First Ask Then Answer: A Framework Design for AI Dialogue Based on Supplementary Questioning with Large Language Models

Anonymous EMNLP submission

Abstract

Large Language Models (LLMs) often struggle to deliver accurate and actionable answers when user-provided information is incomplete or ill-specified. We propose a new interaction paradigm, *First Ask Then Answer* (FATA), in which, through prompt words, LLMs are guided to proactively pose multidimensional supplementary questions to users before answering. Then, using the information received by the users and the original questions to jointly construct prompt words for questioning, a high-quality question-and-answer is ultimately achieved.In contrast to existing clarification approaches—such as the CLAM framework oriented to ambiguity and the self-interrogation Self-Ask method—FATA emphasizes completeness (beyond mere disambiguation) and user participation (inviting human input instead of relying solely on model-internal reasoning). It also adopts a single-turn strategy: all clarifying questions are produced at once, thereby reducing dialogue length and improving efficiency. Conceptually, FATA uses the reasoning power of LLMs to scaffold user expression, elevating ordinary users to an expert-level questioning ability. To evaluate FATA, we constructed a multi-domain benchmark and compared it with two controls: a baseline prompt (B-Prompt) and a context-enhanced expert prompt (C-Prompt). Experimental results show that FATA outperforms B-Prompt by $\sim 40\%$ in aggregate metrics and exhibits a coefficient of variation 8% lower than C-Prompt, indicating superior stability.

1 Introduction

In recent years, LLMs have shown excellent performance on single-round tasks. However, real-world users often cannot describe the complete context as well as domain experts: health consultation misses drug doses, government administration ignores budget constraints, or programming questions miss error logs. If LLMs attempt to answer despite gaps in the user's information, they risk hallucinations and produce irrelevant results, which slows decision-making and undermines user trust. Existing research generally follows three paths:

- **Selective clarification** — the CLAM framework first detects ambiguity and only queries the user when a threshold is crossed [Kuhn and et al., 2022]. Follow-up question prediction further reduces the number of clarification turns. [Zhang et al., 2024]. Undetected gaps still lead to wrong answers reaching end-users.
- **In-model chain-of-thought** — Chain-of-Thought prompting [Wei et al., 2022] and its variants Self-Consistency [Wang et al., 2022] and Tree-of-Thought [Yao et al., 2023] improve explicit reasoning. Self-Ask lets the model generate and answer sub-questions internally before concluding [Press and et al., 2022]. These approaches depend on internal chain-of-thought reasoning or external retrieval and do not facilitate direct interaction with the user's background.
- **Reason–act coupling** — ReAct [Yao and et al., 2022], Toolformer [Schick and et al., 2023], MRKL [Karpas and et al., 2022], and Planner-Executor [Deng and et al., 2025] interleave reasoning with tool-calls or planning. They excel at retrieval and analysis but assume a fully specified input context and often require additional self-refinement cycles.

To reconcile the tension between incomplete information and interaction overhead, we introduce FATA: prior to providing an answer, the LLMs, from an expert's perspective, produce a structured checklist of additional questions covering multiple dimensions. After the user responds, the LLMs generate a personalized solution. FATA offers four core advantages:

1. **Information completeness** – transforms uncertain problem space into answerable parameters, thereby reducing search entropy.

2. **Single-step interaction** – avoids prolonged multi-turn dialogues and context drift.

3. **Ease of deployment** – prompt-only; no fine-tuning required.

4. **Tool-agnostic extensibility** – after entropy reduction, FATA can seamlessly attach any tools or agents.

2 Overview of the First-Ask-Then-Answer (FATA) Framework

The *First-Ask-Then-Answer* (FATA) framework introduces a novel dialogue paradigm that emphasizes the structured generation of supplementary questions before providing answers. Unlike existing frameworks like the CLAM model, which only triggers clarification when ambiguity is detected, FATA proactively generates a comprehensive set of questions for all complex or incomplete queries. These questions are designed to maximize information gain by collecting key details in one go. The entire process is driven by publicly reproducible prompt templates and does not require any fine-tuning.

We introduce the concept of **FATA-Prompt**, a specific set of prompts tailored for this framework. An example of the FATA prompt is as follows:

```
User request: <original query> \ To better assist me, before offering advice, please adopt the perspective of an expert in the relevant field and ask questions to help you identify any missing key information. Please ensure the problem is structured clearly and expressed concisely, with example guidance. Just like how experts ask users questions during consultations to gather key information before providing solutions. After I provide additional information, please then offer a more personalized and practical solution as an expert in that field. If all key information has already been provided, please directly give the solution.Note: Maintain a positive attitude, and do not request phone numbers, ID numbers, or other sensitive data.
```

2.1 Prompt Components

The functions and combinations of prompt words are as follows:

1 Determine the domain & information gap: To better assist me, before offering advice, please adopt the perspective of an expert in the relevant field and ask questions to help you identify any missing key information. *Function:* Role positioning + integrity check.

2 Optimize output structured problem: Please ensure the problem is structured clearly and expressed concisely, with example guidance. *Function:* Facilitate users' understanding and quick filling.

3 Provide examples: Just like how experts ask users questions during consultations to gather key information before providing solutions. *Function:* Clarify the questioning style.

4 Generate personalized solutions: After I provide additional information, please then offer a more personalized and practical solution as an expert in that field. *Function:* Ensure answers are relevant and feasible.

5 Reduce redundancy: If all key information has already been provided, please directly give the solution. *Function:* Avoid unnecessary interactions.

6 Privacy & Tone: Note: Maintain a positive attitude, and do not request phone numbers, ID numbers, or other sensitive data. *Function:* Protect users + Encourage.

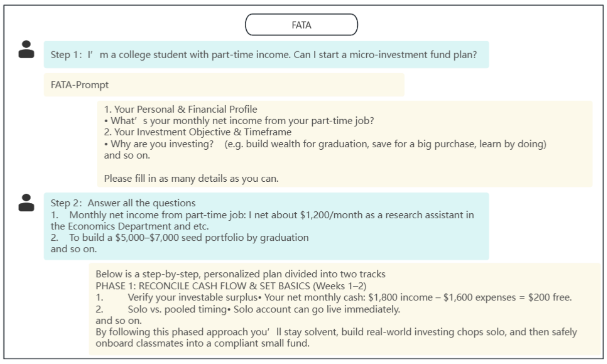
The core point of the FATA framework lies in items 1–4: by automatically determining information gaps, performing single-round structured supplementary questioning, and integrating context to generate answers, it significantly improves the quality of multi-domain dialogues without additional training, while maintaining high human–computer interaction efficiency.

2.2 User-Side Workflow

From the user's perspective, interaction requires only two steps:

1. Submit the original question with FATA prompt words (system supplement).

2. Answer each supplementary question one by one based on the model's question list. This low learning cost yields high answer gains for end users.



Step 1: I' m a college student with part-time income. Can I start a micro-investment fund plan?

FATA-Prompt

1. Your Personal & Financial Profile
• What' s your monthly net income from your part-time job?
2. Your Investment Objective & Timeframe
• Why are you investing? (e.g. build wealth for graduation, save for a big purchase, learn by doing)
and so on.

Please fill in as many details as you can.

Step 2: Answer all the questions
1. Monthly net income from part-time job: I net about $1,200/month as a research assistant in the Economics Department and etc.
2. To build a $5,000–$7,000 seed portfolio by graduation
and so on.

Below is a step-by-step, personalized plan divided into two tracks
PHASE 1: RECONCILE CASH FLOW & SET BASICS (Weeks 1–2)
1. Verify your investable surplus• Your net monthly cash: $1,800 income – $1,600 expenses = $200 free.
2. Solo vs. pooled timing• Solo account can go live immediately.
and so on.
By following this phased approach you' ll stay solvent, build real-world investing chops solo, and then safely onboard classmates into a compliant small fund.

Figure 1: User perspective using the FATA process (simplified answer version).

2.3 Question Generation & Answering

After applying FATA cue words, LLMs generate a set of supplementary questions. In practice, these questions are often presented as numbered lists that cover the various dimensions of information needed to complete a user's task. The number of questions proposed by the model depends on the complexity of the task, generally about 3-7, to avoid too many questions to increase the burden of users and to cover the main aspects. From the perspective of information entropy theory, multiple well-targeted supplementary questions are equivalent to obtaining additional information about the user's intention, which significantly reduces the entropy value of the answer search space. Multiple questions presented at the same time also facilitate the user to consider them as a whole, providing more consistent and coherent answers.

This mechanism explains why FATA achieves robust quality improvement on problems in different domains - each complementary question interaction guides the model to converge to the correct solution. At the same time, the responsibility can be traced, and the three steps of initial answer-follow-up question-final answer make the source of the error clear. If the answer is still not the question, it can be located as insufficient model reasoning. If the information is missing, the user can be prompted to supplement it again.

2.4 Representative Cases

Health Management Scenario. **User request:** "How to better manage my diabetes?"
FATA questions: Current blood glucose/HbA1c, medication type and dosage, diet/exercise habits, comorbidities.
User reply: "HbA1c 7.5%, metformin 500 mg daily, high intake of staple foods, light exercise once a week, no other complications."
Final recommendation: Tailored dietary plan (reduce refined carbs, alternative recipes), exercise schedule (moderate intensity $\geq 3\times/\text{week}$), medication adherence, and blood pressure monitoring.

Urban Governance Decision-Making Scenario.
User request: "Help me develop a KPI plan to improve the level of urban governance."
FATA questions: Focus areas (e.g. traffic, environment, safety), benchmark data (e.g. AQI, commute time), targets and timeframes, resources/policy priorities.
User reply: Environmental protection and housing; current recycling rate 20%→50% in one year; baseline housing satisfaction; special budget allocated.
Final KPI plan: Increase recycling rate to 50% in 1 year, reduce PM2.5 by X%, boost housing satisfaction score by Y, with implementation measures and policy suggestions.

2.5 Paper Contributions

To summarize, our contributions are as follows:

First, Method Innovation: First Ask Then Answer (FATA), a two-stage paradigm ("supplementary requirements" & "solution generation") that leverages LLMs for both information collection and problem solving.Second, Comparative Analysis: Systematic comparison with CLAM and Self-Ask to delineate their respective strengths and limitations.Third, Experimental Findings: FATA outperforms B-Prompt by $\sim$ 40% on overall metrics and reduces coefficient of variation by 8% compared to C-Prompt.Fourth, General Implications: Demonstrates applicability of FATA across 12 dialogue domains (e.g. education, government, health).

3 Experiments

Due to the fact that the character profiles in the existing datasets cannot contain all the information required for supplementary questions and the structured supplementary questions are often open-ended questions, the evaluation effect is not good. In order to answer the core question of "how can FATA improve the quality of personalized answering in multiple fields", we design three strictly controlled input conditions: 1.The Baseline prompt (B-Prompt) simulates the questioning of key information that users often overlook in the real environment, and is used to test the lower limit performance and examine the model performance when information is lacking. 2. Context-enhanced expert prompt (C-Prompt), which is used to test the upper bound performance and simulate the questions of users who have mastered the ability to ask questions of large models with complete background information. 3. FATA questioning method (Question F) : Simulate the questioning of a user adopting FATA questioning method.All other factors remain exactly the same. In this process, we only need to input B-Prompt, and the large model automatically generates rich profiles for evaluation. The following sections describe the data and model, dataset construction, and evaluation methodology. The necessary prompt words for the specific process are placed in the attachment at the end of the article.

The generation steps all invoke the ChatGPTo4-mini-2025-04-16 model, which is a small model optimized for fast, cost-effective reasoning and suitable for cost-effective use cases. Using the same configuration at each stage ensures that the model can dynamically generate content based on the input context, rather than relying on any hard-coded rules. To avoid information leakage, the generation of Supplementary questions (F1), persona profile and final answer were completed in an independent session. Evaluation was performed using ChatGPT-O3, isolated from the answer model.

3.1 Dataset Construction

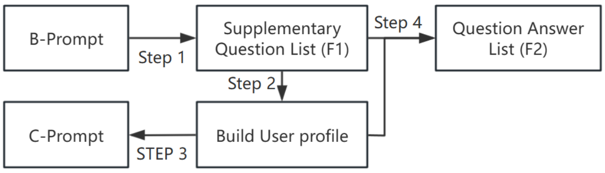
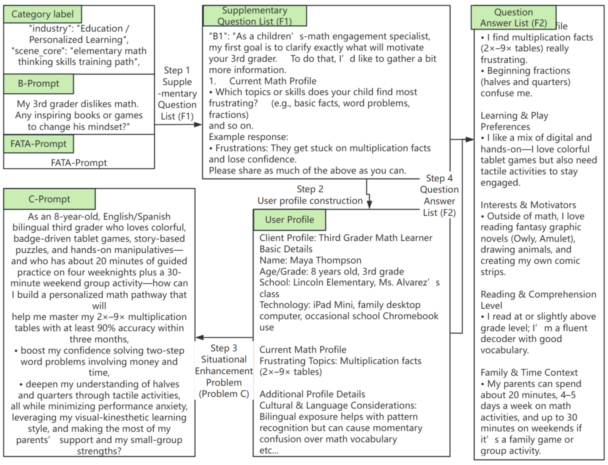


Figure 2: Flowchart of Dataset Construction.

We curated 300 user cases (12 industries × 5 scenarios × 5 B-Prompt variants) with deliberately incomplete information. Each JSON entry has fields: `"industry": "...", "scene_core": "...", "B-Prompt": "..."`

Only `B-Prompt` enters generation; `industry` and `scene_core` guide scene identification.

Supplementary Question list (F1) : The large model generates a set of supplementary question list (F1) by merging questions with B-Prompt and FATA prompt words, which is used to mine key information not yet provided by the user. The goal of this stage is to answer the question, "What other key information do I need to know to better answer B-Prompt?"



"industry": "Education / Personalized Learning", "scene_core": "elementary math thinking skills training path",

My 3rd grader dislikes math. Any inspiring books or games to change his mindset?"

"B1": "As a children' s-math engagement specialist, my first goal is to clarify exactly what will motivate your 3rd grader. To do that, I' d like to gather a bit more information.
1. Current Math Profile
• Which topics or skills does your child find most frustrating? (e.g., basic facts, word problems, fractions)
and so on.
Example response:
• Frustrations: They get stuck on multiplication facts and lose confidence.
Please share as much of the above as you can.

Client Profile: Third Grader Math Learner
Basic Details
Name: Maya Thompson
Age/Grade: 8 years old, 3rd grade
School: Lincoln Elementary, Ms. Alvarez' s class
Technology: iPad Mini, family desktop computer, occasional school Chromebook use

Current Math Profile
Frustrating Topics: Multiplication facts (2×–9× tables)

Additional Profile Details
Cultural & Language Considerations: Bilingual exposure helps with pattern recognition but can cause momentary confusion over math vocabulary
etc...

As an 8-year-old, English/Spanish bilingual third grader who loves colorful, badge-driven tablet games, story-based puzzles, and hands-on manipulatives—and who has about 20 minutes of guided practice on four weeknights plus a 30-minute weekend group activity—how can I build a personalized math pathway that will
help me master my 2×–9× multiplication tables with at least 90% accuracy within three months,
• boost my confidence solving two-step word problems involving money and time,
• deepen my understanding of halves and quarters through tactile activities,
all while minimizing performance anxiety, leveraging my visual-kinesthetic learning style, and making the most of my parents' support and my small-group strengths?

• I find multiplication facts (2×–9× tables) really frustrating.
• Beginning fractions (halves and quarters) confuse me.

Learning & Play Preferences
• I like a mix of digital and hands-on—I love colorful tablet games but also need tactile activities to stay engaged.

Interests & Motivators
• Outside of math, I love reading fantasy graphic novels (Owly, Amulet), drawing animals, and creating my own comic strips.

Reading & Comprehension Level
• I read at or slightly above grade level; I' m a fluent decoder with good vocabulary.

Family & Time Context
• My parents can spend about 20 minutes, 4–5 days a week on math activities, and up to 30 minutes on weekends if it' s a family game or group activity.

Figure 3: Actual Data Flow (Simplified Version).

User profile construction: For each question in the Supplementary Question List (F1), the model automatically generates a person profile including the user's basic information based on the scene information. The archive is equivalent to simulating the complete information profile built by the user on demand.

Question answering list (F2) : Combining the user profile with the supplementary question list (F1), the large model is required to answer the question one by one from the perspective of the user to obtain the question list answer (F2), which is equivalent to the information content supplemented by the user in the consultation

C-Prompt: The model rewrites the B-Prompt into a more complete context-enhanced expert question based on the user profile, incorporating key information directly into the question to form the question "How would the question be formulated if the user with the ability to ask a large model provided this context in the first place?"

Each output sample contains: user profile, B-Prompt, F1 (supplementary question list), F2 (supplementary answer), C-Prompt.

3.2 Answer Generation

Our core experimental process is divided into three stages to simulate LLMs' behavior under incomplete questions:

1. **FATA Question Answer (Answer F):** Takes the supplementary question list (F1) and their answers (F2) as context, sends a new prompt,

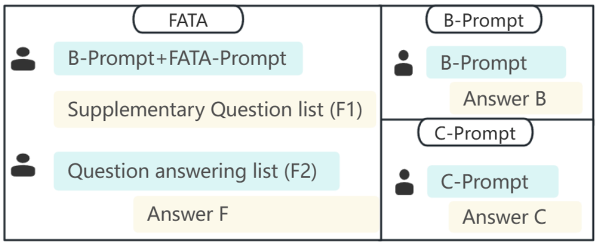


Figure 4: Simplified diagram for specific Answer Generation stages.

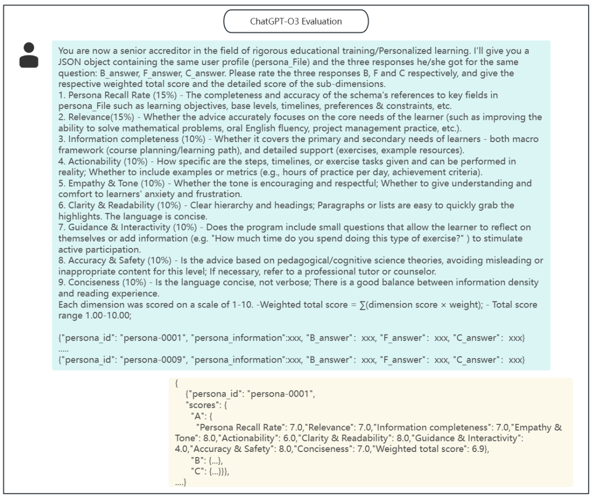
and generates the personalized *Answer F*.

2. **B-Prompt Answer (Answer B):** Generates a direct response to the original B-Prompt without further clarification.

3. **C-Prompt Answer (Answer C):** Generates a response to the C-Prompt, which already includes full context and background details.

These three output forms (Answer B, Answer F, Answer C) allow us to evaluate the impact of different context enrichment methods on answer quality.

4 Evaluation Protocol



You are now a senior accreditor in the field of rigorous educational training/Personalized learning. I'll give you a JSON object containing the same user profile (persona_File) and the three responses he/she got for the same question: B_answer, F_answer, C_answer. Please rate the three responses B, F and C respectively, and give the respective weighted total score and the detailed score of the sub-dimensions.
1. Persona Recall Rate (15%) - The completeness and accuracy of the schema's references to key fields in persona_File such as learning objectives, base levels, timelines, preferences & constraints, etc.
2. Relevance(15%) - Whether the advice accurately focuses on the core needs of the learner (such as improving the ability to solve mathematical problems, oral English fluency, project management practice, etc.).
3. Information completeness (10%) - Whether it covers the primary and secondary needs of learners - both macro framework (course planning/learning path), and detailed support (exercises, example resources).
4. Actionability (10%) - How specific are the steps, timelines, or exercise tasks given and can be performed in reality; Whether to include examples or metrics (e.g., hours of practice per day, achievement criteria).
5. Empathy & Tone (10%) - Whether the tone is encouraging and respectful; Whether to give understanding and comfort to learners' anxiety and frustration.
6. Clarity & Readability (10%) - Clear hierarchy and headings; Paragraphs or lists are easy to quickly grab the highlights. The language is concise.
7. Guidance & Interactivity (10%) - Does the program include small questions that allow the learner to reflect on themselves or add information (e.g. "How much time do you spend doing this type of exercise?") to stimulate active participation.
8. Accuracy & Safety (10%) - Is the advice based on pedagogical/cognitive science theories, avoiding misleading or inappropriate content for this level; If necessary, refer to a professional tutor or counselor.
9. Conciseness (10%) - Is the language concise, not verbose; There is a good balance between information density and reading experience.
Each dimension was scored on a scale of 1-10. -Weighted total score = ∑(dimension score × weight); - Total score range 1.00-10.00;

```
{"persona_id": "persona-0001", "persona_information":xxx, "B_answer":  xxx, "F_answer":  xxx, "C_answer":  xxx}
.....
{"persona_id": "persona-0009", "persona_information":xxx, "B_answer":  xxx, "F_answer":  xxx, "C_answer":  xxx}
```

```
{
  {"persona_id": "persona-0001",
  "scores": {
    "A": {
      "Persona Recall Rate": 7.0,"Relevance": 7.0,"Information completeness": 7.0,"Empathy & Tone": 8.0,"Actionability": 6.0,"Clarity & Readability": 8.0,"Guidance & Interactivity": 4.0,"Accuracy & Safety": 8.0,"Conciseness": 7.0,"Weighted total score": 6.9},
    "B": {...},
    "C": {...}}},
....}
```

Figure 5: Automated evaluation pipeline using ChatGPT-O3.

In this section, we will introduce a fully automated evaluation protocol that employs the ChatGPT-O3 model on the OpenAI website as the reviewer, completing the evaluation process automatically through its API calls.ChatGPT-O3 model on OpenAI's website—OpenAI's most powerful inference model to date, setting records in benchmarks across multiple dimensions: programming, math, science, visual perception, and more. It is used to evaluate the dimensions of relevance, completeness, accuracy, and pertinence of the generated answers.

The data generated for structured supplementary questions are often open-ended. Human evaluation results cannot be easily replicated and are evaluated *ad hoc* in ways that are difficult for external researchers to observe and criticize. Considering the complexity of evaluating user profiles combined with B-Promt, FATA, C-Prompt responses one by one, we adopt a large model to automatically evaluate the performance of selecting and contrasting their responses. The user profile, B_answer, F_answer, and C_answer are organized into a test unit in JSON format, such as:

```
"persona_id": "persona-0001", "persona_information": xxx,
"B_answer": xxx, "F_answer": xxx, "C_answer": xxx
```

Group the characters into batches of 8 to 9 (to avoid long input), and ChatGPT-O3 will rate all the output based on the same rating cue words. This evaluation method can not only avoid the deviation of prompt words, but also greatly improve the efficiency, and the results can be replicated at a low cost. Finally, we count the average score across all scenarios to quantify the overall performance of the different response forms. Scoring requirements vary from field to field, but the dimensions remain consistent.

4.1 Evaluation Considerations

In general, the nine dimensions can be divided into three categories, from "content first, then implementation, and with interaction and style" understanding.

1 Content Relevance (bottom line): Persona recall determines whether the solution is truly based on a user profile and is the basis for all subsequent dimensions. High recall equals high personalization and interpretability. Pertinence checks whether the plan focuses on the "core pain points" that users need to solve most, and avoids going off-topic or answering secondary needs. Information completeness measures the depth and breadth of coverage of all user needs (primary and secondary). Only after recalling enough information and focusing on the core pain points can we talk about "complete" and "detailed".

2 Implementability Related (falling formation): Operability tests whether the suggestions can be transformed into specific actions, focusing on "landing". Accuracy & Safety guarantees that the operation advice given is in line with professional common sense and not misleading. If the operability is strong but not safe, it will cause negative consequences for the user. They need to be parallel. On the premise of ensuring operability and safety, brevity avoids excessive verbosity, improves execution efficiency, and balances the amount of information and reading load.

3 Style and Interaction (experience layer): Empathy & Tone provides users with emotional support and reduces anxiety in addition to content and operation. When the plan is accurate and clear enough, tone and empathy make users more willing to practice it. Guide & Interactivity asks questions and feedback prompts to get users involved and turn passive acceptance into active thinking. Organization & Legibility ensures a clear structure and hierarchy for quick grasp of key points.

5 Results and Analysis

5.1 Industry-level Radar Chart

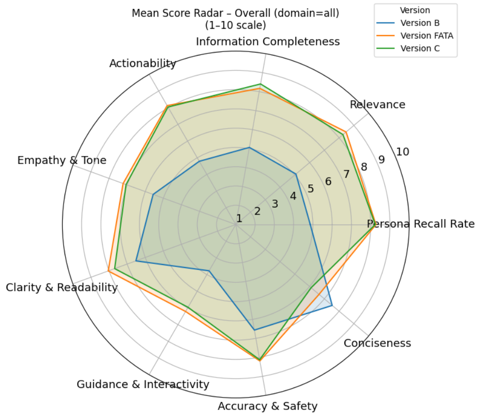


Figure 6: Industry-wide overall score comparison (radar chart).

Overall score and improvement: The FATA method had the highest score in 8 out of 12 vertical categories, leading the B-Prompt by approximately +41.7% and the C-Prompt by approximately +2.1% in overall score. It is indicated that the effects of FATA and C-Prompt are comparable. Both are approximately 40% better than B-Prompt, and the improvement effect is very significant.

5.2 Dimension-level Radar Chart

Overall response quality: Overall, the response quality of FATA and C-Prompt is close. Both bring significant overall improvements, especially in terms of pertinence, information completeness, organization, and readability. Baseline issue B, which is often overlooked by users in real environments for key information, leads only in “simplicity,” but sacrifices almost all core experience indicators. There is a gap of about 2 to 3 points between it and Type B in key dimensions. C-Prompt is slightly higher than FATA in terms of accuracy and safety, as well as persona recall rate, with little difference. FATA leads in the other seven dimensions, but both sacrifice some simplicity. When analyzing, only the final response of FATA is included in the evaluation; in fact, in terms of overall conciseness, FATA will be lower.

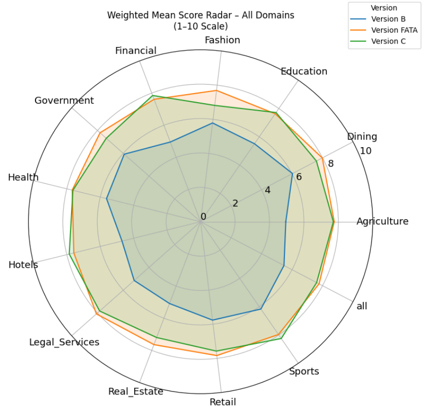


Figure 7: Detailed dimension-level comparison (radar chart).

5.3 Ranking Flow

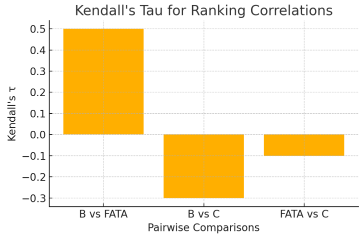


Figure 8: Kendall τ rank correlation across industries.

Industry ranking reshuffling: C-Prompt and B-Prompt are in reverse order—some categories rank lower than others. FATA reshuffled the rankings so that both ends were less concentrated. This shows that FATA improves the competitiveness of weak industries while maintaining the advantages of strong industries. This strategy aligns with the economic principle of “balanced development,” improving underperforming parts of the system while minimizing negative impact on dominant industries.

5.4 Statistical Analysis

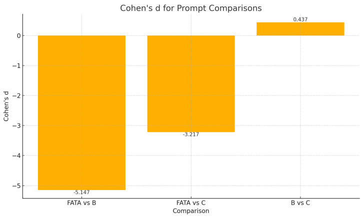


Figure 9: Cohen’s d effect size for pairwise comparisons.

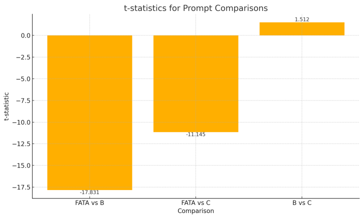


Figure 10: t-test results for pairwise score differences.

Statistical comparisons:

1. **B-Prompt vs FATA:** A t-value of -17.831 and a p-value < 0.001 indicate that the difference between these two groups is highly significant. Cohen’s d = -5.147 (greater than 0.8) indicates a very large effect size. This shows that FATA substantially outperforms B-Prompt in all dimensions, with an almost certain significant positive effect.

2. **B-Prompt vs C-Prompt:** t = -11.145, p < 0.001 indicates a significant difference. Cohen’s d = -3.217 (greater than 0.8) also indicates a large effect size. This demonstrates that C-Prompt significantly outperforms the baseline B-Prompt, leading to marked performance gains.

3. **FATA vs C-Prompt:** t = 1.512, p = 0.1586 indicates no statistical significance (p > 0.05). Cohen’s d = 0.437 (small effect) shows a small effect size. Therefore, there is little practical difference between the two methods.

5.5 Coefficient of Variation (CV) Analysis

Stability analysis: Taking the number of “stable” dimensions (CV $\leq$ 0.10) as the condition, FATA is the most stable, with the lowest mean CV (approximately $\approx$ 0.0803), indicating minimal overall score fluctuation. Nine out of ten dimensions satisfy CV $\leq$ 0.10 (90% stability). B-Prompt has a mean CV of approximately $\approx$ 0.2009, with only one stable dimension (10% stability). C-Prompt has a mean CV of approximately $\approx$ 0.1604, with no stable dimensions (0%). FATA is highly stable in almost all dimensions—especially “targeting” and “accuracy & security.” Although the total scores of FATA and C-Prompt are similar, FATA’s improved stability and consistency suggest that the additional information was effective, which likely explains its slight edge in overall score.

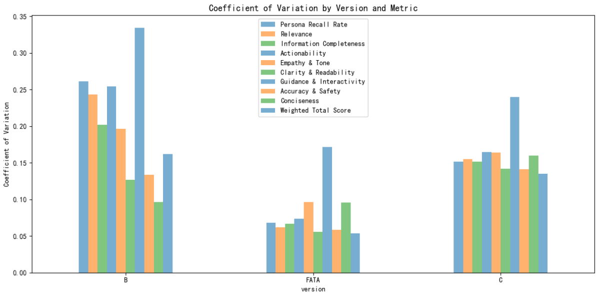


Figure 11: Standard deviation and CV comparison across methods.

6 Conclusion and Future Work

FATA fills an important gap in the existing spectrum of questioning strategies through the interaction paradigm of “first ask and supplement, then answer”. Unlike traditional supplementary-question methods, FATA not only introduces a breakthrough in interaction mode, but also significantly optimizes the questioning strategy by generating a comprehensive list of follow-up questions in a single round of dialogue. This innovation greatly enhances both interaction efficiency and answer quality. Our experimental results demonstrate that FATA outperforms the baseline method B-Prompt across multiple dimensions—particularly information completeness, relevance, and coherence—and exhibits greater stability than the context-enhanced expert prompt C-Prompt.

Looking to the future, FATA holds exciting potential for a wide range of practical applications. By empowering non-expert users to articulate their needs more clearly, it promises a higher-quality interactive experience. Moreover, as related technologies continue to advance, we are confident that integrating FATA with retrieval-augmented methods, tool invocation, and other state-of-the-art techniques will further expand the scope and capabilities of LLMs.

We envision several promising directions for further research:1. Adaptive Interaction Strategy: Develop a general framework enabling the model to choose the optimal questioning strategy based on real-time feedback (e.g., when to trigger method B). 2.Human-Centered Alignment: Incorporate human preferences and values into the interaction paradigm to ensure that the model remains beneficial, truthful, and fair throughout multi-turn conversations. 3. Richer Evaluation Metrics: Expand dialogue-quality metrics—especially those targeting the interaction process—by devising methods to quantify the contribution of each supplementary question to overall task success.

7 Additional Prompt Variants

1. Hybrid Strategy: Combine B-Promp's clarification with C-Prompt's expert keyword search chain. 2. Dual-Expert Strategy: Solicit parallel inquiries from two domain experts. 3. Simplification Strategy: Ensure concise structure and questions with guiding examples. 4. Minimalist Strategy: Pose only essential questions when needed.

8 Related Work

Selective Clarification and Questioning Strategy: Early dialogue systems add questions only after detecting ambiguity. Kuhn et al. [Kuhn and et al., 2022] proposed CLAM, which uses a two-step discrimination–generation process to issue clarification questions only when a threshold is crossed, balancing interaction cost and accuracy. Subsequent works enable models to predict future dialogue turns [Zhang et al., 2024] or directly generate follow-up questions [Tix and Binsted, 2024], reducing the number of clarification rounds. These methods improve accuracy on ambiguous queries but may still miss hard-to-detect information gaps, leading to incorrect answers.

Chain Reasoning and Self-Questioning within the Model: Chain-of-Thought (CoT) prompts allow LLMs to solve complex tasks via explicit reasoning paths [Wei et al., 2022]. Improvements such as Self-Consistency [Wang et al., 2022] and Tree of Thoughts [Yao et al., 2023] enhance robustness by sampling multiple reasoning traces or exploring tree-structured solution spaces. Parallelly, Self-Ask [Press and et al., 2022] lets the model generate and answer sub-questions internally before summarizing its conclusions. These approaches rely on model parameters or external retrieval and lack direct interaction with user background, limiting personalization.

Reasoning–Action Coupling and Tooling Enhancements: Recent frameworks enable LLMs to call external APIs or environments:

ReAct alternates "reasoning–action" steps to guide tool calls and correct hallucinations [Yao and et al., 2022].

Toolformer expands the tool ecosystem by learning when and how to insert API-call tags via self-supervised labeling [Schick and et al., 2023].

MRKL Systems introduce routers to distribute subtasks between neural and symbolic modules, improving composability and interpretability [Karpas and et al., 2022].

Planner-Executor (Plan-and-Act) first creates a high-level plan, then executes subtasks to solve long-chain workflows [Deng and et al., 2025].

Self-Refine adds iterative "self-reflection" to fill remaining information gaps [Press et al., 2023].

While effective for retrieval or computation, these methods assume complete input context and often incur extra rounds for refinement.

Retrieval Enhancement and Knowledge Externalization: Retrieval-Augmented Generation (RAG) reduces hallucinations by retrieving external knowledge, evolving from vanilla RAG to multi-index and adaptive-weighting paradigms [Gao et al., 2023]. However, RAG focuses on knowledge gaps and presumes queries are fully specified; if user questions lack key constraints, retrieved evidence may still miss true requirements.

9 Reproducibility Statement

Models: ChatGPTo4-mini-2025-04-16 for generation; ChatGPT-O3 for evaluation.

Dataset: 300 personas (12×5×5), prompt templates to be open-sourced.

Resources: All prompts, templates, and evaluation scripts will be released upon publication.

10 Limitations

While the *First Ask Then Answer* (FATA) framework demonstrates significant improvements in information completeness and interaction efficiency, there are several limitations to consider:

- **Limited Scenario Coverage:** While FATA has shown effectiveness across 12 dialogue domains, there may be additional complex scenarios or niche areas where the supplementary questions may not be fully exhaustive or aligned with expert needs. Certain highly specialized domains may require deeper context or domain-specific knowledge that FATA's current prompt generation mechanism may not fully capture.

- **User Understanding and Engagement:** The effectiveness of FATA depends on how well users respond to the supplementary questions. In cases where users fail to provide clear or accurate responses, the framework's ability to generate high-quality answers may be hindered. This is particularly relevant for non-expert users who might struggle with interpreting or fully answering the supplementary questions, affecting the final answer quality.
- **Scalability in Complex Systems:** The current design of FATA assumes that the supplementary questions can cover the necessary dimensions of a problem in a single round. However, in highly complex scenarios with numerous interrelated variables, the single-turn questioning strategy may lead to information overload or gaps in the collected data.
- **Dependence on Model Performance:** The success of FATA is heavily dependent on the underlying model's ability to generate accurate and coherent supplementary questions. In the presence of biases, model limitations, or insufficient fine-tuning, the questions generated may not always be optimal, potentially reducing the overall effectiveness of the system.
- **Ethical and Privacy Concerns:** While the framework avoids requesting sensitive data, there is still the potential for privacy concerns, especially in domains like healthcare or personal finance. Ensuring that supplementary questions are phrased in a way that avoids inadvertently collecting sensitive information remains a critical challenge.
- **Evaluation Metrics:** The evaluation process of FATA relies on automated models like ChatGPT-O3 for assessing the relevance, completeness, and accuracy of the generated answers. While this approach is efficient, it may not fully capture the nuances of human evaluation or subjective user experiences. Further human-centered evaluation is necessary to understand the broader impact of FATA on user satisfaction and trust.

References

Y. Deng and et al. Plan-and-Act: Structured Agent Reasoning. arXiv preprint arXiv:2504.04717, 2025.

Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv preprint arXiv:2312.10997, 2023.

E. Karpas and et al. MRKL Systems. arXiv preprint arXiv:2205.00445, 2022.

A. Kuhn and et al. CLAM: Clarify Language Model Ambiguity. arXiv preprint arXiv:2212.07769, 2022.

O. Press and et al. Self-Ask: Chain of Thought Prompting With Self-Generated Questions. arXiv preprint arXiv:2210.03350, 2022.

O. Press, M. Zhang, S. Min, L. Schmidt, N. A. Smith, and M. Lewis. Measuring and Narrowing the Compositionality Gap in Language Models. In *Findings of ACL: EMNLP 2023*, 2023.

T. Schick and et al. Toolformer: Language Models Can Teach Themselves to Use Tools. arXiv preprint arXiv:2302.04761, 2023.

B. Tix and K. Binsted. Better Results Through Ambiguity Resolution: Large Language Models that Ask Clarifying Questions. In D. D. Schmorrow and C. M. Fidopiastis, editors, *Augmented Cognition (Lecture Notes in Computer Science, Vol. 14695)*, pages 72–87. Springer, 2024.

X. Wang, J. Wei, D. Schuurmans, Q. Le, E. H. Chi, S. Narang, and D. Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. arXiv preprint arXiv:2203.11171, 2022.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. Le, and D. Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv preprint arXiv:2201.11903, 2022.

S. Yao and et al. ReAct: Synergizing Reasoning and Acting in Language Models. arXiv preprint arXiv:2210.03629, 2022.

S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. arXiv preprint arXiv:2305.10601, 2023.

M. J. Q. Zhang, W. B. Knox, and E. Choi. Modeling Future Conversation Turns to Teach LLMs to Ask Clarifying Questions. arXiv preprint arXiv:2410.13788, 2024.