Advances in Engineering Innovation Vol.16 Issue 11 EWA Publishing
Available Online: 27 November 2025 DOI: 10.54254/2977-3903/2025.29989

A study on a multi-agent interaction responsibility determination method based on prompt engineering

Xinxin Zhao[1], Fei Chen[1]*

[1]Beijing University of Posts and Telecommunications, Beijing, China

*Corresponding Author. Email: cflyhigh@bupt.edu.cn

Abstract. In response to the challenges of complex multi-agent responsibility division and low efficiency in manual judgment in telecommunications operators' customer complaint handling, this paper proposes a multi-agent interaction responsibility determination method based on Large Language Models (LLMs) and Prompt Engineering. Taking complaint texts as input, the method constructs a hierarchical reasoning framework consisting of an individual layer and an interaction layer. The individual layer analyzes each agent's responsibility fulfillment behavior, while the interaction layer depicts responsibility transfer relationships among agents. A responsibility fusion mechanism then integrates these analyses to generate a comprehensive responsibility distribution. This method achieves automated and interpretable multi-agent responsibility determination, offering a new technical approach and theoretical foundation for intelligent customer service, responsibility tracing, and service quality evaluation.

Keywords: prompt engineering, Large Language Models, multi-agent responsibility determination, telecommunications operator complaints

1. Introduction

In the process of handling customer complaints in the telecommunications industry, incidents often involve multiple responsible agents such as customers, customer service representatives, technical support personnel, and backend systems. The behaviors of these agents are interdependent and intricately intertwined, making the division of responsibility highly multi-causal and ambiguous. Traditional responsibility determination methods largely rely on manual expertise or rule-based text matching. Although such methods may yield acceptable results in specific scenarios, they struggle to cope with the unstructured nature, semantic ambiguity, and multi-party interactions inherent in real complaint texts. As a result, traditional approaches often lead to highly subjective and poorly scalable outcomes. In recent years, Large Language Models (LLMs) have demonstrated remarkable capabilities in semantic understanding, logical reasoning, and text generation, offering new technical paradigms for responsibility recognition and determination. However, the reasoning performance of LLMs is highly sensitive to prompt design: different task instructions may lead to significant variations in reasoning logic and judgment results. Therefore, a key challenge lies in how to employ effective Prompt Engineering to guide LLMs toward structured and multi-layered responsibility reasoning, enabling intelligent and interpretable decision-making.

To address this issue, this paper proposes a multi-agent interaction responsibility determination method tailored to telecommunications customer complaint scenarios. The method integrates the reasoning capability of LLMs with the design principles of Prompt Engineering to construct a hierarchical reasoning framework composed of an individual layer and an interaction layer. At the individual layer, the model evaluates each agent's direct responsibility across dimensions such as controllability of behavior, fulfillment of obligations, and impact on outcomes. At the interaction layer, it analyzes responsibility transmission and dependency relationships among agents, establishing a responsibility matrix to represent the inter-agent influence. Through a responsibility fusion mechanism, the model ultimately generates a comprehensive responsibility distribution, achieving automated and interpretable determination of multi-agent responsibility in complex events. This approach not only enhances the accuracy and consistency of responsibility identification but also provides theoretical and practical references for applications in intelligent customer service systems, service quality assessment, and responsibility tracing.

2. Related research

2.1. Theoretical foundations

Social Exchange Theory [1] posits that social behavior is essentially an interactional process based on the exchange of resources and obligations among individuals, with responsibility determined by the roles and duties each party fulfills during the interaction. The formation and transfer of responsibility constitute a dynamic feedback mechanism rather than a static judgment. In the context of telecommunications complaints, the behaviors of multiple parties—such as customers, customer service representatives, and technical support staff—are interdependent during the occurrence and resolution of issues. Therefore, modeling these interactive relationships is necessary to describe the mechanism of responsibility transmission. Based on this theory, this study proposes an interactional responsibility reasoning model that takes the "agent pair" (pairwise) as the basic analytical unit to reflect the generation and evolution of multi-party responsibilities.

Collaborative Reasoning Theory [2] emphasizes information sharing and multi-path reasoning mechanisms among multiple agents engaged in complex cognitive tasks. The local reasoning results formed by different agents, each based on their own roles and perspectives, must be integrated collaboratively to arrive at a coherent and rational conclusion. Drawing on this concept, this paper divides the responsibility determination process into two stages: individual-layer reasoning and interaction-layer reasoning. The former focuses on the behavioral responsibility of single agents, while the latter concentrates on interactive responsibility among agents. The results of both layers are then fused to produce a globally consistent responsibility judgment.

2.2. Related work

2.2.1. Traditional responsibility determination methods

Early studies on responsibility determination mainly adopted rule-matching and statistical analysis methods [3], such as logistic regression, analytic hierarchy process (AHP), and fuzzy comprehensive evaluation [4]. These approaches quantify each agent's responsibility through pre-defined indicator systems. Although they exhibit good interpretability, they struggle to handle the complexity of natural language contexts and multi-agent interactional information [5]. With the advancement of text mining techniques, researchers began transforming responsibility determination into classification or regression tasks [6], employing algorithms such as Support Vector Machines (SVM) and Random Forests (RF) for automated analysis. While these methods improved efficiency, their capacity to model semantic and contextual dependencies remained limited.

2.2.2. Deep learning methods

The introduction of deep learning has strengthened responsibility determination at the semantic level. Models based on BERT, BiLSTM, and Attention mechanisms can capture syntactic dependencies and semantic associations, thereby improving the accuracy of responsibility assessment. He et al. [6,7] proposed models such as the Responsibility Chain and Event Causal Graph to analyze logical relationships among agents' behaviors. However, these models typically depend on large-scale annotated datasets and show limited interpretability when dealing with complex causal and interactional structures, making it difficult to generalize effectively to multi-agent responsibility allocation scenarios.

2.2.3. Prompt engineering and large model reasoning methods

In recent years, Large Language Models (LLMs) have demonstrated powerful capabilities in language understanding and logical reasoning, becoming a key technological foundation for responsibility determination tasks. Prompt Engineering [8] has been widely adopted to guide models toward structured reasoning. Its core concept [8,9] lies in explicitly controlling the model's reasoning path through carefully designed task descriptions, instructions, or examples (few-shot examples), thereby aligning the model's reasoning logic with human attribution patterns. Research on prompt-based methods can be divided into three main categories:

(1) Task-Reconstruction Prompts [8,10,11] (Instruction-style Prompts): These employ natural language instructions to guide the model in performing responsibility attribution tasks, such as "Analyze the behaviors of the customer and the service representative in this case, and determine which party should bear the main responsibility." This approach is suitable for zero-shot or few-shot tasks.

(2) Context-Augmented Prompts [12-14] (Context-aware Prompts): These incorporate domain knowledge or behavioral norms (e.g., service procedures, response time limits) within the prompt to improve the model's accuracy and interpretability in responsibility determination.

(3) Hierarchical Reasoning Prompts [15,16] (Chain-of-Thought / Multi-step Prompts): These guide the model to reason step-by-step through causal chains—first identifying behaviors, then analyzing interactions, and finally quantifying responsibility proportions—to enhance logical consistency.

Existing studies have shown that the quality of prompt design directly influences the performance of LLMs in responsibility determination. Multi-stage or multi-perspective prompt structures can significantly improve reasoning consistency and output stability while maintaining interpretability. The current research trend integrates Prompt Engineering with cognitive theories, extending model reasoning from single-point judgments to multi-path collaborative reasoning to achieve interpretable responsibility attribution.

In summary, most existing studies remain focused on single-agent responsibility identification, with insufficient modeling of responsibility interaction and transfer among agents. Building upon this gap, the innovations of this paper are as follows:

(1) Theoretical Innovation: Integrates Social Exchange Theory and Collaborative Reasoning Theory to establish an interactive cognitive model for responsibility formation and transmission.

(2) Methodological Innovation: Proposes a multi-agent hierarchical responsibility determination framework based on Prompt Engineering, employing pairwise reasoning and responsibility matrix fusion to achieve multi-layered structured attribution.

(3) Applied Innovation: Uses telecommunications customer complaints as a case scenario to validate the framework's feasibility and interpretability in multi-agent responsibility determination tasks.

Through structured prompt design and interactional responsibility modeling, this study transitions from traditional single-agent judgment toward multi-agent collaborative reasoning, providing a new methodological pathway for responsibility cognition modeling and explainable artificial intelligence research.

3. Methodology design

3.1. Multi-agent interaction responsibility determination method

This study proposes a multi-agent interaction responsibility determination method based on Prompt Engineering and Large Language Models (LLMs), with the customer complaint scenarios of telecommunications operators as the research background. The method aims to fully leverage the natural language understanding and reasoning capabilities of LLMs through hierarchical prompt design and a responsibility fusion mechanism, thereby achieving automated and interpretable responsibility determination among multiple agents involved in an event.

The overall methodological framework comprises three core stages, as illustrated in Figure 1.

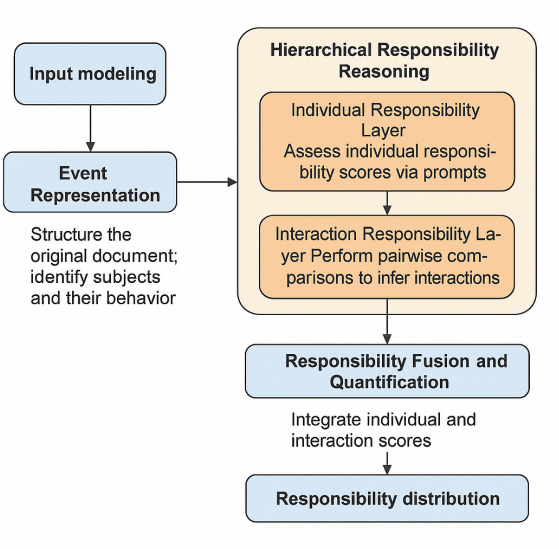

Figure 1. Flowchart of the multi-agent interaction responsibility determination process

3.1.1. Event representation

The original complaint documents are structured and modeled to identify multiple agents (e.g., customers, service representatives, companies, and systems) and their corresponding behavioral descriptions, forming a sequence of agent behaviors $S = \{s_1, s_2, \ldots, s_n\}$.

During this stage, a mapping table between events and agents is established to ensure traceability of subsequent reasoning results.

3.1.2. Hierarchical responsibility reasoning

This constitutes the core of the proposed method, employing a dual-layer reasoning mechanism combining individual-layer reasoning and interaction-layer reasoning.

(1) Individual Responsibility Layer

Guided by prompts, the large language model (LLM) evaluates each entity's individual responsibility score R_i^{ind} from three dimensions: behavioral controllability, obligation compliance, and outcome influence.

This layer focuses on single-entity attribution analysis, capturing the directness and subjective controllability of responsibility.

(2) Interaction Responsibility Layer

In multi-entity scenarios, the model performs pairwise comparisons between entities to infer relative responsibilities, generating an interaction responsibility matrix $M_{ij} \in [-1,1]$ .This matrix depicts the transmission and dependency of responsibilities among entities.

3.1.3. Responsibility fusion and quantification

The individual responsibility scores and the interaction responsibility matrix are integrated through weighted fusion to obtain each entity's overall responsibility score:

$$R_i = \alpha R_i^{ind} + \beta \sum_{j \neq i} M_{ij}$$

where α and β are weighting parameters used to balance the influence of individual and interaction responsibilities in the final results. The total responsibility is normalized to ensure that the sum equals 100.

Through this design, the model outputs both the responsibility distribution across entities and the corresponding reasoning text. It enables hierarchical analysis of multi-entity responsibilities under conditions of limited or no explicit labels, while maintaining interpretability.

3.2. Individual responsibility reasoning

The individual responsibility reasoning layer aims to analyze the behavioral consequences borne by each entity and their causal relationship with the event outcomes from a single-entity perspective.

Leveraging the semantic understanding capabilities of large language models, the reasoning process is guided by prompts across three dimensions:

(1) Behavioral Controllability – Whether the entity has the capacity to control the event outcome.

(2) Behavioral Normativity – Whether the entity's behavior aligns with assigned duties or industry standards.

(3) Outcome Influence – The degree to which the entity's behavior contributes to or affects the final result.

The model outputs a structured JSON file containing the scores for each of the three dimensions and the overall individual responsibility score $R_i^{ind} \in [0,1]$. This mechanism enables the model to semantically characterize each entity's behavioral traits and the dominance of its responsibility.

3.3. Interaction responsibility reasoning

In real-world communication complaint scenarios, the behaviors of different entities are often interdependent. For example, a service agent's response delay may stem from a customer's inaccurate description or insufficient system feedback. Such causal chain dependencies necessitate a dynamic allocation of responsibility.

Therefore, this study introduces an interaction responsibility reasoning mechanism based on the foundation of individual reasoning.

By constructing entity pairs (i,j)(i, j)(i,j), the model performs responsibility inference within a relative comparative semantic space, producing judgments on the relative degree of responsibility between two entities—such as "Entity A is primarily responsible, while Entity B is secondarily responsible," or "Both share equal responsibility."

The system maps these results to matrix elements M_{ij} , thereby forming an interaction responsibility matrix $M = [M_{ij}], M_{ij} \in [-1,1], M_{ij} = -M_{ji}$.

Here, $M_{ij} > 0$ indicates that entity i bears greater responsibility than entity j. This matrix captures the transmission paths and dependency patterns of responsibility in multi-entity systems, providing structured input for the final responsibility fusion process.

4. Experimental design and result analysis

The main objectives of this experiment are as follows:

(1) To verify the feasibility of the proposed hierarchical responsibility reasoning framework (individual layer + interaction layer) in multi-entity scenarios;

(2) To compare the effects of different prompt design strategies on the model's responsibility attribution accuracy and consistency;

(3) To explore the role of the responsibility matrix fusion mechanism in improving model stability and interpretability.

Through these experiments, the study aims to demonstrate the proposed method's innovation and effectiveness in hierarchical responsibility modeling and reasoning.

4.1. Dataset and preprocessing

The experimental dataset consists of 679 customer complaint cases collected from a telecommunications operator. Each document includes text segments such as the reason for complaint, verification process, preliminary handling, and rectification opinions, which reflect the interactive behaviors among multiple entities (e.g., customer, service representative, company, and system). Domain experts manually annotated a subset of samples to produce a gold-standard responsibility distribution, including the entity name, role type, responsibility score, and reasoning explanation. The annotated data contain the following structured information:

- Entity set: $S = \{s_1, s_2, \ldots, s_n\}$;
- Behavior segments and corresponding semantic labels for each entity;
- Expert-assigned responsibility scores and textual reasoning explanations.

4.2. Experimental design

4.2.1. Comparative models

To evaluate the effectiveness of the proposed approach, three comparative models were designed, as shown in Table 1.

Table 1. Comparative model settings

Model ID	Description	Key Features
M1	Direct question-answering prompt	Directly asks the model to output responsibility scores for each entity without structural constraints
M2	Structured prompt with explicit entity list	Specifies all entities and requires output in standardized JSON format to ensure uniformity
M3 (Proposed)	Hierarchical prompt structure: individual reasoning + interaction reasoning + fusion computation	Evaluates responsibility dimensions (controllability, obligation fulfillment, and outcome influence) at the individual layer; captures responsibility transfer relationships at the interaction layer; and fuses results to generate a comprehensive responsibility score

Note: Both M1 and M2 are single-stage prompt structures differing only in guidance and output constraints. M3, the proposed method, introduces a hierarchical reasoning mechanism that simulates the multi-level causal attribution process of human experts.

4.2.2. Experimental procedure

First, the preprocessed complaint samples were input into the Spark Lite large language model. The model generated individual responsibility scores (R_i^{ind}) for each entity based on the textual evidence. Next, an interactive prompt mechanism was employed to perform pairwise reasoning among entities, generating a responsibility matrix (M_{ij}) that characterizes their relative responsibility relationships. Then, the individual scores and interaction matrix were fused and normalized to obtain the

final comprehensive responsibility distribution. Finally, the model outputs were compared against the expert annotations, evaluating performance using Accuracy, Cohen's Kappa, F1-score, and Mean Absolute Error (MAE) metrics.

4.3. Experimental results and analysis

4.3.1. Overall performance comparison, as shown in table 2

Table 2. Overall performance comparison

Model	Accuracy	Kappa	F1-score	Explainability
M1	0.551	0.253	0.472	0.328
M2	0.684	0.446	0.581	0.546
M3	0.775	0.572	0.719	0.814

The results show that the proposed M3 model outperforms both baseline methods across all metrics. Specifically, M3 achieved an Accuracy of 0.775, representing improvements of approximately 13.3% over M2 and 22.4% over M1. The Cohen's Kappa increased from 0.446 (M2) to 0.572 (M3), indicating substantially stronger alignment between the model's results and expert judgments. The F1-score of 0.719 reflects better balance between precision and recall in multi-entity responsibility identification.

This performance improvement primarily stems from M3's hierarchical prompt design and responsibility matrix fusion mechanism. By explicitly reasoning at both the individual and interaction levels, the model can capture not only the independent responsibilities of each entity but also their causal interdependencies, thereby more accurately reflecting the complex, chain-like nature of real-world responsibility attribution.

4.3.2. Interpretability analysis

In terms of Explainability, M3 achieved a score of 0.814, significantly surpassing M2 (0.546). This indicates that the outputs of M3 are more consistent with expert reasoning logic, offering stronger semantic readability and transparency. Specifically, M3 provides both individual responsibility explanations (e.g., "failed to fulfill obligation" or "had significant influence on the outcome") and interaction reasoning justifications (e.g., "the delay in service handling was caused by the customer's inaccurate description"), presenting a structured and logical causal reasoning process. In contrast, M1—relying solely on direct Q&A prompts—often produces inconsistent or irreproducible outputs lacking traceability. While M2 improves stability through output structuring, it still cannot adequately model the interaction dynamics among entities.

4.3.3. Error and consistency analysis

A mean-square error analysis between model outputs and expert-annotated responsibility distributions revealed that M3 exhibits significantly lower variance across multi-entity cases than M1 and M2. This demonstrates that the weighted fusion of the interaction responsibility matrix effectively reduces bias from single-entity reasoning and ensures convergence across multiple inference rounds. Furthermore, the improvement in Cohen's Kappa indicates that M3 maintains high consistency and robustness across different complaint categories (e.g., network failures, billing disputes, service experience), validating the generalizability of the proposed framework.

5. Conclusion and future prospects

5.1. Research conclusions

This study addresses the complex issue of multi-agent responsibility determination in the context of customer complaints within telecommunications operations, where multiple stakeholders are often intertwined and responsibility criteria are inconsistent. A large language model (LLM)-based multi-agent interactive responsibility determination method was proposed, leveraging prompt engineering to enhance interpretability and reasoning capacity. Integrating social exchange theory and collaborative reasoning theory, this work offers both theoretical and technical innovations in responsibility modeling. The main conclusions are as follows:

(1) A multi-level responsibility reasoning framework was established. The proposed dual-layer architecture—comprising an individual layer and an interaction layer—captures both individual behavioral responsibility and inter-agent interaction

responsibility. This framework overcomes the limitations of traditional single-agent approaches and provides a structured foundation for hierarchical reasoning in complex events.

(2) An interpretable reasoning mechanism based on prompt engineering was developed. By designing prompts aligned with principles of social cognition, the model can generate reasonable responsibility inferences even under zero-shot or few-shot conditions. The individual-level prompts focus on controllability and duty fulfillment, while the interaction-level prompts emphasize behavioral dependency and feedback mechanisms. This design ensures logical consistency and interpretability in responsibility attribution.

(3) A quantitative representation of inter-agent responsibility was proposed. The introduction of a responsibility matrix M_{ij} effectively models the relative responsibility relationships among agents. Through a weighted fusion mechanism, comprehensive responsibility scores for each agent are derived, transforming qualitative judgments into quantitative analysis. This mechanism captures the dynamics of responsibility transfer and complementarity, significantly improving precision and stability in responsibility determination.

(4) The applicability and effectiveness of prompt engineering in responsibility reasoning were validated. Experimental results in the telecommunications complaint scenario demonstrate that the proposed structured prompt design substantially improves the accuracy of responsibility recognition and the agreement with expert annotations. This confirms the potential of prompt engineering in multi-agent responsibility reasoning tasks.

5.2. Limitations and future directions

Despite its promising outcomes, this study still faces several limitations:

(1) Limited automation in prompt design. The current structured prompts rely heavily on manual optimization guided by task semantics, lacking adaptive adjustment capabilities. Future research should focus on developing automated prompt generation mechanisms to enhance generalizability across tasks.

(2) Restricted experimental scope. The present experiments were confined to telecommunications complaint scenarios. Subsequent studies could extend this approach to other multi-agent domains—such as government accountability, customer service coordination, and supply chain collaboration—to evaluate its cross-domain robustness.

Future work can be further developed along the following directions:

(1) Incorporating adaptive prompt optimization mechanisms. Techniques such as Reinforcement Learning from Human Feedback (RLHF) or genetic algorithms may be employed to dynamically adjust prompt structures, improving model generalization and robustness across diverse reasoning tasks.

(2) Extending the framework to multi-domain, multi-agent applications. The proposed responsibility reasoning framework can be applied to broader fields, including government auditing, collaborative customer service, and intelligent manufacturing, to validate its domain adaptability and scalability.

(3) Constructing a responsibility cognition knowledge graph. By extracting knowledge and modeling causal associations, future research could build a structured knowledge graph of responsibility relationships to support interpretable reasoning by LLMs, thereby realizing a hybrid paradigm of explicit knowledge + implicit reasoning for more transparent and reliable responsibility determination.

References

[1] Song, X. (2021). Comparison and consistency test of corporate social responsibility measurement methods. *Finance and Accounting Monthly*, 17, 114–121. https: //doi.org/10.19641/j.cnki.42-1290/f.2021.17.016

[2] Wang, Y. (2020). Judicial practice analysis of liability determination for shared car platforms in traffic accidents [Master's thesis, Xinjiang University].

[3] Zhang, Y., Zhang, Q., Xu, Q., Wang, H., & Yu, D. (2021). Responsibility assessment of multiple harmonic sources based on statistical correlation analysis. *Applied Science and Technology*, 48(04), 47–53.

[4] Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Podstawski, M., Gianinazzi, L., Gajda, J., Lehmann, T., Niewiadomski, H., Nyczyk, P., & Hoefler, T. (2024). Graph of thoughts: Solving elaborate problems with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16), 17682–17690. https: //doi.org/10.1609/aaai.v38i16.29720

[5] Creswell, A., & Shanahan, M. (2022). Faithful reasoning using large language models (arXiv: 2208.14271). arXiv. https: //doi.org/10.48550/arXiv.2208.14271

[6] He, J., Rungta, M., Koleczek, D., Sekhon, A., Wang, F. X., & Hasan, S. (2024). Does prompt formatting have any impact on LLM performance? (arXiv: 2411.10541). arXiv. https: //doi.org/10.48550/arXiv.2411.10541

[7] Heider, F. (1982). The psychology of interpersonal relations. Psychology Press.

[8] Jiang, H., Wu, H., Wang, T., & Yan, X. (2022). A knowledge-enabled data management method towards intelligent police applications. In B. Li, L. Yue, J. Jiang, W. Chen, X. Li, G. Long, F. Fang, & H. Yu (Eds.), Advanced Data Mining and Applications, ADMA 2021, Part II (Vol. 13088, pp. 162–174). Springer International Publishing. https: //doi.org/10.1007/978-3-030-95408-6_13

[9] Robbins, S. P., & Judge, T. (2018). Organizational behavior. Pearson.

[10] Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications (arXiv: 2402.07927). arXiv. https: //doi.org/10.48550/arXiv.2402.07927

[11] Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models (arXiv: 2203.11171). arXiv. https: //doi.org/10.48550/arXiv.2203.11171

[12] Wang, Y., Gao, J., Chen, J., & IEEE. (2020). Deep learning algorithm for judicial judgment prediction based on BERT. In Proceedings of the 2020 5th International Conference on Computing, Communication and Security (ICCCS-2020). IEEE. https: //doi.org/10.1109/icccs49678.2020.9277068

[13] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), Advances in Neural Information Processing Systems, 35 (NeurIPS 2022). Neural Information Processing Systems (NeurIPS).

[14] Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models (arXiv: 2305.10601). arXiv. http: //arxiv.org/abs/2305.10601

[15] Zhang, Y., Du, L., Cao, D., Fu, Q., & Liu, Y. (2024). Prompting with divide-and-conquer program makes large language models discerning to hallucination and deception (arXiv: 2402.05359). arXiv. http: //arxiv.org/abs/2402.05359

[16] Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q., & Chi, E. (2023). Least-to-most prompting enables complex reasoning in large language models (arXiv: 2205.10625). arXiv. https: //doi.org/10.48550/arXiv.2205.10625