \cdot P_{\text{bip}} \cdot R_{\text{bip}}}{P_{\text{bip}} + R_{\text{bip}}} \tag{B.9}$$

D. Average Cosine Similarity:

$$\text{AvgCos}(L_{\text{CIR3}}, L_{\text{consensus}}) = \frac{1}{m \cdot n} \sum_{i=1}^{m} \sum_{j=1}^{n} \text{sim}(s_{1,i}, s_{2,j}) \tag{B.10}$$

B.3. Approach 2: holistic two-stage consensus evaluation

This approach establishes the consensus between models through semantic clustering and then evaluates CIR3 against this consensus baseline.

B.3.1. Stage 1: consensus establishment via semantic clustering

For each document d, all subtopics from all consensus models are pooled:

$$S_d = \bigcup_{m \in \{\text{LLMs}\}} S_d^m \tag{B.11}$$

Agglomerative clustering is applied using cosine distance with average linkage:

$$d_{\text{cosine}}(E(s_i), E(s_j)) = 1 - \text{sim}(s_i, s_j) \tag{B.12}$$

$$d_{\text{avg}}(C_p, C_q) = \frac{1}{|C_p| \cdot |C_q|} \sum_{s_i \in C_p} \sum_{s_j \in C_q} d_{\text{cosine}}(E(s_i), E(s_j)) \tag{B.13}$$

Clustering proceeds with distance thresholds $\tau \in \{0.30, 0.35, 0.40, 0.45, 0.50\}$ to ensure optimal semantic granularity. Each cluster represents a conceptual topic, forming the consensus baseline.

Consensus Reliability Assessment: Krippendorff's Alpha is computed over the resulting conceptual topic clusters:

$$\alpha = 1 - \frac{D_o}{D_e} \tag{B.14}$$

where D_o is observed disagreement and D_e is expected disagreement under independence. The final reliability score averages across threshold configurations:

$$\alpha_{\text{final}} = \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} \alpha_\tau \tag{B.15}$$

B.3.2. Stage 2: CIR3 evaluation against consensus

CIR3's semantic agreement with the established consensus is evaluated using two complementary pairwise semantic metrics:

A. Soft-F1 Evaluation: CIR3 subtopics are compared against consensus-derived conceptual topics using Soft-F1 scoring (Eqs. (B.4)–(B.6)). For each threshold configuration, Soft-F1 scores are computed and averaged:

$$\text{Soft-F1 Agreement}_{\text{final}} = \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} F1_{\text{soft}}(\tau) \tag{B.16}$$

B. Bipartite Matching Evaluation: CIR3 subtopics are optimally matched against consensus conceptual topics using the Hungarian algorithm (Eqs. (B.7)–(B.9)). For each threshold configuration, bipartite F1 scores are computed and averaged:

$$\text{Bipartite F1 Agreement}_{\text{final}} = \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} F1_{\text{bip}}(\tau) \tag{B.17}$$

The holistic evaluation provides two complementary consensus-validated semantic agreement scores that account for different matching strategies: Soft-F1 allows multiple matches per subtopic while bipartite matching enforces one-to-one optimal assignment (Algorithm 2).

B.4. Algorithm: holistic two-stage evaluation

Algorithm 2: Cross-Model Semantic Evaluation Framework.

Data: Document set D, consensus models $\mathcal{M} = \{$Claude Sonnet, GPT-4o-mini, Gemma-27B$\}$, CIR3 classifier, thresholds $\mathcal{T} = \{0.3, 0.35, 0.4, 0.45, 0.5\}$
Result: Cross-model reliability α_{final}, CIR3 semantic agreement scores

Stage 1: Consensus Establishment;
foreach *document* $d \in D$ **do**
  $S_d \leftarrow \bigcup_{m \in \mathcal{M}} S_d^m$ `// Pool all subtopics (Eq. (B.11))`
  $E_d \leftarrow \{E(s) : s \in S_d\}$ `// Compute embeddings (Eq. (B.1))`
  foreach *threshold* $\tau \in \mathcal{T}$ **do**
    $C_d(\tau) \leftarrow \text{AgglomerativeClustering}(E_d, \tau)$ `// Eq. (B.12), B.13`
    $\alpha_d(\tau) \leftarrow \text{KrippendorffAlpha}(C_d(\tau), \mathcal{M})$ `// Eq. (B.14)`
  end
end
$\alpha_{\text{final}} \leftarrow \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} \frac{1}{|D|} \sum_{d \in D} \alpha_d(\tau)$ `// Eq. (B.15)`
Stage 2: CIR3 Evaluation;
foreach *document* $d \in D$ **do**
  $S_{\text{CIR3}}^d \leftarrow \text{CIR3}(d)$ `// Get CIR3 subtopics`
  foreach *threshold* $\tau \in \mathcal{T}$ **do**
    $S_{\text{consensus}}^d(\tau) \leftarrow \text{GetClusterRepresentatives}(C_d(\tau))$;
    $F1_{\text{soft}}^d(\tau) \leftarrow \text{SoftF1}(S_{\text{CIR3}}^d, S_{\text{consensus}}^d(\tau))$ `// Eq. (B.6)`
    $F1_{\text{bip}}^d(\tau) \leftarrow \text{BipartiteF1}(S_{\text{CIR3}}^d, S_{\text{consensus}}^d(\tau))$ `// Eq. (B.9)`
  end
end
$\text{Soft-F1}_{\text{final}} \leftarrow \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} \frac{1}{|D|} \sum_{d \in D} F1_{\text{soft}}^d(\tau)$ `// Eq. (B.16)`
$\text{Bipartite-F1}_{\text{final}} \leftarrow \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} \frac{1}{|D|} \sum_{d \in D} F1_{\text{bip}}^d(\tau)$ `// Eq. (B.17)`

B.5. Implementation parameters

- **Embedding Model:** BAAI/bge-large-en-v1.5 (1024-dimensional);
- **Clustering:** Agglomerative with cosine distance, average linkage;
- **Distance Thresholds:** $\tau = \{0.3, 0.35, 0.4, 0.45, 0.5\}$;
- **Similarity Threshold:** $\theta = 0.7$ for semantic matching;
- **Consensus Models:** GPT-4o-mini, Google/Gemma-27B-IT, Claude Sonnet 4;
- **Alpha Computation:** `simpledorff`[29] library for nominal data;

B.6. Interpretation guidelines

Approach 1 (Pairwise): Provides direct comparison metrics between CIR3 and individual consensus models using Eqs. (B.3), (B.6), (B.9), and (B.10). Higher values across all four metrics indicate stronger agreement.

Approach 2 (Holistic): Establishes validated consensus baseline ($\alpha \geq 0.8$ indicates excellent reliability per Eq. (B.14)) then measures CIR3's semantic alignment using Soft-F1 (Eq. (B.16)) and bipartite matching (Eq. (B.17)). This approach accounts for cross-model variability and provides consensus-validated evaluation with complementary matching strategies.

Soft-F1 vs. Bipartite Matching: Soft-F1 allows flexible many-to-many semantic matching, while Bipartite matching enforces optimal one-to-one assignment. Both metrics provide different perspectives on the quality of semantic alignment.

Complementary Nature: Approach 1 offers granular pairwise insights while Approach 2 provides a consensus-validated holistic assessment with dual matching strategies, together allowing for comprehensive evaluation of classifier performance.

Appendix C. CIR3: algorithm implementation details

C.1. Module input/output specifications

- **classify_subtopics**(c : str, M : int) $\rightarrow$ P : List[str]
  Identifies subtopics within context.
- **generate_QAs**(c : str, M_{prev} : InnerMemory, H_{prev} : OuterMemory) $\rightarrow$ $\mathcal{G}^+$: List[QAPair]
  Generates QA pairs from context and memory.
- **refine_QAs**($\mathcal{G}^+$: List[QAPair], $\mathcal{M}$: InnerMemory) $\rightarrow$ $\mathcal{F}$: Feedback $\cup$ $\{\emptyset\}$
  Produces refinement feedback or $\emptyset$ for consensus.
- **curmudgeon_QAs**($\mathcal{G}^-$: List[QAPair], $\mathcal{H}$: OuterMemory) $\rightarrow$ $\mathcal{F}'$: Feedback $\cup$ $\{\emptyset\}$
  External evaluation feedback or $\emptyset$ for acceptance.

C.2. Error handling

- **Subtopic identification failure**: Fallback to domain-specific default writer W_{p_0}.
- **Agent timeout**: 30-second timeout with retry mechanism (max 3 attempts).
- **Memory overflow**: Truncate oldest entries when memory exceeds max input tokens.

[29] https://pypi.org/project/simpledorff

- **API failures**: Exponential back-off with graceful degradation to available models.
  (a) Retry with exponential back-off (to handle transient failures).
  (b) If still failing after a threshold (e.g. 5 attempts), gracefully degrade by using an alternative model / service.

Appendix D. Human evaluation guidelines

This section represents a partial excerpt of the evaluation guidelines.

Box 2. **Partial Excerpt of the Human Evaluation Guidelines.**

This document describes the instructions of human evaluation for AI-based question-answer generation task. The results will be made publicly available within 12 months of study completion.

By proceeding with this task, you agree that any resulting work product may be shared publicly.

You will be provided with two datasets, each containing a collection of text documents and corresponding question-answer pairs. Your task is to evaluate the question-answer pairs in terms of their accuracy and completeness in relation to the information presented in the associated text documents.

Instructions:

Familiarise yourself with the context: Carefully read the provided context to understand the topic and key information it contains.

Review the generated question-answer pairs: Examine each question-answer pair.

Evaluate each criterion: For each of the following criteria, rate the question-answer pair on a scale of **1** to **5** (5 being the highest), and provide a brief explanation for your rating:

- **Coverage**: Does the question-answer pair address the main ideas and important details in the context?
- **Depth:** Does the question prompt deeper understanding of the context, or is it superficial? Does the answer provide sufficient detail and explanation?
- **Accuracy:** Is the answer factually correct and complete based on the information in the context?
- **Coherence:** Does the question-answer pair flow logically? Does the question naturally lead to the answer, and do they together contribute to a better understanding of the topic?
- **Representation:** Does the question-answer pair distort or present misleadingly any facts in the context?
- **Diversification:** Does the collection of question-answer pairs provide diverse and unique insights, or is there significant overlap in the knowledge they convey?

Appendix E. Evaluation prompts

Following figures show the illustrative prompts used in our evaluation.

We also release CIR3's source code on GitHub[30].

[30] https://github.com/anonym-nlp-ai/cirr

```python
"""
Create an Enhanced G-EVAL metric for Comprehensiveness
    evaluation based on Coverage, Depth, Accuracy, and
    Coherence aspects.
"""
EnhancedGEval(
    name="Comprehensiveness",
    evaluation_aspects=["Coverage", "Depth", "Accuracy",
       "Coherence"],
    evaluation_steps=[
        # Coverage
        "Examine the source document to identify all key
           topics, concepts, and important information
           covered",
        "Review the set of question-answer pairs to
           determine what aspects of the document they
           address",
        "Check if the questions cover all major themes
           and subtopics from the document",

        # Depth
        "Consider depth of coverage - are complex topics
           explored adequately or only superficially?",
        "Assess whether important details,
           relationships, and nuances are captured in
           the QA pairs",

        # Accuracy
        "Verify that the questions accurately reflect
           the document's content and answers are
           factually correct",

        # Coherence
        "Evaluate the logical flow and connection
           between questions and their relationship to
           document structure",

        # Final Assessment
        "Score HIGH if the QA set demonstrates
           comprehensive coverage, adequate depth,
           accuracy, and coherence",
        "Score LOW if major topics are missing, coverage
           is superficial, inaccurate, or lacks
           coherence",
    ],
    evaluation_params=[LLMTestCaseParams.INPUT,
       LLMTestCaseParams.ACTUAL_OUTPUT],
    threshold=0.6,
    model="gpt-4o",
    top_logprobs=20,
    async_mode=True,
    verbose_mode=False,
    _include_g_eval_suffix=True
)
```

Listing 1: Comprehensiveness metric with four evaluation aspects: Coverage, Depth, Accuracy, and Coherence.

```python
"""
Create an Enhanced G-EVAL metric for Faithfulness evaluation based on Accuracy, Exaggeration, Consistency, Justification, Plausibility, and Misrepresentation aspects.
"""
EnhancedGEval(
    name="Faithfulness",
    evaluation_aspects=[
        "Accuracy", "Exaggeration", "Consistency",
        "Justification", "Plausibility", "Misrepresentation"
    ],
    evaluation_steps=[
        # Accuracy
        "Carefully read the source document to understand the factual information presented",
        "Examine each answer to verify factual accuracy against the source document",

        # Exaggeration
        "Check for any statements that overstate or embellish information from the document",

        # Consistency
        "Look for contradictions or deviations from facts presented in the source material",

        # Justification
        "Verify that all claims in answers are well-supported by evidence from the document",

        # Plausibility
        "Assess whether answers represent reasonable inferences based on the document content",

        # Misrepresentation
        "Check for any distortion or misleading presentation of facts from the source",

        # Final Assessment
        "Score HIGH if answers demonstrate accuracy, avoid exaggeration, maintain consistency, are well-justified, plausible, and avoid misrepresentation",
        "Score LOW if answers contain inaccuracies, exaggerations, inconsistencies, poor justification, implausible claims, or misrepresentations",
    ],
    evaluation_params=[LLMTestCaseParams.INPUT, LLMTestCaseParams.ACTUAL_OUTPUT],
    threshold=0.6,
    model="gpt-4o",
    top_logprobs=20,
    async_mode=True,
    verbose_mode=False,
    _include_g_eval_suffix=True
)
```

Listing 2: Faithfulness metric with six evaluation aspects: Accuracy, Exaggeration, Consistency, Justification, Plausibility, and Misrepresentation.

Appendix F. Case study: QA evolution trajectory analysis

To demonstrate how iterative feedback tangibly improves QA diversity and alignment, we present a detailed trajectory analysis of a representative financial document undergoing CIR3's refinement process.

F.1. Qualitative curmudgeon feedback analysis

The curmudgeon's iterative feedback (Listing: Multi-Iteration Refinement Example) reveals concrete improvement mechanisms:

- **Iteration 1**: Identifies specific coverage gaps (*"misses the role of life insurance"*) and structural limitations (*"question set lacks diversity, primarily focusing on factual recall"*). The feedback provides actionable guidance: *"add more analytical and comparative questions"* and *"varying the linguistic patterns and cognitive demands."*.

Multi-Iteration Refinement Example

Context: "A strong financial plan balances growth, stability, and protection by strategically combining diverse assets. Bonds provide steady income and preserve capital, though they are subject to inflation and interest rate risks. For growth, index ETFs offer low-cost, diversified exposure to the stock market, accepting the associated market volatility. A structured retirement is built on a foundation of pensions, while annuities provide guaranteed lifetime income, guarding against the possibility of outliving one's savings. Finally, life insurance protects dependents and can build cash value for future needs. By aligning these tools with one's personal risk tolerance, time horizon, and specific goals, a resilient financial strategy can be built to adapt to economic changes and secure both a current lifestyle and future legacy."

Identified Subtopics:

- financial planning
- retirement
- investment
- risk management

Writer Agent Assignment:

- $W_{\text{financial planning}}$: General financial strategy perspective;
- $W_{\text{retirement}}$: Retirement-focused analysis;
- $W_{\text{investment}}$: Investment and growth perspective;
- $W_{\text{risk management}}$: Risk assessment and mitigation;

// ...

Outer Iteration k=1
Inner Iteration l=4 - Writers Collaboration: QA generation:

- Q: What is the main objective of a strong financial plan?
  A: To balance growth, stability, and protection by strategically combining diverse assets.
- Q: Which investment provides steady income and preserves capital?
  A: Bonds provide steady income and preserve capital.
- Q: What is the main role of index ETFs in a financial plan?
  A: They offer low-cost, diversified exposure to the stock market.

- Q: What forms the foundation of a structured retirement?
  A: Pensions form the foundation of a structured retirement.
- Q: What type of investment offers guaranteed lifetime income?
  A: Annuities offer guaranteed lifetime income.

Vendi Diversity Scores

- "balanced_g_score": 0.6598721146583557,
- "score_a": 1.3656857,
- "score_ca": 1.0813369,
- "score_q": 1.4364766

Curmudgeon reflection: The current QA set provides a good foundation, covering key topics such as the main objective of a financial plan, the roles of bonds, index ETFs, pensions, and annuities. However, it falls short in exploring the nuances of aligning financial strategies with personal circumstances and misses the role of life insurance. The answers accurately reflect the information in the document, but the question set lacks diversity, primarily focusing on factual recall with similar structures and complexity. To improve, it's essential to add more analytical and comparative questions that require deeper analysis and synthesis, as well as varying the linguistic patterns and cognitive demands.
Curmudgeon status: "refine"
// ...

Outer Iteration k=2
// ...

Inner Refinement Feedback (l=2): $W_{\texttt{retirement}}$: "Missing coverage of annuities and longevity risk protection." $W_{\texttt{risk management}}$: "Need specific risk details: inflation, interest rate, market volatility impacts." $W_{\texttt{financial planning}}$: "Lacks personalization factors: risk tolerance, time horizon alignment." $W_{\texttt{investment}}$: "Should clarify ETF advantages beyond basic market exposure."

Inner Iteration l=2 - Writers Collaboration: QA generation:

- Q: How does a strong financial plan integrate different asset types?
  A: It combines growth assets like index ETFs, stability assets like bonds, and protection assets like pensions, annuities, and life insurance.
- Q: What risks can affect the income from bonds?
  A: Bonds are subject to inflation and interest rate risks.

- **Iteration 2**: Demonstrates error detection capabilities by identifying a critical factual inaccuracy in Q7 (*"the document does not state that bonds guarantee higher returns over time"*). This shows the curmudgeon's ability to catch hallucination errors that could compromise faithfulness.
- **Iteration 3**: Confirms convergence with comprehensive validation: "*covers all major themes and subtopics*", *"answers align perfectly with the source document"* and *"demonstrates strong variety in types, including factual, analytical, and comparative questions."*

F.1.1. Common error mitigation evidence

The trajectory demonstrates CIR3's effectiveness against the four primary error categories:

- **Duplication**: The increasing Vendi Score for questions (1.436 → 1.590) indicates successful reduction of redundant or overly similar questions. The curmudgeon's observation about *"similar structures and complexity"* in Iteration 1 implicitly addresses this concern.
- **Hallucination**: Explicitly detected by the curmudgeon in Iteration 2, identifying Q7's factual inaccuracy (*"the document does not state that bonds guarantee higher returns over time"*) as hallucinated information not present in the source document.
- **Irrelevance**: The curmudgeon's focus on coverage gaps (*"misses the role of life insurance"*) and requirement for document alignment (*"answers align perfectly with the source document"*) addresses relevance, even if this error type is not explicitly flagged in this particular trajectory.
- **Specificity**: Addressed through the curmudgeon's feedback requesting *"varying the linguistic patterns and cognitive demands"* and *"more analytical and comparative questions"* which counters template driven or overly narrow question generation.

- Q: What trade-off comes with investing in index ETFs?
  A: They provide market growth potential at low cost but come with market volatility.
- Q: How do annuities support retirement security?
  A: They provide guaranteed lifetime income, guarding against the possibility of outliving one's savings.
- Q: What is the role of life insurance in financial protection?
  A: It protects dependents and can build cash value for future needs.
- Q: How can financial tools be matched to personal needs?
  A: By aligning them with personal risk tolerance, time horizon, and specific goals.
- Q: What is a key advantage of bonds compared to index ETFs?
  A: Bonds guarantee higher returns over time with minimum risk of loss, making them a safer alternative for long-term investors seeking consistent growth.

Vendi Diversity Scores

- "balanced_g_score": 0.7028632760047913,
- "score_a": 1.6247656,
- "score_ca": 1.1747185,
- "score_q": 1.5361243

Curmudgeon reflection: The QA set are mostly aligned with the document and covers most major themes from the document, including the integration of different asset types, risks associated with bonds, trade-offs of index ETFs, and roles of annuities and life insurance. It also addresses aligning financial tools with personal needs. However, Q7 contains a factual inaccuracy, as the document does not state that bonds guarantee higher returns over time or that they are a safer alternative for consistent growth. This misrepresentation reduces the depth of coverage and overall alignment with the source material. To further refine, it's crucial to ensure the accuracy of all questions and maintain a diverse range of question types and complexities.
Curmudgeon status: "refine"

Inner Iteration l=2 - Writers Collaboration: QA generation:

- Q: How does strategic asset allocation create a resilient financial plan?
  A: By balancing growth through index ETFs, stability through

Table F.12
Vendi Score diversity metrics for questions, answers, context-answers, and balanced G score across iterations.

		Vendi Scores			
Iteration	QA Count	Questions	Answers	Context-Answers	Balanced G
#1	5	1.436	1.366	1.081	0.660
#2	7	1.536	1.625	1.175	0.703
#3	8	1.590	1.500	1.062	0.742

F.2. Quantitative evolution patterns

Tables F.12 and F.13 reveal a systematic improvement in key metrics with notable intermediate dynamics:

- **Comprehensiveness** shows substantial growth from 0.734 (Iteration 1) to 0.953 (Iteration 3), representing a 29.85 % improvement. QA Count expands strategically from 5 to 8 questions, indicating a controlled expansion of coverage rather than arbitrary multiplication.

bonds, and protection through pensions, annuities, and life insurance.

- Q: In what way do bonds contribute to stability but face risks?
  A: They provide steady income and preserve capital but are affected by inflation and interest rate changes.
- Q: How do index ETFs contribute to long-term growth?
  A: They provide diversified stock market exposure at low cost while accepting market volatility.
- Q: Why are pensions important in retirement planning?
  A: They provide a stable foundation for a structured retirement.
- Q: What dual benefit do annuities offer in retirement?
  A: They ensure guaranteed lifetime income and protect against outliving one's savings.
- Q: How does life insurance support both protection and future planning?
  A: It safeguards dependents and can accumulate cash value for future needs.
- Q: What personal factors shape an adaptive financial strategy?
  A: Risk tolerance, time horizon, and specific goals.
- Q: How can a financial plan remain effective during economic changes?
  A: By aligning growth, stability, and protection tools with personal needs.

Vendi Diversity Scores

- "balanced_g_score": 0.741766095161438
- "score_a": 1.5003632
- "score_ca": 1.0616497
- "score_q": 1.5900005

Curmudgeon reflection: ""
Curmudgeon: "agreement"
Curmudgeon agreement feedback: The QA set has achieved a high level of quality, comprehensively covering all major themes and subtopics from the document, including strategic asset allocation, personal factors, and the roles of various financial instruments. The answers align perfectly with the source document, maintaining the tone and perspective without contradictions or inaccuracies. The question set demonstrates a strong variety in types, including factual, analytical, and comparative questions, and explores diverse aspects of financial planning from different angles. The set effectively covers different levels of detail and maintains linguistic diversity, making it ready for agreement without further refinement.

Table F.13
Comprehensiveness and Faithfulness scores across iterations for QA trajectory.

Iteration	QA Count	Comprehensiveness	Faithfulness
#1	5	0.734	1.000
#2	7	0.803	0.767
#3	8	0.953	0.999

- **Faithfulness** demonstrates an interesting trajectory: starting at perfect alignment (1.00), it drops to 0.767 in Iteration 2 as new content is added, then recovers to near-perfect alignment (0.999) in Iteration 3. Similarly, Vendi Score A (answer diversity) shows a comparable pattern, peaking at 1.625 in Iteration 2 before stabilizing at 1.5 in Iteration 3. These temporary fluctuations reflect the natural trade-off between coverage expansion and maintaining document fidelity, a pattern that validates the curmudgeon's role in balancing comprehensiveness, diversity, and faithfulness through iterative refinement.
- **Vendi Score** patterns demonstrate the curmudgeon's dual optimization: Question diversity increases consistently from 1.436 to 1.59, while the balanced generation score (G) improves from 0.66 to 0.742, showing enhanced overall system performance despite the intermediate adjustments in answer diversity.

Appendix G. Computational cost analysis

CIR3's multi-agent architecture with iterative refinement requires careful consideration of computational resources for practical deployment. We provide detailed cost breakdowns across different infrastructure configurations to enable informed feasibility assessments.

Table G.14
Average CIR3 computational resource utilization for 1000 documents, comparing Groq cloud API deployment and self-hosted p5.48xlarge instance with vLLM, showing token consumption and round-trip inference times per component.

Component	Tokens	Round trip inference time (s)	
		Groq	AWS p5.48xlarge + vLLM
Classifier Agent	25	0.2	0.26
Moderator Agent	2007	1.7	2.17
Writer Agents	17,287	13.71	17.54
Curmudgeon Agent	617	11.28	14.43
Diversity (Encoder)	-	6.55	6.55
Total	**19937**	**38**	**48.51**

Table G.14 presents comprehensive resource utilization metrics for CIR3 across two deployment scenarios: cloud-based API services (Groq[31]) and self-hosted infrastructure (p5.48xlarge[32] with vLLM). The total token consumption per document averages 19,937 tokens across all agents, with writer agents consuming the majority (17287 tokens, 86.7 %) due to their iterative QA generation and refinement processes.

G.1. Component-level analysis

The computational distribution reveals strategic resource allocation: the classifier agent requires minimal resources (25 tokens, 0.2-0.26s), enabling efficient subtopic identification. The moderator agent consumes 2007 tokens (1.7-2.17s) for coordination tasks, while the diversity encoder requires 6.5 seconds for Vendi Score computation via a self-hosted embeddings API[33]. The curmudgeon agent runs in 11.28–14.43 seconds across configurations, reflecting its evaluation and feedback generation steps.

G.2. Infrastructure performance analysis

To provide comprehensive feasibility assessment, we evaluated CIR3 across two distinct deployment architectures with detailed performance characteristics.

[31] https://groq.com/pricing
[32] https://instances.vantage.sh/aws/ec2/p5.48xlarge?currency=USD
[33] https://github.com/huggingface/text-embeddings-inference

Table G.15
Comparative performance analysis of CIR3 deployment configurations showing throughput, latency characteristics, and operational trade-offs.

Metric	Groq (LPU)	AWS p5.48xlarge + vLLM
Throughput (tokens/sec)	525	411
Total Processing Time (s)	38.0	48.5
Time to First Token (s)	0.22-0.3	0.2-0.4
Latency Consistency	Deterministic	Variable
Operational Complexity	Minimal	High
Deployment Flexibility	Limited	Full Control

Table G.16
Cost analysis for processing 1000 documents across deployment configurations. **Note:** Cost and time estimates are based on sequential execution and do not account for batch processing optimizations, which would reduce these numbers considerably.

Configuration	Total Cost	Cost per Document	Processing Time
Groq Cloud (Llama 3 70B)	$16.90	$0.0169	10.6 hours
AWS p5.48xlarge (8xH100, Spot)	$253.07	$0.25	13.5 hours
AWS p4de.24xlarge (8xA100, Spot)	$158.65	$0.16	13.5 hours*

*Estimated processing time; actual performance may vary.

Groq Cloud Deployment: Groq's Language Processing Unit (LPU) architecture delivered consistent performance, processing 19,937 tokens in 38.0 seconds with sustained throughput of approximately 525 tokens per second (Table G.15). The deterministic latency characteristics (time-to-first-token: 0.22-0.3s) prove particularly valuable for multi-agent systems where round-trip delays compound across iterative cycles.

Self-Hosted AWS Configuration: Our optimized AWS p5.48xlarge deployment (8xH100 GPUs with vLLM) achieved comparable performance: 19937 tokens in 48.5 seconds, sustaining approximately 411 tokens per second. Key optimizations included continuous batching for pipeline utilization, KV cache management to prevent memory thrashing, CUDA graphs with FlashAttention for latency reduction, and tuned EFA/NCCL communication across GPUs[34].

G.3. Practical implications

The performance differential between configurations is modest (22 % throughput difference), with Groq achieving superior consistency while our AWS setup provides greater customization capabilities. Both configurations demonstrate CIR3's practical feasibility for production deployment. The throughput rates (411–525 tokens/second) support real-time document processing applications, while the total processing time (38–49 seconds per document) remains reasonable for comprehensive QA generation tasks. Despite the multi-agent complexity, CIR3 demonstrates favorable cost-benefit ratios when considering the substantial quality improvements. The modular architecture enables selective deployment optimization, such as caching classifier results or parallelizing writer agent operations, making the framework adaptable to various computational budget constraints.

G.3.1. Cost analysis for knowledge base processing

To evaluate CIR3's economic feasibility for large-scale deployment, we analyze the costs of processing a knowledge base of 1000 documents across multiple infrastructure configurations.

Groq Cloud Pricing: Based on Groq's current pricing structure [77], using Llama 3 70B (our primary model) costs $0.59 per million input

[34] Elastic Fabric Adapter (EFA) enables low-latency interconnect on AWS; NCCL provides high-performance multi-GPU/multi-node collectives: https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/efa.html, https://developer.nvidia.com/nccl

tokens and \$0.79 per million output tokens. For CIR3's token distribution (19,937 tokens per document with approximately 70 % input and 30 % output), the cost per document is approximately \$0.0169. Processing 1000 documents would cost approximately **\$16.90** in token fees, with processing completing in approximately 10.6 hours (Table G.16).

AWS Self-Hosted Configurations: We evaluated two AWS configurations for self-hosted deployment:

- **p5.48xlarge (8xH100):** At \$18.746 per hour spot pricing[35], processing 1000 documents (13.5 hours) costs approximately \$253.07
- **p4de.24xlarge (8xA100):** At \$11.752 per hour spot pricing, the same workload costs approximately \$158.65

It is important to note that in practical deployments, self-hosted infrastructure typically serves multiple applications beyond CIR3, effectively amortizing costs across various workloads. This shared utilization significantly improves the cost-effectiveness of dedicated GPU clusters for organizations running multiple AI applications.

G.3.2. Cost-benefit analysis

The substantial quality improvements achieved by CIR3 (+23 % comprehensiveness, +17 % faithfulness compared to baselines) justify the computational overhead for applications requiring high-quality QA generation. CIR3 is particularly well-suited for batch processing scenarios such as:

- Knowledge base creation and augmentation for enterprise documentation.
- Dataset indexing and enrichment for information retrieval systems.
- Educational content processing for automated quiz and assessment generation.
- Research literature analysis for systematic review and knowledge extraction.
- Financial document processing for banking, investment analysis, and regulatory compliance.
- Legal and compliance document processing for searchable QA databases.
- Scheduled document processing pipelines for content management systems.

For batch processing workflows, AWS spot pricing offers significant cost advantages, with the p4de.24xlarge configuration providing a favorable balance between performance and cost. When infrastructure is shared across multiple applications, the effective cost per CIR3 document decreases substantially, making large-scale deployment highly economical.

[35] AWS pricing may vary across regions: https://aws.amazon.com/ec2/spot/pricing

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, 2024. arXiv:2308.07107.

[2] L. Silva, L. Barbosa, Improving dense retrieval models with LLM augmented data for dataset search, Knowl. Base. Syst. 294 (2024) 111740. https://doi.org/10.1016/j.knosys.2024.111740

[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, Knowl. Base. Syst. 309 (2025) 112789. https://doi.org/10.1016/j.knosys.2024.112789

[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. 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, Knowl. Base. Syst. 309 (2025) 112837. https://doi.org/10.1016/j.knosys.2024.112837

[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. 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. 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. 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. Krisnadhi (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. 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 Trans. Inf. Syst. 40 (1) (2022) 1–43. 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. 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, 2020. 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. https://doi.org/10.1145/3539618.3591960

[15] X. Yuan, T. Wang, Y.-H. Wang, E. Fine, R. Abdelghani, H. Sauzéon, 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. https://doi.org/10.18653/v1/2023.findings-acl.820

[16] Z. Zhao, W. Fan, J. Li, Y. Liu, X. Mei, Y. Wang, Z. Wen, F. Wang, X. Zhao, J. Tang, Q. Li, Recommender systems in the era of large language models (LLMs), IEEE Trans. on Knowl. and Data Eng. 36 (11) (2024) 6889–6907. https://doi.org/10.1109/TKDE.2024.3392335

[17] 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, 2024. https://doi.org/10.48550/arXiv.2401.03428

[18] 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.

[19] Y. Du, S. Li, A. Torralba, J.B. Tenenbaum, I. Mordatch, Improving Factuality and Reasoning in Language Models through Multiagent Debate, 2023. arXiv:2305.14325.

[20] 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, 2023. arXiv:2305.19118.

[21] 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, 2024. arXiv:2402.01680.

[22] 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. https://doi.org/10.1016/j.jmathb.2014.02.001

[23] 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. https://api.semanticscholar.org/CorpusID:261334778.

[24] E. Kamalloo, 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. https://doi.org/10.18653/v1/2023.acl-long.307

[25] X. Zeng, A. Zubiaga, Aggregating pairwise semantic differences for few-shot claim verification, PeerJ Comput. Sci. 8 (2022) e1137. https://doi.org/10.7717/peerj-cs.1137

[26] A. Amiri-Margavi, I. Jebellat, E. Jebellat, S.P.M. Davoudi, Enhancing Answer Reliability Through Inter-Model Consensus of Large Language Models, 2024. https://doi.org/10.48550/arXiv.2411.16797
[27] C.M. Giannantonio, Book review: krippendorff, k. (2004). content analysis: an introduction to its methodology (2nd ed.). thousand oaks, CA: sage, Organ. Res. Method. 13 (2) (2010) 392–394. Publisher: SAGE Publications Inc. https://doi.org/10.1177/1094428108324513
[28] R. Artstein, M. Poesio, Survey article: inter-coder agreement for computational linguistics, Comput. Linguist. 34 (4) (2008) 555–596. https://doi.org/10.1162/coli.07-034-R2
[29] 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 Comput. Linguist., Association for Computational Linguistics, Los Angeles, California, 2010, pp. 609–617.
[30] J. Mostow, W. Chen, Generating instruction automatically for the reading strategy of self-Questioning, International Conference on Artificial Intelligence in Education - (-) (2009).
[31] Y. Huang, L. He, Automatic generation of short answer questions for reading comprehension assessment, Nat. Lang. Eng. 22 (3) (2016) 457–489. https://doi.org/10.1017/S1351324915000455
[32] 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 Comput. Linguist. (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 1342–1352. https://doi.org/10.18653/v1/P17-1123
[33] K.D. Dhole, S. Bajaj, R. Chandradevan, E. Agichtein, QueryExplorer: An Interactive Query Generation Assistant for Search and Exploration, 2024. arXiv:2403.15667.
[34] 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. https://doi.org/10.18653/v1/D16-1264
[35] 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. https://doi.org/10.18653/v1/K17-1028
[36] 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 Comput. Linguist. (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 1725–1735. https://doi.org/10.18653/v1/P18-1160
[37] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, et al., 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 Comput. Linguist.: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423
[38] 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 Comput. Linguist., Association for Computational Linguistics, Florence, Italy, 2019, pp. 6168–6173. https://doi.org/10.18653/v1/P19-1620
[39] 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. https://doi.org/10.18653/v1/2020.emnlp-main.468
[40] 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. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Comput. Linguist., Association for Computational Linguistics, Online, 2020, pp. 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703
[41] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, et al., RoBERTa: A Robustly Optimized BERT Pretraining Approach, 2019. https://doi.org/10.48550/arXiv.1907.11692
[42] L. Murakhovs'ka, C.-S. Wu, P. Laban, T. Niu, W. Liu, C. Xiong, MixQG: Neural Question Generation with Mixed Answer Types, 2022. arXiv:2110.08175.
[43] 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. https://doi.org/10.18653/v1/2021.emnlp-main.696
[44] 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 Comput. Linguist.: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 14262–14272. https://doi.org/10.18653/v1/2023.findings-acl.899
[45] W. Zhang, W. Hua, K. Stratos, EntQA: Entity Linking as Question Answering, 2022. https://doi.org/10.48550/arXiv.2110.02369
[46] 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 Comput. Linguist.: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 5449–5463. https://doi.org/10.18653/v1/2023.findings-emnlp.362
[47] 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, 2023. https://doi.org/10.48550/arXiv.2308.00352
[48] G. Li, H.A. A.K. Hammoud, H. Itani, D. Khizbullin, B. Ghanem, CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society, 2023. arXiv:2303.17760.
[49] 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, 2023. arXiv:2307.07924.
[50] 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, 2023. arXiv:2308.07201.
[51] 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, 2024. https://doi.org/10.48550/arXiv.2402.14207
[52] 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 Comput. Linguist.: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 7572–7590. https://doi.org/10.18653/v1/2023.findings-emnlp.508
[53] H. Chen, W. Ji, L. Xu, S. Zhao, Multi-Agent Consensus Seeking via Large Language Models, 2023. arXiv:2310.20151.
[54] P. Li, V. Menon, B. Gudiguntla, D. Ting, L. Zhou, Challenges Faced by Large Language Models in Solving Multi-Agent Flocking, 2024. arXiv:2404.04752.
[55] D.J. Hall, R.A. Davis, Engaging multiple perspectives: a value-based decision-making model, Decis. Support Syst. 43 (4) (2007) 1588–1604. https://doi.org/10.1016/j.dss.2006.03.004
[56] M. Ihori, 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 Comput. Linguist., International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 6128–6138.
[57] I. Momennejad, Collective minds: social network topology shapes collective cognition, Philosoph. Transact. Roy. Soc. B: Biolog. Sci. 377 (1843) (2021) 20200315. https://doi.org/10.1098/rstb.2020.0315
[58] 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. ,. https://api.semanticscholar.org/CorpusID:17063782.
[59] 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. https://doi.org/10.1007/978-3-642-01991-3_4
[60] D. Friedman, A.B. Dieng, The Vendi Score: A Diversity Evaluation Metric for Machine Learning, 2023. https://doi.org/10.48550/arXiv.2210.02410
[61] 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, 2018. https://doi.org/10.48550/arXiv.1611.09268
[62] 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, Transact. Assoc. Comput. Linguist. 7 (2019) 452–466. https://doi.org/10.1162/tacl_a_00276
[63] 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. https://doi.org/10.1145/3184558.3192301
[64] 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. https://doi.org/10.1109/ASRU.2015.7404872
[65] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, P. Szolovits, What disease does this patient have? A large-scale open domain question answering dataset from medical exams, Appl. Sci. 11 (14) (2021) 6421.
[66] A. Pal, L.K. Umapathi, M. Sankarasubbu, MedMCQA: a large-scale multi-subject multi-choice dataset for medical domain question answering, in: Proceedings of the Conference on Health, Inference, and Learning, PMLR, 2022, pp. 248–260. ISSN: 2640–3498, https://proceedings.mlr.press/v174/pal22a.html.
[67] 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. ,. https://api.semanticscholar.org/CorpusID:233296016.
[68] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C.H. Yu, J. Gonzalez, H. Zhang, I. Stoica, Efficient memory management for large language model serving with PagedAttention, in: Proceedings of the 29th Symposium on Operating Systems Principles, ACM, Koblenz Germany, 2023, pp. 611–626. https://doi.org/10.1145/3600006.3613165
[69] 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.

[70] 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.
[71] L.d.F. Costa, Further generalizations of the jaccard index, ArXiv abs/2110.09619 (2021). https://api.semanticscholar.org/CorpusID:239024336.
[72] T. Zhang, V. Kishore, F. Wu, K.Q. Weinberger, Y. Artzi, BERTScore: Evaluating Text Generation with BERT, 2020. https://doi.org/10.48550/arXiv.1904.09675
[73] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, C. Zhu, G-Eval: NLG Evaluation Using GPT-4 with Better Human Alignment, 2023. https://doi.org/10.48550/arXiv.2303.16634
[74] 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, 2023. https://doi.org/10.48550/arXiv.2201.11903
[75] 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, 2023. arXiv:1910.10683.
[76] Y. Du, S. Li, A. Torralba, J.B. Tenenbaum, I. Mordatch, Improving factuality and reasoning in language models through multiagent debate, in: Proceedings of the 41st International Conference on Machine Learning, 235 of *ICML'24*, JMLR.org, Vienna, Austria, 2024, pp. 11733–11763.
[77] Groq, On-demand Pricing for Tokens-as-a-Service, 2025. Accessed January 2025, https://groq.com/pricing.