Optimizing Prompt Refinement: Algorithmic Strategies for Large Language Model-based Text Classification

Ziqiao Ao
Applied Scientist at Microsoft
Microsoft Corporate
One Microsoft Way, Redmond, WA

Juhi Singh
Principal Applied Scientist at Microsoft
Microsoft Corporate
One Microsoft Way, Redmond, WA
*Contributed equally as first author

Sebastian Antinome
Director of Business Insights at Microsoft
Microsoft Corporate
One Microsoft Way, Redmond, WA

ABSTRACT

The performance of Large Language Models (LLMs) for text classification depends on how well prompts are designed and refined. This paper presents a structured framework for improving prompt refinement strategies for LLM-based classification, with a focus on question-type classification for Microsoft technical certification exams. Several prompt optimization techniques were evaluated, including Chain of Thought (CoT), Self-Consistency, Tree of Thought (ToT), and different configurations of Retrieval-Augmented Generation (RAG) were evaluated. A modular prompt structure was also developed to support category-specific evaluation and improve decision consistency. Experiments were conducted in three stages: (1) tuning and comparing prompt refinement techniques, (2) optimizing RAG retrieval parameters, and (3) applying a modular rule-based approach to enhance classification reliability. Experimental results indicate that the proposed framework enhances classification performance, achieving an absolute improvement of approximately 40 percentage points in F1 score compared to baseline prompting methods. The methodology can be adapted to educational assessment, automated content analysis, and other text classification applications.

General Terms

Artificial Intelligence, Natural Language Processing, Machine Learning

Keywords

Prompt Engineering, Retrieval-Augmented Generation, Large Language Models, Text Classification, Modularized Prompting

1. INTRODUCTION

1.1 Research Background

Large language models have significantly advanced natural language processing by demonstrating strong performance across a wide range of text based tasks, including classification, sentiment analysis, and question answering. Their adaptability and reasoning capabilities make them particularly suitable for complex language understanding tasks. However, achieving reliable performance in text classification remains challenging due to factors such as ambiguous prompt design, limited contextual grounding, and insufficient domain specific optimization.

Prompt engineering can be an effective method for improving model behavior, by guiding large language models through iterative prompt refinement. Well designed prompts can improve accuracy, consistency, and interpretability of model outputs. This study focuses on prompt engineering for quality based exam question classification in technical certification assessments, where the objective is to automatically identify and filter low quality questions while maintaining alignment with domain specific evaluation criteria.

1.2 Problem Statement

Crafting assessment items for Microsoft credentials and certifications is an intricate, expensive, and time-consuming endeavor involving many stakeholders, ranging from internal content developers to external specialized vendors. With nearly 90 certifications, each requiring 100–300 questions for practice assessments, exams, and certification renewals that demand regular refresh, the demand for efficient, cost-effective, and high-quality assessment question creation is paramount.

In the context of Microsoft Learn, the challenge is amplified by the need to adhere to stringent rules and guidelines for exam standardization, as detailed in a comprehensive 45-page document. Generated questions must be accurate, unbiased, and strictly based on Microsoft's official documentation or training materials. The system must produce items for a range of low- to high-stakes assessments, including diverse question types beyond multiple-choice, reflecting increased complexity and adherence to higher-order cognitive skills as per Bloom's taxonomy. Additionally, it is essential to identify duplicates and enemy pairs within the item bank, where enemy pairs are questions that could hint at each other's answers.

Conventional assessment authoring workflows are labor intensive, inconsistent, and difficult to scale. While AI assisted authoring offers the potential to improve efficiency and reduce manual effort, its effectiveness depends on reliable classification and quality control mechanisms. Large language models combined with structured prompt engineering present a promising solution, but their performance varies based on prompt design and refinement. This motivates the need for a modular and systematic prompt engineering framework that improves classification accuracy and provides actionable feedback for item authors or as input for an automated item filtering system.

1.3 Research Objectives

This study introduces a comprehensive framework for optimizing LLM prompts tailored to exam question quality classification tasks. The research aims to:

(1) Develop and evaluate advanced prompt engineering strategies, including Chain of Thought (CoT), Self-Consistency, and Tree of Thought (ToT) techniques.
(2) Optimize Retrieval-Augmented Generation (RAG) to enhance retrieval and generation processes.
(3) Propose a modularized prompt structure to enable category-specific evaluations (e.g., relevancy, format, clarity, complexity, accuracy).
(4) Assess the effectiveness of the proposed methodologies through quantitative experiments measuring precision, recall, and F1 scores.

Through iterative experimentation and performance evaluation, a scalable methodology is established for achieving high accuracy and consistency in LLM driven classification tasks. This work provides a practical guide for prompt optimization across diverse and complex application scenarios.

2. RELATED WORK

2.1 Prompt Engineering for LLM Optimization

Prompt engineering is a key approach for improving the performance of large language models on classification tasks. Early studies focused on zero shot and few shot learning, where models were evaluated on unseen tasks with limited or no examples. More recent work has explored zero shot text classification by integrating prompt based keyword extraction and knowledge graph embeddings, demonstrating improved performance in low data settings [21]. Other studies have questioned whether in context learning alone is sufficient, suggesting that retrieval mechanisms or additional tuning may be required for reliable classification [4].

To improve interpretability and accuracy, Chain of Thought prompting introduced structured reasoning steps [13], while Self Consistency Chain of Thought aggregated multiple reasoning paths to enhance reliability [21]. Tree of Thought prompting further extended this approach by evaluating alternative reasoning paths to support decision making [22]. More recently, Rank Prompting has been proposed to improve decision consistency in classification tasks [29]. Additional research has examined lightweight large language models for classification, showing that prompt learning techniques can reduce computational cost while maintaining competitive accuracy [3].

Prompt engineering has also been applied to domain specific tasks such as legal text classification [6] and financial natural language processing, where combining prompt strategies with fine tuning has improved performance [15]. Prompt chaining approaches, in which classification precedes extraction, have further demonstrated improvements in structured reasoning and information retrieval [10]. Building on these studies, this work integrates multiple prompt engineering strategies within a unified classification framework. Through iterative experimentation, Chain of Thought, Tree of Thought, Role Prompting, and Rank Prompting were applied to the same task and their effectiveness was evaluated using precision, recall, and F1 score metrics.

2.2 LLM Evaluation

Evaluation plays a critical role in assessing the effectiveness of large language models for natural language processing tasks. Traditional evaluation methods rely on metrics such as BLEU, ROUGE, and METEOR, which primarily measure surface level similarity between generated text and reference outputs [17, 12]. While useful, these metrics often fail to capture semantic correctness, logical reasoning, and factual accuracy. More advanced evaluation methods have been introduced, including BERTScore, which leverages contextual embeddings to assess semantic similarity [26], MAUVE, which compares the distribution of generated text with human written text [18], and GPTScore, a model based metric shown to correlate with human judgment [5].

Evaluating large language models in retrieval augmented settings presents additional challenges, particularly for long context inputs. Prior work has proposed dual perspective retrieval augmented generation frameworks to improve retrieval efficiency and context utilization [27]. Other studies have examined how model performance degrades as input context length increases, resulting in reduced accuracy and consistency [24]. Evaluation efforts have also expanded to include truthfulness and reliability metrics. The TruthfulQA benchmark focuses on factual correctness and hallucination avoidance [12], while the Holistic Evaluation of Language Models framework incorporates dimensions such as bias, fairness, and calibration [11].

In contrast to prior approaches, this study evaluates prompt engineering strategies using expert labeled ground truth data within a structured classification task. Instead of relying solely on generic text similarity metrics, performance was assessed using precision, recall, and F1 score metrics to ensure practical relevance.

2.3 Quality based Exam Question Classification

Quality based exam question classification evaluates assessment items according to predefined criteria such as clarity, complexity, and relevance [13]. Recent research has applied large language models with adaptive techniques including prompt tuning and active learning to improve classification performance [7]. Zero shot classification methods have also been explored for domain specific tasks, demonstrating that prompt based approaches can generalize effectively with limited labeled data [21].

This study extends prior work by introducing a structured and data driven framework to identify effective prompt tuning strategies for exam question classification. Unlike general purpose classification models, the proposed approach explicitly incorporates domain specific evaluation criteria. Related research in legal text classification highlights the importance of tailoring prompt design to specialized domains [6], while recent findings suggest that in context learning alone may be insufficient without additional tuning or retrieval mechanisms [4]. Similar conclusions have been reported in financial text classification, where hybrid prompt based approaches improved domain specific performance [15].

To support systematic evaluation, this work uses expert labeled ground truth data to compare multiple prompt tuning strategies. In addition, a Rule Compliance Score was introduced to measure how closely model outputs align with predefined classification rules, ensuring both accurate and contextually meaningful predictions. Prior research on prompt chaining shows that classification before extraction improves performance in structured retrieval tasks [10]. The resulting modular evaluation framework supports scalability and adaptability, enabling application beyond educational assessments to broader text classification scenarios.

3. METHODOLOGY

The development of optimized prompts for exam question quality classification followed a structured roadmap that used classification metrics as evaluation criteria for comparative analysis. Iterative experiments were conducted to assess effectiveness and adaptability, with ground truth data used to validate improvements in classification performance. This roadmap enabled systematic integration and refinement of prompt engineering techniques through controlled experimentation.

The framework emphasized refined prompt patterns and advanced in context strategies, including hybrid combinations such as Few Shot Prompting with Chain of Thought and Zero Shot Prompting with Chain of Thought, as well as Role Prompting and Rank Prompting. Further optimization was achieved through Retrieval Augmented Generation, with experiments on chunk size selection, overlap configuration, and indexing strategies. Finally, a modularized approach was introduced to improve scalability and support structured evaluation across classification categories.

3.1 Prompt Engineering Techniques

3.1.1 Zero-Shot Prompting. In the zero-shot prompt technique, specific to the use case the model classified exam questions based on task instructions without any guiding examples [19, 16]. The prompt defined the role, rules for classification, and an output format in JSON. The technique established a baseline using pretrained knowledge but struggled with nuanced rule violations and logical connections between question stems. These limitations highlighted the need for contextual guidance, paving the way for improved techniques such as Few-Shot Prompting.

3.1.2 Few-Shot Prompting. Few-shot prompting significantly enhances LLM's ability to classify questions by embedding task-specific examples within the prompt [14]. This technique leverages a small, curated set of examples to establish context, guiding the model's responses toward alignment with the predefined classification rules [28]. For instance, in the classification task, sample questions and responses are included to illustrate compliance or violation of quality criteria.

3.1.3 Chain of Thought Prompting. Chain of Thought (CoT) prompting is a structured approach that guides an LLM through a sequential reasoning process to enhance its decision-making capabilities. Specific to the use case (Figure 1), CoT prompting involves breaking down the evaluation into discrete, logical steps corresponding to predefined rules.

Each step requires the model to assess specific rule categories and provide detailed reasoning for its classification decisions. CoT prompting combines systematic evaluation, granular reasoning, and enhanced consistency to significantly improve classification accuracy. By structuring the evaluation process through Sequential Analysis, CoT prompts guide the model to address each rule methodically, minimizing the risk of overlooking critical quality aspects. The structured breakdown of criteria further enables the identification of nuanced rule violations, enabling the model to tackle intricate tasks.

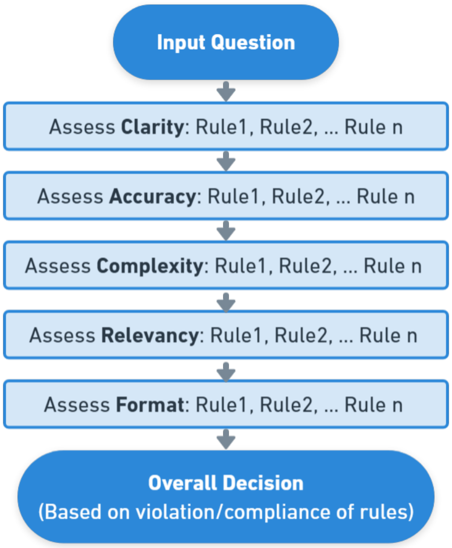
Input Question
Assess Clarity: Rule1, Rule2, ... Rule n
Assess Accuracy: Rule1, Rule2, ... Rule n
Assess Complexity: Rule1, Rule2, ... Rule n
Assess Relevancy: Rule1, Rule2, ... Rule n
Assess Format: Rule1, Rule2, ... Rule n
Overall Decision
(Based on violation/compliance of rules)

Fig. 1. Chain-of-Thought(CoT) Prompt Flow

Prompt Excerpt

Task:
For each exam question provided, please follow the steps to evaluate it sequentially against a set of defined rules.

Step 1: Assess Question Clarity and Subtlety
- Rules: <list of rules>
- Provide a decision indicating compliance or violation in **Rule 1 Decision**, with a brief explanation in **Rule 1 Explanation**.

Step 2: Evaluate Answer Quality and Challenge
- Rules: <list of rules>
- Provide a decision indicating compliance or violation in **Rule 2 Decision**, with a brief explanation in **Rule 2 Explanation**.
...

Overall Classification Decision Process:
- After assessing each step, compile the decisions. Conclude with an overall classification decision of the question as "Level1" or "Other" in "ItemLevel", based on the assessments made for each rule. Provide an "Overall Reasoning" for the final decision.
- If the question violates any rule, classify it as "Level1" and provide reasoning highlighting which specific rule(s) were violated. This indicates the question is unsuitable for use due to issues in clarity, challenge, depth, or logic.
- If the question adheres to all the **Rules**, classify it as "Other" and affirm the question's adherence to the quality standards.

3.1.4 Self-Consistency Chain of Thought Prompting. Self-Consistency Chain of Thought (SC-CoT) prompting leverages the principles of SC-CoT reasoning to systematically evaluate exam questions in this use case [20] (See Figure 2). The method emphasizes the rigorous assessment of question quality across rules within predefined categories. This technique ensures robust and consistent classification by generating multiple reasoning paths for the same question and consolidating these paths into a final decision based on modal consensus.

3.1.5 Tree of Thought Breadth-First Search Prompting. The Breadth-First Search (BFS) approach in Tree of Thought(ToT)

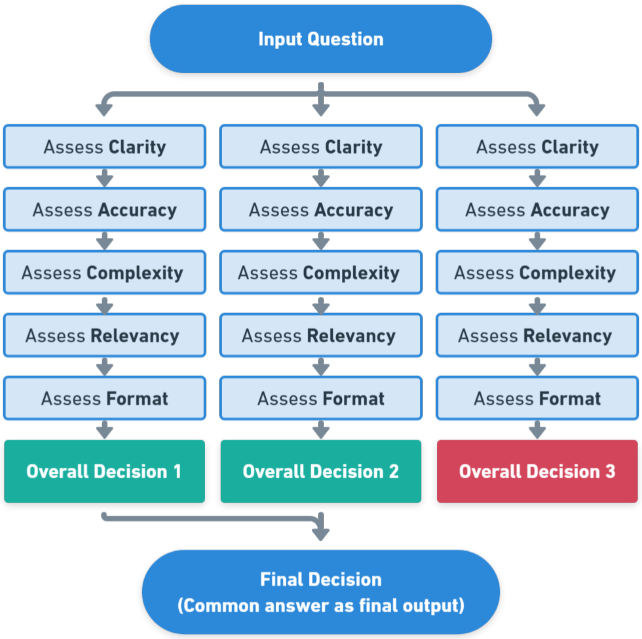

Fig. 2. Self-Consistency Chain-of-Thought (SC-CoT) Prompt Flow

prompt systematically evaluates questions against predefined rules [22] (See Figure 3).

Each category of rule was treated as a distinct branch. The BFS strategy ensured that all rules were evaluated concurrently, allowing for a comprehensive assessment of questions quality. Few-shot examples illustrated compliant and non-compliant scenarios, enabling the model to better understand the evaluation process.

Each question was simultaneously evaluated against all rules in parallel, leveraging BFS to prevent oversight in the decision-making process. For example, Rule 1 focused on whether the question stem was clear and free from overt clues, while Rule 2 checked the quality of distractors and alignment with technical requirements. Rule 3 assessed the question's complexity to ensure it extended beyond basic memorization, and Rule 4 confirmed the logical connection between the stem and answers. Aggregating the results from all branches ensured comprehensive evaluation and accurate classification. Questions failing any rule were classified as 'violation' while those complying with all were deemed 'compliant'. The BFS strategy improved transparency by requiring detailed explanations for each decision, facilitating validation against predefined criteria.

3.1.6 Tree of Thought Depth-First Search Prompting. In Depth-First Search (DFS) ToT prompting, rules are assessed sequentially (See Figure 4).

If a rule is violated, the classification halts, reducing unnecessary evaluations [22]. With DFS, the model evaluated Rule 1 first. If the question stem failed this rule, further evaluation ceased, classifying the questions immediately. This method optimized resource use by focusing on problematic rules. This DFS strategy proved effective in identifying low-quality questions early in the evaluation process, significantly reducing computational overhead. By progressing rule-by-rule, the model ensured thorough and focused assessments, avoiding distractions from irrelevant checks. The method demonstrated a strong capacity for early error detection, contributing to a more robust and scalable evaluation framework for question stem.

3.1.7 Role-Play Prompting. Role-Play Prompting involves assigning the model a specific role, such as a subject-matter expert, to guide its behavior and align responses with domain-specific expectations [9]. This method improves task-specific performance by embedding contextual relevance into the prompt. The prompt directed the model to evaluate questions according to categorized rules, simulating the decision-making process of a professional psychometrician.

> **Prompt Excerpt**
>
> Role:
> You are a renowned Certification Exam Psychometrician tasked with safeguarding the integrity and efficacy of the certification exams. Your expertise is crucial in maintaining the highest quality standards for the examination process. Today, you will perform a critical analysis of several exam questions to ensure they meet defined strict criteria.
>
> Task:
> As the psychometrician, your role involves meticulously evaluating each exam question to identify those that are unsuitable for use, referred to as Level 1 questions. You will apply your expertise to examine each question against the comprehensive quality criteria. These criteria are designed to ensure that the questions are relevant, sufficiently challenging, and align with the expected knowledge and skills of the candidates.
> ...

3.1.8 Rank Prompting. The Rank Prompting approach was utilized to evaluate question stems by systematically generating and comparing multiple reasoning paths for each rule [7]. The primary objective was to identify unsuitable questions based on violations of rigorous quality rules. For each rule, the model generated at least three distinct reasoning paths and ranked them based on rule categories. For example, Rule 1 evaluated whether the question stem was clear and free of answer hints. The model generated different interpretations of the stem's clarity and ranked these paths to identify the most consistent and logical assessment. The technique's efficiency was evident in the early identification of low-quality questions, reducing redundant evaluations and streamlining the process.

All prompt engineering techniques were applied under identical experimental conditions on the same labeled dataset, ensuring fair comparison. By standardizing evaluation across Zero Shot, Few Shot, Chain of Thought, Self Consistency, Tree of Thought, Role Prompting, and Rank Prompting, the effect of each technique on classification accuracy was isolated.

3.2 Retrieval Augmented Generation

Retrieval Augmented Generation techniques were evaluated to improve retrieval quality and support more accurate response generation[25]. The approach involved systematic variation of chunk size and chunk overlap to identify effective context windows for retrieval. Chunk configuration directly affects retrieval performance, as larger chunks provide broader context but may introduce irrelevant information, while smaller chunks improve precision at the risk of losing contextual continuity.

Multiple retrieval strategies were examined, including keyword based search, vector based search, Maximal Marginal Relevance, and hybrid search methods. Keyword based retrieval performs well

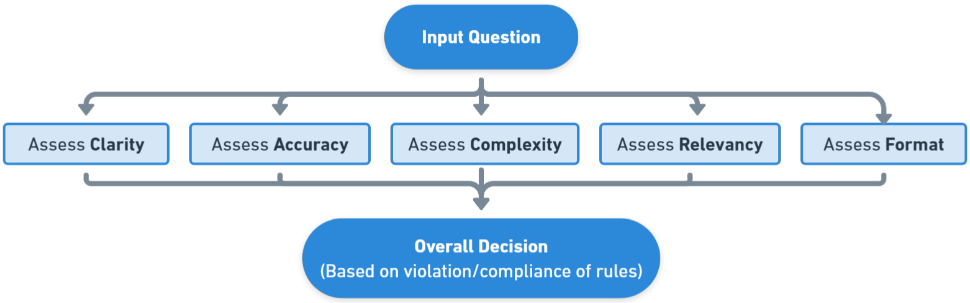

Fig. 3. Tree of Thought (ToT) Breadth-First Search (BFS) Prompt Flow

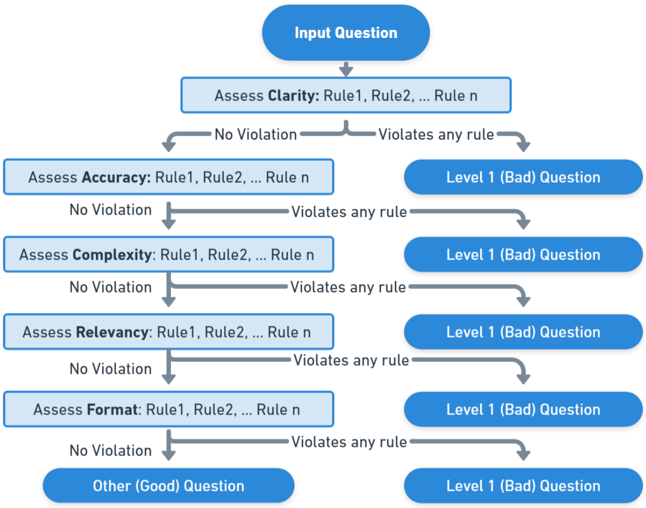

Fig. 4. Tree of Thought (ToT) Depth-First Search (DFS) Prompt Flow

for structured queries but often lacks semantic depth. Vector based retrieval improves semantic matching through dense embeddings but may be less effective for specialized domain queries. Maximal Marginal Relevance balances relevance and diversity, reducing redundancy in retrieved results. Hybrid search combines keyword and vector based methods to leverage the strengths of both approaches. These retrieval designs align with recommendations from Rankify, which integrates dense and sparse retrievers with re ranking for systematic evaluation of Retrieval Augmented Generation pipelines [1].

Recent studies indicate that standard Retrieval Augmented Generation architectures face challenges when handling long context inputs [27]. LongRAG introduces a dual perspective retrieval strategy that improves context selection and response synthesis. Motivated by this work, re ranking strategies were evaluated to improve retrieval quality and reduce hallucinations. Re ranking prioritizes relevant documents and limits the inclusion of misleading context. Prior research such as RankRAG [23] and DSLR [8] demonstrates that fine grained ranking improves factual consistency while preserving informativeness. Additional work shows that attention patterns in large language models can support efficient zero shot re ranking with low latency [2]. The experiments applied these insights to improve factual accuracy in generated responses.

Different embedding models were further evaluated, and the Chroma indexer was employed to support efficient retrieval at scale. Embedding selection plays a critical role in capturing semantic relationships, while the Chroma indexer enables scalable indexing and retrieval across large text corpora.

Overall, these experiments provide insights into effective retrieval configuration for item classification tasks. By systematically evaluating retrieval strategies, chunking methods, embedding models, and re ranking techniques, this study contributes to the development of a robust and scalable Retrieval Augmented Generation pipeline for large language model driven classification.

3.3 Modularized Approach

During the initial prompt engineering and Retrieval Augmented Generation experimentation using a single system prompt, all rules were incorporated into one large prompt. While this approach was straightforward to implement, it presented several challenges: (1) the prompt became excessively long, straining the model's attention and context window, (2) multiple nuanced rules had to be summarized or merged, sometimes causing a loss in granularity, and (3) the large language model (LLM) often tended to overlook or skip certain rules due to the sheer volume of information. These limitations motivated the development of a modularized approach, in which the rules were divided across five distinct prompts, each dedicated to a specific category. By modularizing the rules, the approach ensures that the large language model addresses each requirement in a more systematic and accurate manner.

3.3.1 Approach Design. The modularized approach divides the overarching "question quality" rules into five categories:

(1) **Accuracy:** Focuses on ensuring factual correctness and alignment with the intended learning outcomes.

(2) **Clarity:** Addresses linguistic clarity, unambiguous language, and straightforward question stems.

(3) **Complexity:** Assesses whether the question's difficulty and cognitive load align with the targeted skill level.

(4) **Format:** Looks at formatting guidelines, such as question structure, choice layout, and readability.

(5) **Relevancy:** Evaluates the question's pertinence to the course content or learning objectives.

For each of these categories, a separate prompt was developed, which includes: 1) A concise set of rules specific to that category (e.g., "There is no technical error in question stem" under Accuracy); 2) Instructions on how to determine compliance or violation of each rule; and 3) Examples of compliance or violation if necessary. By doing so, the LLM processes a reduced set of instructions per prompt, lowering the likelihood of ignoring or conflating certain rules.

3.3.2 Implementation Details. According to Figure 5:

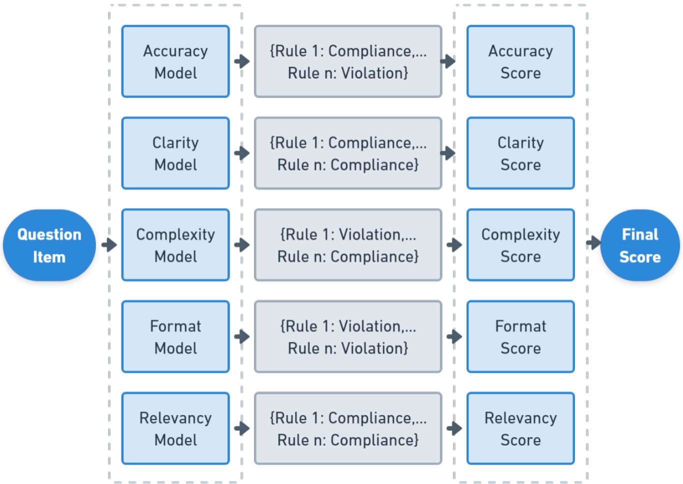
Question Item
Accuracy Model
Clarity Model
Complexity Model
Format Model
Relevancy Model
{Rule 1: Compliance,... Rule n: Violation}
{Rule 1: Compliance,... Rule n: Compliance}
{Rule 1: Violation,... Rule n: Compliance}
{Rule 1: Violation,... Rule n: Violation}
{Rule 1: Compliance,... Rule n: Compliance}
Accuracy Score
Clarity Score
Complexity Score
Format Score
Relevancy Score
Final Score

Fig. 5. Flow Chart of Modularized Approach

(1) **Five-Prompt Evaluation:** Each question (or item) is sequentially passed through five prompts—one for each category. The model outputs a verdict of either “violation” or “compliance” for each relevant rule under that category.

(2) **Rule Violation/Compliance Scoring:** Each category (Accuracy, Clarity, Complexity, Format, Relevancy) yields a category score. The default for a fully compliant category is a maximum score (category-specific maximum). For every rule violated:
   (a) A penalty is applied that reduces the category score.
   (b) An “easy-to-fix” factor (ranging from 1 to 3) is considered to scale the penalty. Rules that are deemed easier to fix incur a smaller penalty than more complex or critical rules.
   (c) If a rule is a “deal breaker,” violation of that rule sets the entire category’s score to 0, reflecting the significance of that rule.

(3) **Summation of Scores:** Once the LLM has evaluated compliance/violation across all five categories, a Final Score was computed by summing the 5 individual category scores. This Final Score provides a holistic view of the item’s overall quality.

(4) **Final Classification Decisions:** Drawing on insights from subject matter experts, specifically content authors, thresholds and conditions were defined to classify items as Level 1 bad questions or Acceptable. The classification logic, based on grading criteria provided by content authors, includes conditions such as:
   (a) **Score Thresholds:** Different thresholds are imposed for both the overall Final Score and individual category scores. For example, if the Final Score falls below a defined cutoff, such as 56 or 58, and category scores in Accuracy or Format drop below predefined thresholds, the item is automatically classified as Level 1.
   (b) **Deal Breakers:** If any deal-breaker rule is violated (resulting in a 0 score in any category), the item is flagged as Level 1.

By combining these rules, penalties, and threshold checks, a structured and transparent classification process is established, enabling precise identification of the category or categories that require improvement.

The modularized approach improves maintainability by organizing rules into category specific prompts, simplifying updates and ongoing management. It provides granular diagnostic feedback that enables item writers to identify problematic categories and address issues more efficiently, while preserving detailed rule coverage and subject matter expert priorities. The structure also supports scalability by allowing new categories or rule changes to be incorporated without redesigning the entire system.

4. EXPERIMENTS AND EVALUATION RESULTS

4.1 Experimental Data and Evaluation Design

To evaluate the effectiveness of the proposed classification approaches across diverse scenarios, the study employs three standard evaluation metrics, Precision, Recall, and F1 Score, computed against a labeled dataset of certification exam questions. The dataset contains over 520 questions drawn from 20 Microsoft technical certification exams, with each exam contributing between 20 and 40 questions. These exams span six major solution areas, including Infrastructure, Data & AI, Digital & App Innovation, Business Applications, Modern Work and Security. This design enables evaluation across varied technical domains, question formats, and cognitive complexity levels.

Each question was independently reviewed and labeled by subject matter experts using predefined quality criteria that include accuracy, clarity, complexity, format, and relevancy. Based on these criteria, questions were classified as either Level 1 or Acceptable. These expert annotations serve as ground truth labels for all reported performance metrics, ensuring that evaluation reflects real world assessment standards.

Rather than treating the dataset as a single uniform corpus, the evaluation design explicitly incorporates variability across exams, solution areas, and content styles. The selected certification exams represent distinct technical contexts and authoring conventions, allowing performance to be assessed under multiple realistic usage scenarios. As a result, the reported metrics capture model behavior across heterogeneous datasets that differ in language structure, domain constraints, and assessment intent.

4.1.1 Iterative and Multi Scenario Evaluation Strategy. An iterative and multi scenario evaluation strategy was adopted to strengthen robustness and generalizability. Initial prompt engineering and retrieval configurations were evaluated on smaller subsets ranging from 50 to 100 questions. These subsets were intentionally sampled across different certification exams and solution areas to ensure coverage of diverse question types, difficulty levels, and domain specific constraints.

Findings from early stage evaluations informed refinements to prompt structures, retrieval parameters, and classification logic before scaling to larger subsets. Once stable performance trends were observed across multiple subsets and scenarios, the most effective configurations were evaluated on the full dataset of 520 questions. This staged approach enabled systematic identification of failure cases and reduced the risk of overfitting to specific exams or question styles.

By repeatedly evaluating performance across exams from different solution areas at each stage, the study demonstrates that observed improvements are consistent across multiple datasets and assessment scenarios. Performance gains were found to be stable across domains with varying technical focus and linguistic structure, supporting the applicability of the proposed framework to large scale and diverse certification assessment settings.

4.1.2 Experimental Setup and Model Configuration. All experiments were conducted using the Azure OpenAI Service, primarily leveraging the GPT-4.0 model. To ensure comparability across prompt engineering techniques, inference parameters were held constant across all runs. The temperature was set to 0.5 to balance determinism with controlled variability, the maximum token limit was capped at 1,500, top-p was set to 1.0, and both frequency and presence penalties were fixed at 0.5.

Each experimental configuration was evaluated three times to mitigate stochastic variation, and all reported results represent averages across runs. In addition to prompt-only experiments, Retrieval-Augmented Generation (RAG) configurations were systematically evaluated by varying chunk sizes (ranging from 1,000 to 8,191 tokens), overlap values (0 to 150 tokens), and retrieval strategies, including Keyword-based search, Vector-based search, Maximal Marginal Relevance (MMR), and Hybrid approaches with reranking.

For retrieval experiments, both Ada-002 and MiniLM-L6-v2 embedding models were employed, with all embeddings indexed using the Chroma vector database. This experimental design enabled controlled comparison of prompt engineering techniques, retrieval strategies, and modular evaluation approaches under consistent conditions.

4.2 Definitions and Importance of Evaluation Metrics

(1) **Recall:** Recall measures the proportion of problematic questions correctly identified by the model. High recall is critical to prevent flawed items from compromising exam validity and learner outcomes.

(2) **Precision:** Precision indicates the proportion of questions classified as problematic that are truly problematic. High precision minimizes unnecessary manual review by reducing false positive classifications.

(3) **F1 Score:** The F1 score combines precision and recall into a single metric, providing a balanced assessment of overall classification performance when both false negatives and false positives are costly.

Given the high stakes of exam reliability, recall is slightly prioritized over precision in the final selection. Identifying as many flawed questions as possible (even if it means some additional manual checks for false positives) is generally deemed more acceptable than letting defective questions remain in circulation.

4.3 Comprehensive Prompt Engineering Results

Table 1 summarizes the performance of various prompt engineering methods, which gives the average scores of all the testing exams:

On the prompt engineering front, Role Prompting combined with Few-Shot examples and Self-Consistency in Chain of Thought (SC-CoT) reasoning delivered the best performance, attaining similar F1 scores. Both methods underscore the value of combining innovative strategies and precise configurations for achieving optimal results in their respective applications. Considering the higher Recall score of SC-CoT prompt, it was selected for further RAG experiments.

Table 1. Prompt Engineering Methods Performance

Method	**Prec. (%)**	**Rec. (%)**	**F1 (%)**
Zero-Shot Prompting	15.38	4.65	7.14
Few-Shot Prompting (1)	29.03	20.93	24.32
Few-Shot Prompting (2)	28.57	23.26	25.64
CoT Prompting	23.08	6.98	10.72
CoT + Few-shot (Self-Consistency)	**38.10**	**55.81**	**45.29**
ToT (BFS)	15.38	4.65	7.14
ToT (DFS)	13.33	4.65	6.89
ToT (BFS) + Few-shot	25.00	20.93	22.78
Role Prompting	25.00	16.28	19.72
Role + Few-shot	41.07	53.49	46.46
Rank + Few-shot	29.79	32.56	31.11

4.4 RAG Experimental Results

Table 2 is a cross-tabulation of the RAG experimental results on SC-CoT prompt, highlighting the performance metrics achieved under each configuration.

For retrieval techniques, Hybrid (0.5-0.5) Search + Re-ranking with Ada-002 embedding and Chroma indexer at a chunk size of 8191 and overlap of 150 emerged as the top configuration, achieving the highest F1 score of 56.88%, with a precision of 45.21% and recall of 76.74%. The balanced hybrid search weights, re-ranking, and advanced embedding methods proved highly effective.

4.5 Modularized Approach Results

Table 3 presents the performance of the modularized evaluation approach:

The Modularized Approach achieved strong overall performance, evidenced by a high recall of 88.89% and a balanced F1 score of 66.67%, despite a moderate precision of 53.33%. This indicates that while the system occasionally includes some false positives, it is exceptionally effective at capturing the majority of relevant instances. The balanced F1 score underscores that the benefits of high recall outweigh the trade-offs in precision, making the Modularized Approach a robust solution for applications where comprehensive retrieval is paramount.

5. DISCUSSION

5.1 Interpretation of Results

In this study, a systematic approach was established to address the classification task, emphasizing the importance of precise and efficient methods in automated rule identification. The process began with extensive experimentation on various prompt engineering techniques to determine the most effective structure for enhancing classification accuracy. Through iterative testing, the Self Consistency Chain of Thought prompt was identified as the most effective approach, as it encourages the model to explicitly reason through its decisions, thereby reducing ambiguity and improving interpretability. As a result, this method emerged as the most suitable choice for complex classification scenarios.

Table 2. RAG Experimental Results

Chunk Size	Overlap	Search Method	Embedding	Indexer	Precision (%)	Recall (%)	F1 Score (%)
1000	0	Keyword	MiniLM-l6-v2	Chroma	12.50	4.65	7.78
1000	0	Vector	MiniLM-l6-v2	Chroma	32.73	41.86	36.73
4096	150	MMR	MiniLM-l6-v2	Chroma	34.55	44.19	38.78
8191	150	MMR	Ada-002	Chroma	38.60	51.16	40.00
8191	**150**	**Hybrid (0.5–0.5) + Re-ranking**	**Ada-002**	**Chroma**	**45.21**	**76.74**	**56.88**
8191	150	Hybrid (0.8–0.2) + Re-ranking	Ada-002	Chroma	40.63	60.47	48.56

Table 3. Modularized Approach Results

Approach	Precision (%)	Recall (%)	F1 Score (%)
Modularized Approach	53.33	88.89	66.67

Following the prompt tuning experiments, different parameter configurations for Retrieval Augmented Generation were systematically explored to optimize retrieval accuracy. By evaluating multiple configurations, RAG was fine tuned to ensure that generated responses were both contextually accurate and relevant, thereby strengthening the overall classification framework.

Building upon these advancements, a modularized approach to question classification was introduced. This approach segmented the classification task into discrete categories aligned with specific rule categories, streamlining the identification and rectification of rule violations. For content authors and domain experts, this modular strategy provides a clear and actionable framework, allowing them to efficiently pinpoint and resolve specific issues without navigating cumbersome or monolithic systems.

Another critical aspect of the applied methodology was the iterative nature of experimentation. Rather than applying each method to the dataset in a single pass, prompts were refined through multiple smaller scale trials before being scaled to the full dataset. This iterative design allowed us to (a) detect performance bottlenecks early, (b) incorporate improvements dynamically, and (c) converge toward stronger final results. This stepwise refinement not only strengthens the validity of the reported metrics but also enhances the reproducibility of this study for future researchers.

Overall, the findings highlight the effectiveness of this sequential approach—progressing from prompt tuning to RAG optimization and culminating in a modular classification framework. This structured methodology delivers accurate, interpretable, and adaptable classification solutions tailored for dynamic environments.

5.2 Further Application and Impact

The proposed item classification framework has been integrated into the Microsoft Learn assessment authoring workflow. By automating the evaluation of exam questions across key quality criteria including accuracy, clarity, complexity, format, and relevance, the system streamlines item review and provides targeted feedback to item authors. This integration has reduced manual review effort for subject matter experts and improved both efficiency and scalability within the authoring process.

The framework currently supports the development and maintenance of hundreds of assessment items across nearly ninety certifications. Early evaluations indicate that approximately forty percent of AI generated items are suitable for practice assessments without significant revision, while the system effectively identifies most problematic items through high recall. Category specific feedback enables item authors to resolve issues more efficiently, contributing to improved quality and consistency across Microsoft Certification assessments.

5.3 Broader Implications

The techniques introduced in this study extend beyond technical certification exams. The modularized prompt engineering approach can be applied to a wide range of text classification tasks, including content moderation, customer support routing, and automated grading. Retrieval Augmented Generation techniques may also enhance systems that require accurate retrieval and classification of textual information, such as search and recommendation platforms. Furthermore, this work emphasizes the value of incorporating domain specific rules into large language model workflows, which is particularly important in regulated fields such as healthcare, legal services, and finance.

6. CONCLUSION AND FUTURE WORK

This study introduces a novel, sequentially structured roadmap for optimizing question classification tasks using Large Language Models (LLMs), emphasizing an innovative modularized prompt engineering framework. The research methodology followed a clear sequential roadmap: initially focusing on identifying the optimal prompt structure through rigorous prompt engineering, subsequently refining Retrieval-Augmented Generation (RAG) configurations, and finally developing and validating a modularized approach. Extensive experimentation confirmed that this sequential, modularized strategy delivered substantial improvements in precision, recall, and F1 scores compared to conventional single-prompt methods. The structured, modular design not only enhances accuracy and interpretability but also significantly simplifies rule violation detection and correction. This empowers content authors and domain experts to efficiently address specific rule violations and integrate evolving rules or additional categories seamlessly.

Future research directions include:

(1) Expanding to Broader Domains: Applying the modularized approach to diverse scenarios, such as sentiment analysis, intent detection, and customer feedback analysis, to assess and validate generalizability.

(2) Active Learning Integration: Exploring active learning methods to further enhance prompt optimization, minimizing the manual effort required to maintain and update classification systems.

(3) Model Efficiency through Distillation: Employing distillation techniques to create lightweight models that retain classification performance while reducing computational demands and improving inference efficiency.

Ultimately, this research establishes a valuable foundation for prompt engineering methodologies and RAG optimization and encourages further exploration into efficient, adaptable, and interpretable classification strategies using LLMs.

7. REFERENCES

[1] Abdelrahman Abdallah, Bhawna Piryani, Jamshid Mozafari, Mohammed Ali, and Adam Jatowt. Rankify: A comprehensive python toolkit for retrieval, re-ranking, and retrieval-augmented generation, 2025.

[2] Shijie Chen, Bernal Jiménez Gutiérrez, and Yu Su. Attention in large language models yields efficient zero-shot re-rankers, 2024.

[3] ChunLiu ChunLiu, Hongguang Zhang, Kainan Zhao, Xinghai Ju, and Lin Yang. Llmembed: Rethinking lightweight llm's genuine function in text classification. *arXiv (Cornell University)*, pages 7994–8004, 01 2024.

[4] Aleksandra Edwards and Jose Camacho-Collados. Language models for text classification: Is in-context learning enough? *ACL Anthology*, pages 10058–10072, 05 2024.

[5] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. GPTScore: Evaluate as you desire. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[6] Ali Hakimi Parizi, Yuyang Liu, Prudhvi Nokku, Sina Gholamian, and David Emerson. A comparative study of prompting strategies for legal text classification. In Daniel Preoțiuc-Pietro, Catalina Goanta, Ilias Chalkidis, Leslie Barrett, Gerasimos Spanakis, and Nikolaos Aletras, editors, *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 258–265, Singapore, December 2023. Association for Computational Linguistics.

[7] Chi Hu, Yuan Ge, Xiangnan Ma, Hang Cao, Qiang Li, Yonghua Yang, Tong Xiao, and Jingbo Zhu. Rankprompt: Step-by-step comparisons make language models better reasoners, 2024.

[8] Taeho Hwang, Soyeong Jeong, Sukmin Cho, SeungYoon Han, and Jong C Park. Dslr: Document refinement with sentence-level re-ranking and reconstruction to enhance retrieval-augmented generation, 07 2024.

[9] Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, and Xin Zhou. Better zero-shot reasoning with role-play prompting, 08 2023.

[10] Alice Kwak, Clayton Morrison, Derek Bambauer, and Mihai Surdeanu. Classify first, and then extract: Prompt chaining technique for information extraction. In Nikolaos Aletras, Ilias Chalkidis, Leslie Barrett, Cătălina Goanță, Daniel Preoțiuc-Pietro, and Gerasimos Spanakis, editors, *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 303–317, Miami, FL, USA, November 2024. Association for Computational Linguistics.

[11] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models, 11 2022.

[12] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries, 07 2004.

[13] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55, 09 2022.

[14] Robert Logan IV, Ivana Balazevic, Eric Wallace, Fabio Petroni, Sameer Singh, and Sebastian Riedel. Cutting down on prompts and parameters: Simple few-shot learning with language models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2824–2835, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[15] Soumya Mishra. ESG impact type classification: Leveraging strategic prompt engineering and LLM fine-tuning. In Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, Hsin-Hsi Chen, Hiroki Sakaji, and Kiyoshi Izumi, editors, *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing*, pages 72–78, Bali, Indonesia, November 2023. Association for Computational Linguistics.

[16] Gabriel Orlanski. Evaluating prompts across multiple choice tasks in a zero-shot setting, 2022.

[17] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 2001.

[18] Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. Mauve: Measuring the gap between neural text and human text using divergence frontiers, 2021.

[19] Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Hyperscience Teven, Le Scao, Arun Raja, Manan Dey, Sap Saiful, Bari Ntu, Singapore Xu, Urmish Thakker, Shanya Sharma, Walmart Labs, Eliza Szczechla, Bigscience Taewoon, Kim Vu, Amsterdam Gunjan, Chhablani Bigscience, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian, Jiang Zeals, Japan Wang, Matteo Manica, Sheng Shen, U Berkeley, Zheng-Xin Yong, Harshit Pandey, Bigscience Michael, Mckenna Parity, Rachel Inria, France Thomas, Wang Inria, France Trishala, Neeraj Bigscience, Jos Rozen, Andrea Santilli, Thibault Fevry, Bigscience Jason, Alan Fries, Snorkel Ai, Ryan Teehan, Charles River, Analytics Bers, Stella Biderman Booz, Eleutherai Leo, Gao Eleutherai, Thomas Wolf, and Alexander Rush. Multitask prompted training enables zero-shot task generalization, 03 2022.

[20] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv:2203.11171 [cs]*, 10 2022.

[21] Yuqi Wang, Wei Wang, Qi Chen, Kaizhu Huang, Anh Nguyen, and Suparna De. Prompt-based zero-shot text classification with conceptual knowledge. In Vishakh Padmakumar, Gisela Vallejo, and Yao Fu, editors, *Proceedings of the*

61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 30–38, Toronto, Canada, July 2023. Association for Computational Linguistics.

[22] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models, 05 2023.

[23] Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. Rankrag: Unifying context ranking with retrieval-augmented generation in llms, 2024.

[24] Lei Zhang, Yunshui Li, Ziqiang Liu, Jiaxi Yang, Junhao Liu, Longze Chen, Run Luo, and Min Yang. Marathon: A race through the realm of long context with large language models. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5201–5217, 2024.

[25] Tianjun Zhang, Shishir G Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E Gonzalez. Raft: Adapting language model to domain specific rag. *arXiv (Cornell University)*, 03 2024.

[26] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv:1904.09675 [cs]*, 02 2020.

[27] Qingfei Zhao, Ruobing Wang, Yukuo Cen, Daren Zha, Shicheng Tan, Yuxiao Dong, and Jie Tang. Longrag: A dual-perspective retrieval-augmented generation paradigm for long-context question answering. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22600–22632, 2024.

[28] Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. *arXiv:2102.09690 [cs]*, 06 2021.

[29] Yujia Zhou, Yan Liu, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Zheng Liu, Chaozhuo Li, Zhicheng Dou, Tsung-Yi Ho, and Philip S Yu. Trustworthiness in retrieval-augmented generation systems: A survey, 2024.