arXiv:2402.18439v3 [cs.CL] 19 Jun 2024

Beyond Natural Language: LLMs Leveraging Alternative Formats for Enhanced Reasoning and Communication

Weize Chen[1*], **Chenfei Yuan**[1*], **Jiarui Yuan**[1*], **Yusheng Su**[1], **Chen Qian**[1], **Cheng Yang**[3†], **Ruobing Xie**[2], **Zhiyuan Liu**[1†], **Maosong Sun**[1]
[1] Tsinghua University
[2] Tencent
[3] Beijing University of Posts and Telecommunications
{chenwz21,yuancf21,yuanjr22}@mails.tsinghua.edu.cn

Abstract

Natural language (NL) has long been the predominant format for human cognition and communication, and by extension, has been similarly pivotal in the development and application of Large Language Models (LLMs). Yet, besides NL, LLMs have seen various non-NL formats during pre-training, such as code and logical expression. NL's status as the optimal format for LLMs, particularly in single-LLM reasoning and multi-agent communication, has not been thoroughly examined. In this work, we challenge the default use of NL by exploring the utility of non-NL formats in these contexts. We show that allowing LLMs to autonomously select the most suitable format before reasoning or communicating leads to a 3.3 to 5.7% improvement in reasoning efficiency for different LLMs, and up to a 72.7% reduction in token usage in multi-agent communication, all while maintaining communicative effectiveness. Our comprehensive analysis further reveals that LLMs can devise a format from limited task instructions and that the devised format is effectively transferable across different LLMs. Intriguingly, the structured communication format decided by LLMs exhibits notable parallels with established agent communication languages, suggesting a natural evolution towards efficient, structured communication in agent communication. Our code is released at https://github.com/thunlp/AutoForm.

1 Introduction

Natural language (NL) has long been recognized as a fundamental format for human thought expression and communication, underscored by its pivotal role in the cognitive processes and information exchange of humans (Chomsky, 2006; Lakoff, 2008; Whorf, 2012). However, the human mind's capabilities often extend beyond the scope of NL, as

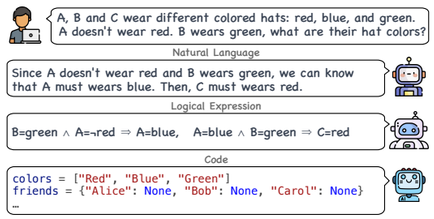


Figure 1: LLMs may leverage non-NL thought format.

suggested by the concept of *mentalese*, a mental language posited by linguists (Fodor, 1975; Pinker, 2003). Recent advancements in LLMs (OpenAI, 2023b; Google et al., 2023; Anthropic, 2023) have been remarkable, leading to their increasingly sophisticated application in language agents (Yao et al., 2023b; Park et al., 2023; Significant Gravitas). These advancements, while impressive, predominantly utilize NL for both single-LLM reasoning via Chain-of-Thought (CoT) (Wei et al., 2022; Kojima et al., 2022) and multi-agent communication (Wu et al., 2023; Park et al., 2023). Given the human mind's proficiency in transcending NL, critical inquiries emerge: Is NL the optimal format for LLMs in reasoning and inter-agent communication? If not, how should we determine the most suitable format for these applications (Fig. 1)?

Recent research challenges the notion that NL is the ideal intermediate format for LLM reasoning and multi-agent communication. Emerging variants of CoT, such as Program-of-Thought (Chen et al., 2022; Gao et al., 2023) and X-of-Thought (Liu et al., 2023) have explored the use of alternative formats like code and mathematical equations, expanding the LLMs' reasoning capabilities. However, these approaches often integrate external tools, where the alternative formats primarily serve as a means to facilitate tool execution (e.g., prompting LLM to generate code and use code interpreter execution result as the answer). This introduces complexity in discerning whether the performance improvements are attributable to the

* Equal Contribution
† Corresponding author.

arXiv:2402.18439v3 [cs.CL] 19 Jun 2024

超越自然语言：大语言模型利用替代格式增强推理与通信

陈伟泽[1*], **袁晨飞**[1*],**袁佳睿**[1*],**苏宇声**[1],**钱陈**[1],**杨程**[3†], **谢若冰**[2], **刘知远**[1†], **孙茂松**[1] **清华大学**[2] **腾讯**[3] **北京邮电大学**
{chenwz21,yuancf21,yuanjr22}@mails.tsinghua.edu.cn

摘要

自然语言（NL）长期以来一直是人类认知和通信的主要格式，由此，它在大型语言模型（LLMs）的开发和应用中也同样举足轻重。然而，除了自然语言，大型语言模型在预训练阶段还接触了多种非自然语言格式，例如代码和逻辑表达式。自然语言作为大型语言模型的最优格式，特别是在单大语言模型推理和多智能体通信中的应用，尚未得到充分检验。在这项工作中，我们通过探索非自然语言格式在这些上下文中的效用，挑战了默认使用自然语言的做法。我们证明，允许大型语言模型在推理或通信之前自主选择最合适的格式，可为不同的大语言模型带来 3.3% 到 5.7% 的推理效率提升，并在多智能体通信中实现多达 72.7% 的令牌使用量减少，同时保持通信的有效性。我们的综合分析进一步揭示，大型语言模型能够根据有限的任务指令设计格式，且所设计的格式在不同的语言模型之间可以有效迁移。有趣的是，由大型语言模型决定的结构化通信格式与既有的智能体通信语言表现出显著的相似性，这表明在智能体通信中正在向高效、结构化通信自然演进。我们的代码已发布于 https://github.com/thunlp/AutoForm。

1 引言

自然语言（NL）长期以来被视为人类思维表达和通信的基本格式，这在其在人类认知过程和信息交换中的关键作用中得到了体现（Chomsky, 2006；Lakoff, 2008；Whorf, 2012）。然而，人类心智的能力往往超越了自然语言的范畴，因为

* 同等贡献
† 通讯作者。

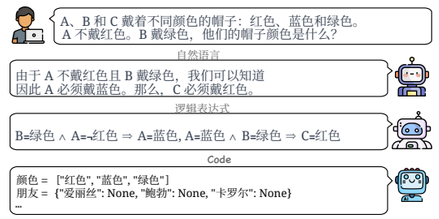


图 1：大语言模型可能利用非自然语言思维格式。

由语言学家提出的心智语（mentalese）概念所启发（Fodor, 1975；Pinker, 2003）。大语言模型（LLM）的近期进展（OpenAI, 2023b；Google et al., 2023；Anthropic, 2023）令人瞩目，促使其在语言智能体中得到日益精细的应用（Yao et al., 2023b；Park et al., 2023；Significant Gravitas）。尽管这些进展令人印象深刻，但它们主要利用自然语言（NL）进行单一大语言模型通过思维链（CoT）的推理（Wei et al., 2022；Kojima et al., 2022）以及多智能体通信（Wu et al., 2023；Park et al., 2023）。鉴于人类心智在超越自然语言方面的熟练能力，关键问题浮现：自然语言是否是大语言模型在推理和智能体间通信中的最优格式？如果不是，我们应如何确定适用于这些应用的最合适格式（图1）？

最近的研究对自然语言是理想中间格式的观点提出了挑战，该格式被认为适用于大语言模型推理和多智能体通信。新兴的思维链变体，如程序思维（陈等人，2022年；高等人，2023年）和X思维（刘等人，2023年），已探索使用代码和数学方程等替代格式，从而扩展了大语言模型的推理能力。然而，这些方法通常集成外部工具，在此过程中，替代格式主要作为促进工具执行的媒介（例如，提示大语言模型生成代码，并将代码解释器的执行结果作为答案）。这增加了区分性能提升是否归因于的复杂性

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

format itself or the accompanying tools. Additionally, while the natural ambiguities and emotions inherent in NL may align well with the nuances of human communication, these may not be desired in agent communication, where precision is more important. Nonetheless, current multi-agent research predominantly utilizes NL (Li et al., 2023a; Wu et al., 2023; Chen et al., 2023), with limited exploration of other potentially more accurate and efficient communication formats.

In this study, we implement a straightforward and effective mechanism that prompts the model to favor non-NL formats for single-LLM reasoning and multi-agent communication tasks. By adding an instruction to the original CoT prompt that directs LLMs to explore a non-NL format appropriate for the current input, we showcase the LLMs' capacity for autonomous format decisions. We observe that the LLMs can leverage many non-NL formats such as ordered lists, logical expressions, and markdown tables to reason better. Also, we observe that agents can use more structured language as their communication language to enhance the efficiency of multi-agent collaboration. In particular, our analyses across various single-LLM reasoning tasks demonstrate an average improvement in performance by 3.3-5.7%. For multi-agent communication, we observe a reduction in token usage by up to 72.7% without sacrificing effectiveness. These results highlight the considerable potential of non-NL formats in amplifying the reasoning capabilities and communicative efficiency of LLMs.

Our investigation further extends to a comprehensive analysis revealing that LLMs can devise a suitable format from a set of task-specific examples. Using the fixed devised format for the whole task also leads to better answers. Moreover, we show that the formats devised by one LLM are transferable to another LLM. And for the multi-agent communication format, we find that the format adopted by LLMs mirrors those of traditional Agent Communication Languages (ACLs) like KQML (Finin et al., 1994), highlighting their clarity, brevity, and structured format for efficient exchanges. Empirically, our approach significantly reduces token usage compared to both ACLs and NL without sacrificing performance. Our work underscores the efficacy of non-NL formats in advancing LLM reasoning and communication.

2 Related Work

LLM Reasoning. LLMs have exhibited impressive reasoning performance, especially when employing Chain-of-Thought (CoT) technique (Wei et al., 2022; Kojima et al., 2022). CoT requires LLMs to articulate their reasoning process step-by-step before arriving at a final answer. Building on the CoT framework, variants have been proposed. Program-of-Thought (PoT) (Chen et al., 2022; Gao et al., 2023) prompts models to generate code as thought, and offloads the answer generation to a code interpreter. X-of-Thought (Liu et al., 2023) integrates CoT, PoT and Equation-of-Thought, dynamically ensembling these methods for improved reasoning. Tree-of-Thought (Yao et al., 2023a) employs depth and breadth-first search techniques to produce high-quality reasoning chains. While some CoT variants explore formats beyond NL for reasoning, the chosen formats' improvements are obscured by the concurrent use of supplementary tools such as code interpreters, blurring the distinction between format efficacy and tool execution. We focus on the format itself, investigating whether alternative formats to NL improve the CoT performance.

Multi-Agent Problem Solving. Advances in Large Language Models (LLMs) have led to the development of autonomous agents like Auto-GPT (Significant Gravitas) and OpenAI Assistant (OpenAI, 2023a), demonstrating success in diverse tasks (Shinn et al., 2023; Mialon et al., 2023; Zhou et al., 2023; Boiko et al., 2023). Recent research extends this to multi-agent systems for collaborative problem-solving (Du et al., 2023; Osika, 2023; Hong et al., 2023; Qian et al., 2023). CAMEL (Li et al., 2023a) explores collaborative problem-solving between two agents through role-playing. ChatEval (Chan et al., 2023) and PRD (Li et al., 2023b) assess model responses using multi-agent debates. AgentVerse (Chen et al., 2023) introduces a comprehensive framework for multi-agent collaboration, highlighting emergent inter-agent behaviors. However, the alternative formats of multi-agent communication remains underexplored and NL is directly adopted across various research. Pham et al. (2023) explores agent communication with hidden states, but is limited to agents with the same open-source LLM. Our work analyze communication formats for both homogeneous and heterogeneous LLMs.

格式本身或配套工具。此外，尽管自然语言中固有的歧义和情感可能与人类通信的细微差别相契合，但在智能体通信中，精确性更为重要，这些特性可能并不被期望。然而，目前的多智能体研究主要使用自然语言（Li 等，2023a；Wu 等，2023；Chen 等，2023），对其他可能更准确、更高效的通信格式的探索有限。

在本研究中，我们实现了一种简洁而有效的机制，用于提示大语言模型在单大语言模型推理和多智能体通信任务中优先采用非自然语言格式。通过在原始思维链提示中添加一条指令，引导大语言模型探索适合当前输入的非自然语言格式，我们展示了大语言模型自主进行格式决策的能力。我们观察到，大语言模型能够利用许多非自然语言格式（如有序列表、逻辑表达式和 Markdown 表格）来进行更好的推理。此外，我们还观察到，智能体可以使用更结构化的语言作为其通信语言，以提高多智能体协作的效率。特别是，我们在各种单大语言模型推理任务上的分析表明，性能平均提升了 3.3% 至 5.7%。对于多智能体通信，我们观察到令牌使用量最高减少 72.7%，且不影响有效性。这些结果突显了非自然语言格式在增强大语言模型推理能力和通信效率方面的巨大潜力。

我们的研究进一步扩展到一项全面分析，表明大语言模型能够根据一组特定任务的示例设计出合适的格式。在整个任务中使用固定的设计格式也能带来更好的答案。此外，我们证明了一个大语言模型设计的格式可以迁移到另一个大语言模型。对于多智能体通信格式，我们发现大语言模型所采用的格式类似于 KQML（Finin等，1994年）等传统智能体通信语言（ACLs），突出了它们在高效交换中的清晰度、简洁性和结构化格式。实证上，与智能体通信语言（ACLs）和自然语言相比，我们的方法显著减少了令牌使用，同时并未牺牲推理性能。我们的工作强调了非自然语言格式在推进大语言模型推理和通信方面的有效性。

2 相关工作

大语言模型推理。大语言模型在推理性能方面表现令人印象深刻，尤其是在采用思维链 (CoT) 技术时（Wei 等, 2022; Kojima等, 2022年）。思维链要求大语言模型在得出最终答案之前逐步阐述其推理过程。在思维链框架的基础上，人们提出了各种变体。程序思维 (PoT) （陈等人，2022 年；高等人，2023 年）提示模型生成代码作为思维，并将答案生成卸载到代码解释器。X思维（刘等人，2023年）整合了思维链、程序思维和方程思维，动态集成这些方法以改善推理。思维树（姚等，2023a）采用深度优先和广度优先搜索技术来生成高质量的推理链。虽然一些思维链变体探索了自然语言以外的格式用于推理，但所选择格式带来的改进被同时使用的补充工具（如代码解释器）所掩盖，模糊了格式有效性和工具执行之间的区别。我们关注格式本身，研究非自然语言的替代格式是否提高了思维链的性能。

多智能体问题解决。大语言模型（LLMs）的进步推动了自主智能体的发展，例如Auto-GPT（Significant Gravitas）和OpenAI助手（OpenAI，2023a），并在多样化任务中展现了成功（Shinn等，2023；Mialon等，2023；Zhou等，2023；Boiko等，2023）。近期研究将此扩展至用于协作问题解决的多智能体系统（Du等，2023；Osika，2023；Hong等，2023；Qian等，2023）。CAMEL（Li等，2023a）通过角色扮演探索了两个智能体之间的协作问题解决。Chateval（Chan等，2023）和PRD（Li等，2023b）利用多智能体辩论来评估模型响应。AgentVerse（Chen等，2023）提出了一个多智能体协作的综合框架，强调了智能体间涌现行为。然而，多智能体通信的替代格式仍未得到充分探索，自然语言（NL）直接应用于各类研究中。Pham等（2023）探索了使用隐藏状态进行智能体通信，但仅限于使用相同开源大语言模型（open-source LLM）的智能体。我们的工作在分析通信格式时，考虑了同质和异质大语言模型（LLMs）。

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

3 Method

3.1 Problem Formulation

Consider an LLM parameterized by θ, denoted as p_θ. In response to a task description $\boldsymbol{x}$ and a prompt $\boldsymbol{p}$, CoT prompting initially guides the model to generate thought $\boldsymbol{t} = \{t_i\}$ utilizing a thought format m_t. While this format is often unspecified and defaults to natural language, alternative formats are feasible. The LLM then formulates an answer $\boldsymbol{y} = \{y_j\}$ based on $\boldsymbol{t}$. This process is mathematically expressed as sampling from the conditional probability distribution:

$$\begin{aligned} p_\theta(\boldsymbol{y}, \boldsymbol{t}|\boldsymbol{p}, \boldsymbol{x}, m_t) &= p_\theta(\boldsymbol{y}|\boldsymbol{t}, \boldsymbol{p}, \boldsymbol{x}, m_t) p_\theta(\boldsymbol{t}|\boldsymbol{p}, \boldsymbol{x}, m_t), \\ p_\theta(\boldsymbol{t}|\boldsymbol{p}, \boldsymbol{x}, m_t) &= \prod_i p_\theta\left(t_i|\boldsymbol{t}_{<i}, \boldsymbol{p}, \boldsymbol{x}, m_t\right), \\ p_\theta(\boldsymbol{y}|\boldsymbol{t}, \boldsymbol{p}, \boldsymbol{x}, m_t) &= \prod_j p_\theta\left(y_i|\boldsymbol{y}_{<i}, \boldsymbol{t}, \boldsymbol{p}, \boldsymbol{x}, m_t\right). \end{aligned} \tag{1}$$

In multi-agent scenarios, we extend this formulation to encompass communication among multiple LLMs, each characterized by θ_k. We consider a set of LLMs $\{p_{\theta_k}\}$ collaborating on a task. Communication among these agents utilizes format m_c, which can be NL or other alternative formats. This communication is formulated as:

$$\begin{aligned} &p_{\theta_k}(\boldsymbol{y}_k, \boldsymbol{t}|\mathcal{Y}, \boldsymbol{p}, \boldsymbol{x}, m_c, m_t) = \\ &\quad p_{\theta_k}(\boldsymbol{y}_k|\boldsymbol{t}, \mathcal{Y}, \boldsymbol{p}, \boldsymbol{x}, m_c) \cdot p_{\theta_k}(\boldsymbol{t}|\mathcal{Y}, \boldsymbol{p}, \boldsymbol{x}, m_t), \end{aligned} \tag{2}$$

here we slightly abuse the notation and use $\boldsymbol{y}_k$ to denote the response generated by agent k, and $\mathcal{Y}$ denote the communication history.

3.2 Format Choosing for LLMs

Building upon the framework delineated in Section 3.1, our work investigates the effectiveness of allowing the LLMs to decide the thought and communication format before actually starting reasoning or communicating. At the heart of our method is the exploration of alternative formats beyond natural language. We hypothesize that various formats, such as structured data formats (e.g., JSON, markdown tables, lists) or symbolic representations (e.g., logical expressions, mathematical equations), can potentially yield more precise and effective reasoning and streamline communication.

We employ a simple yet effective prompting mechanism, where the LLMs are prompted to select and utilize the format most conducive to the task at hand, which we term as **AutoForm** (Autonomously-Decided Format). The overview of AutoForm is illustrated at Fig. 2. Specifically, for single-LLM reasoning, we add an instruction encouraging the use of non-NL formats to the original CoT prompt. In multi-agent scenarios, a similar instruction for format decision is also added. The specifics of these prompts are detailed in Appendix B. In this way, the LLMs implicitly determine and use the thought format $m_t^* = p_\theta(\boldsymbol{x}, \boldsymbol{p}_t)$ for single-LLM reasoning and the communication format $m_c^* = p_\theta(\boldsymbol{x}, \boldsymbol{p}_c)$ for multi-agent communication, where $\boldsymbol{p}_t$ and $\boldsymbol{p}_c$ include instructions for format decision.

4 Experiments

4.1 Experimental Settings

Single-LLM Reasoning. In a preliminary experiment, we prompt various LLMs to use formats other than NL for solving reasoning problems. The results, shown in Table 6, reveal significant variability in performance across different formats. For instance, GPT-3.5 exhibited a 66.8% performance gap between using an ordered list and a multi-level list, highlighting the intrinsic suitability of specific formats for different tasks.

However, in practical applications, selecting optimal formats for each task may be impractical. To address this, we conduct a comprehensively evaluate the impact of automatically chosen thought formats on LLMs' reasoning performance. We select reasoning benchmarks covering different types of reasoning, including logical reasoning (Logic Grid Puzzle (Srivastava et al., 2022) and Information Essentialy (Srivastava et al., 2022)), mathematical reasoning (AQuA (Ling et al., 2017)), causal reasoning (Minute Mysteries QA (Srivastava et al., 2022)) and symbolic reasoning (Coin Flip (Wei et al., 2022)). For all these tasks, we require the LLMs to generate the answer in a particular format, and we extract the answer with a written regular expression. The average accuracy and the standard deviation of each dataset over 3 runs are reported on most of the datasets. For more details on the experimental settings, please refer to Appendix A.

Multi-Agent Communication. To measure whether alternative formats can streamline communication, we consider scenarios where two agents with different knowledge or contexts are tasked with answering a question. The answer to the question should be derived from the knowledge of only one of the agents, or both agents' knowledge collectively, therefore requiring information exchange. The two agents speak in turn to discuss and give

3 方法

3.1 问题陈述

考虑一个由 θ参数化、表示为p_θ的大语言模型。针对任务描述 $\boldsymbol{x}$ 和提示词$\boldsymbol{p}$，思维链提示首先引导模型利用思维格式m_t生成思维 $\boldsymbol{t} = \{t_i\}$ 。虽然该格式通常未明确指定且默认为自然语言，但也可能采用替代格式。随后，大语言模型基于 $\boldsymbol{t}$生成答案$\boldsymbol{y} = \{y_j\}$ 。该过程在数学上表示为从条件概率分布中采样：

$$\begin{aligned} p_\theta(\boldsymbol{y}, \boldsymbol{t}|\boldsymbol{p}, \boldsymbol{x}, m_t) &= p_\theta(\boldsymbol{y}|\boldsymbol{t}, \boldsymbol{p}, \boldsymbol{x}, m_t) p_\theta(\boldsymbol{t}|\boldsymbol{p}, \boldsymbol{x}, m_t), \\ p_\theta(\boldsymbol{t}|\boldsymbol{p}, \boldsymbol{x}, m_t) &= \prod_i p_\theta\left(t_i|\boldsymbol{t}_{<i}, \boldsymbol{p}, \boldsymbol{x}, m_t\right), \\ p_\theta(\boldsymbol{y}|\boldsymbol{t}, \boldsymbol{p}, \boldsymbol{x}, m_t) &= \prod_j p_\theta\left(y_i|\boldsymbol{y}_{<i}, \boldsymbol{t}, \boldsymbol{p}, \boldsymbol{x}, m_t\right). \end{aligned} \tag{1}$$

在多智能体场景中，我们将此公式扩展以涵盖多个大语言模型之间的通信，每个模型由 θ_k表征。我们考虑一组协作执行任务的大语言模型 $\{p_{\theta_k}\}$ 。这些智能体之间的通信采用格式 m_c，可以是自然语言或其他替代格式。该通信表述为：

$$\begin{aligned} &p_{\theta_k}(\boldsymbol{y}_k, \boldsymbol{t}|\mathcal{Y}, \boldsymbol{p}, \boldsymbol{x}, m_c, m_t) = \\ &\quad p_{\theta_k}(\boldsymbol{y}_k|\boldsymbol{t}, \mathcal{Y}, \boldsymbol{p}, \boldsymbol{x}, m_c) \cdot p_{\theta_k}(\boldsymbol{t}|\mathcal{Y}, \boldsymbol{p}, \boldsymbol{x}, m_t), \end{aligned} \tag{2}$$

这里我们略微滥用符号，用 $\boldsymbol{y}_k$ 表示智能体 k 生成的响应，用 $\mathcal{Y}$表示通信历史。

3.2 为大语言模型选择格式

基于第 3.1 节所述的框架，我们的工作研究了允许大语言模型在实际开始推理或通信之前决定思维和通信格式的有效性。我们方法的核心在于探索自然语言之外的替代格式。我们假设各种格式，例如结构化数据格式（如 JSON、markdown 表格、列表）或符号表示（如逻辑表达式、数学方程），可能带来更精确和有效的推理，并简化通信。

我们采用了一种简单而有效的提示机制，其中大语言模型被提示选择并利用最有利于当前任务的格式，我们将此称为AutoForm（自主决定格式）。AutoForm的概述如图2所示。具体而言，

对于单大语言模型推理，我们在原始思维链提示中增加了一条鼓励使用非自然语言格式的指令。在多智能体场景中，也添加了一条类似的关于格式决策的指令。这些提示词的详细信息见附录B。通过这种方式，大语言模型隐式地确定并使用用于单大语言模型推理的思维格式 $m_t^* = p_\theta(\boldsymbol{x}, \boldsymbol{p}_t)$ 以及用于多智能体通信的通信格式 $m_c^* = p_\theta(\boldsymbol{x}, \boldsymbol{p}_c)$，其中 $\boldsymbol{p}_t$ 和 $\boldsymbol{p}_c$ 包含格式决策的指令。

4 实验

4.1 实验设置

单大语言模型推理。在初步实验中，我们提示各大语言模型使用自然语言以外的格式来解决推理问题。结果如表6所示，揭示了不同格式间性能存在显著差异。例如，GPT-3.5在使用有序列表与多级列表时，性能差距达 66.8%，凸显了特定格式对不同任务的内在适用性。

然而，在实际应用中，为每项任务选择最佳格式可能并不切实际。为此，我们全面评估自动选择的思维格式对大语言模型推理性能的影响。我们选取了涵盖不同推理类型的推理基准，包括逻辑推理（逻辑网格谜题 (Srivastava 等, 2022) 与信息本质 (Srivastava 等, 2022))、数学推理（AQuA (Ling 等, 2017)）、因果推理（Minute Mysteries 问答 (Srivastava 等, 2022)) 和符号推理（抛硬币 (Wei 等, 2022))。对于所有这些任务，我们要求大语言模型以特定格式生成答案，并使用预设的正则表达式提取答案。我们报告了大多数数据集在 3 次运行中的平均准确率和标准差。有关实验设置的更多细节，请参阅附录 A。

多智能体通信。为衡量替代格式能否精简通信，我们考虑如下情景：两个拥有不同知识或上文的智能体被指派回答一个问题。问题的答案应仅由其中一个智能体的知识推导得出，或由两个智能体的知识共同推导得出，因此需要信息交换。两个智能体依次发言进行讨论并给出

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

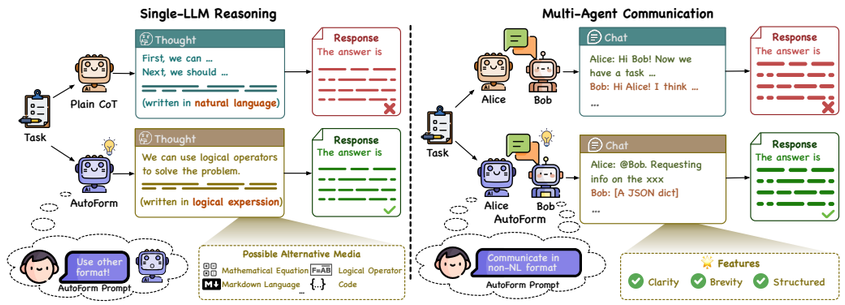


Figure 2: Overview of single-LLM reasoning and multi-agent communication using plain CoT versus the CoT with AutoForm. The left side depicts the shift from natural language to alternative formats in single-LLM reasoning, while the right side illustrates the enhanced efficiency in multi-agent communication.

Model	Logic Grid	Coin Flip	Info Essen	MM QA	AQuA	Average
GPT-3.5 CoT	$46.7_{\pm 1.6}$	$23.1_{\pm 1.0}$	$32.3_{\pm 3.2}$	$24.9_{\pm 0.8}$	$60.9_{\pm 1.2}$	$41.1_{\pm 1.8}$
+*AutoForm*	$\mathbf{48.0}_{\pm 3.9}$	$\mathbf{39.4}_{\pm 1.1}$	$\mathbf{36.7}_{\pm 3.2}$	$\mathbf{26.8}_{\pm 0.6}$	$\mathbf{63.7}_{\pm 0.7}$	$\mathbf{46.0}_{\pm 2.3}$
Gemini Pro CoT	$49.7_{\pm 0.2}$	$47.5_{\pm 0.2}$	$34.3_{\pm 0.7}$	$28.1_{\pm 0.7}$	$56.3_{\pm 0.6}$	$43.2_{\pm 0.5}$
+*AutoForm*	$\mathbf{51.2}_{\pm 0.8}$	$\mathbf{57.6}_{\pm 0.7}$	$\mathbf{39.2}_{\pm 1.8}$	$\mathbf{31.3}_{\pm 1.1}$	$\mathbf{60.0}_{\pm 0.4}$	$\mathbf{47.9}_{\pm 1.1}$
GPT-4 CoT	$61.8_{\pm 1.6}$	$93.4_{\pm 1.0}$	$78.4_{\pm 2.5}$	$38.4_{\pm 1.1}$	$79.1_{\pm 0.3}$	$71.8_{\pm 1.5}$
+*AutoForm*	$\mathbf{65.8}_{\pm 2.2}$	$\mathbf{98.4}_{\pm 0.2}$	$\mathbf{76.9}_{\pm 2.5}$	$\mathbf{41.7}_{\pm 0.9}$	$\mathbf{80.4}_{\pm 0.8}$	$\mathbf{74.1}_{\pm 1.6}$

Table 1: Comparative performance of single LLM reasoning across various datasets. "Information Essentiality" dataset is abbreviated as "Info Essen," and "Minute Mysteries QA" is referred to as "MM QA" for conciseness.

the final answer. To create such scenarios, we utilize three existing datasets: Hotpot QA (Yang et al., 2018), Wiki Hop (Welbl et al., 2018) and Narrative QA (Kociský et al., 2018). Hotpot QA and Wiki Hop are two multi-hop QA datasets, which require multiple sentences or paragraphs to deduce the final answer. We randomly assign the text segments provided in the datasets to two different agents. Communication is thus needed to derive the correct answer. We have also explored assigning each agent part of the supporting facts, thus needing more communication to derive the answer. Narrative QA requires the model to read the whole book and answer a question. The length of a book often exceeds the context limit of the LLMs. We split the books into nearly equal sizes for the two agents, and ask them to answer the question. For evaluation, we use RougeL as the primary metric. More details are elaborated in Appendix A.

To measure whether alternative formats can streamline communication, we consider scenarios where two agents with different knowledge or contexts are tasked with answering a question. The answer to the question should be derived from the knowledge of only one of the agents, or both agents' knowledge collectively, therefore requiring information exchange. The two agents speak in turn to discuss and give the final answer. To create such scenarios, we utilize three existing datasets: Hotpot QA (Yang et al., 2018), Wiki Hop (Welbl et al., 2018), and Narrative QA (Kociský et al., 2018). Hotpot QA and Wiki Hop are two multi-hop QA datasets, which require multiple sentences or paragraphs to deduce the final answer. We randomly assign the text segments provided in the datasets to two different agents. Communication is thus needed to derive the correct answer. We also explore assigning each agent part of the supporting facts, thus needing more communication to derive the answer. Narrative QA requires the model to read the whole book and answer a question. The length of a book often exceeds the context limit of the LLMs. We split the books into nearly equal sizes for the two agents and ask them to answer the question. While supporting facts in Narrative QA are not guaranteed to reside in different segments, this division still introduces critical challenges: agents must identify who possesses the necessary information, necessitating communication. For evaluation, we use RougeL as the primary metric. More details are elaborated in Appendix A.

Research Questions. To comprehensively explore the potential of LLMs in selecting suitable formats for both single-agent reasoning and multi-agent communication, we conduct an in-depth analysis addressing six pivotal research questions

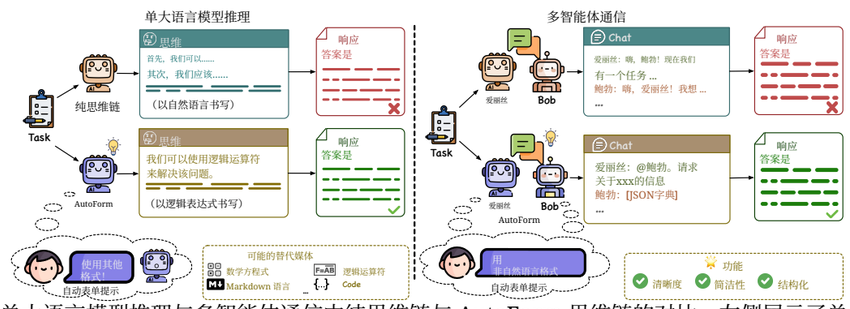


图 2：单大语言模型推理与多智能体通信中纯思维链与 AutoForm 思维链的对比。左侧展示了单大语言模型推理中从自然语言到替代格式的转变，右侧则展示了多智能体通信中效率的提升。

模型	逻辑网格	抛硬币	Info Essen	MM QA	AQuA	平均值
GPT-3.5 思维链	$46.7_{\pm 1.6}$	$23.1_{\pm 1.0}$	$32.3_{\pm 3.2}$	$24.9_{\pm 0.8}$	$60.9_{\pm 1.2}$	$41.1_{\pm 1.8}$
+*AutoForm*	$\mathbf{48.0}_{\pm 3.9}$	$\mathbf{39.4}_{\pm 1.1}$	$\mathbf{36.7}_{\pm 3.2}$	$\mathbf{26.8}_{\pm 0.6}$	$\mathbf{63.7}_{\pm 0.7}$	$\mathbf{46.0}_{\pm 2.3}$
Gemini Pro 思维链	$49.7_{\pm 0.2}$	$47.5_{\pm 0.2}$	$34.3_{\pm 0.7}$	$28.1_{\pm 0.7}$	$56.3_{\pm 0.6}$	$43.2_{\pm 0.5}$
+*AutoForm*	$\mathbf{51.2}_{\pm 0.8}$	$\mathbf{57.6}_{\pm 0.7}$	$\mathbf{39.2}_{\pm 1.8}$	$\mathbf{31.3}_{\pm 1.1}$	$\mathbf{60.0}_{\pm 0.4}$	$\mathbf{47.9}_{\pm 1.1}$
GPT-4 思维链	$61.8_{\pm 1.6}$	$93.4_{\pm 1.0}$	$78.4_{\pm 2.5}$	$38.4_{\pm 1.1}$	$79.1_{\pm 0.3}$	$71.8_{\pm 1.5}$
+*AutoForm*	$\mathbf{65.8}_{\pm 2.2}$	$\mathbf{98.4}_{\pm 0.2}$	$\mathbf{76.9}_{\pm 2.5}$	$\mathbf{41.7}_{\pm 0.9}$	$\mathbf{80.4}_{\pm 0.8}$	$\mathbf{74.1}_{\pm 1.6}$

表 1：单大语言模型推理在各个数据集上的比较性能。“信息必要性”数据集缩写为“InfoEssen”，“Minute Mysteries 问答”简称为“MM QA”，以求简洁。

最终答案。为了构建此类情景，我们利用三个现有数据集：Hotpot QA（Yang等，2018）、Wiki Hop（Welbl等，2018）和叙事问答（Kociský等，2018）。Hotpot QA和Wiki Hop是两个多跳问答数据集，需要多个句子或段落来推导最终答案。我们将数据集中提供的文本片段随机分配给两个不同的智能体，因此需要通信才能得出正确答案。我们还探索了为每个智能体分配部分支持事实，从而需要更多的通信来得出答案。叙事问答要求模型阅读整本书并回答一个问题。书籍的长度通常超过大语言模型的上下文限制。我们将书籍划分为大小大致相等的两部分，分给两个智能体，并要求它们回答问题。在评估方面，我们使用RougeL作为主要指标。更多细节在附录 A中阐述。

为了衡量替代格式能否简化通信，我们考虑两个具有不同知识或背景的智能体被要求回答一个问题的场景。问题的答案应仅从一个智能体的知识中推导出来，或从两个智能体的共同知识中推导出来，因此需要进行信息交换。两个智能体轮流发言进行讨论并给出最终答案。为了创建此类场景，我们使用三个现有数据集：Hotpot QA（Yang 等，2018）、Wiki Hop（Welbl 等，2018）和叙事问答（Kociský 等，2018）。Hotpot QA 和 Wiki Hop 是两个多跳问答数据集，需要多个句子或段落来推导最终答案。我们将数据集中提供的文本片段随机分配给两个不同的智能体。因此，需要通信来得出正确答案。我们还探索将部分支持事实分配给每个智能体，从而需要更多通信来得出答案。叙事问答要求模型阅读整本书并回答问题。书籍的长度通常超过大语言模型的上下文限制。我们将书籍分为大致相等的两部分，分别交给两个智能体，并要求他们回答问题。虽然叙事问答中的支持事实不一定位于不同的片段中，但这种划分仍然引入了关键挑战：智能体必须识别谁拥有必要的信息，从而需要通信。对于评估，我们使用 RougeL 作为主要指标。更多细节在附录 A 中详述。

研究问题。为了全面探索大语言模型在单智能体推理和多智能体通信中选择合适格式方面的潜力，我们进行了一项深入分析，以解决六个关键的研究问题

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

Model	Wiki Hop			Hotpot QA			Narrative QA		
	RougeL	# Tokens	ΔTokens	RougeL	# Tokens	ΔTokens	RougeL	# Tokens	ΔTokens
GPT-4 + GPT-3.5	**0.53**	281.5	-	0.63	345.5	-	0.43	178.3	-
+*AutoForm*	**0.53**	**255.0**	-9.4%	**0.70**	**94.3**	-72.7%	**0.48**	**119.4**	-33.0%
GPT-4 + GPT-4	0.50	237.5	-	0.69	145.2	-	**0.43**	240.7	-
+*AutoForm*	**0.52**	**146.2**	-38.4%	**0.76**	**115.0**	-20.8%	**0.43**	**141.7**	-41.1%

Table 2: Comparative performance in multi-agent communication across various QA datasets. The table highlights RougeL scores, with better performance in different model pairing settings indicated in bold. The ΔTokens column quantifies the token reduction achieved by the AutoForm method.

(RQs). These questions aim to unravel the intricacies of format selection by LLMs and its impact on task performance across various scenarios:

- **RQ1**: Can LLMs select the suitable formats autonomously? (Section 4.2)
- **RQ2**: What formats are chosen in single-LLM reasoning? (Section 4.3)
- **RQ3**: Can LLMs devise a general format for a task based on some task inputs? (Section 4.4)
- **RQ4**: Is the decided format transferable between different LLMs? (Section 4.5)
- **RQ5**: What are the features of the formats used in multi-agent communication? (Section 4.6)
- **RQ6**: Does the autonomously determined multi-agent communication format align with conventional agent communication languages such as KQML (Finin et al., 1994)? (Section 4.7)

4.2 RQ1: The Capability of LLMs on Selecting Suitable Format

Single-LLM Reasoning. The comparative efficacy of the AutoForm approach over the conventional Chain-of-Thought (CoT) methodology in single-LLM reasoning tasks is encapsulated in Table 1. We observe clear performance improvements when employing AutoForm across different datasets, compared to the baseline CoT method.

For GPT-3.5, the implementation of AutoForm leads to a significant improvement in accuracy, particularly notable in the Coin Flip dataset, where accuracy escalates from 22.2% to 38.0%. This substantial increase highlights the model's enhanced ability in symbolic reasoning. Across other datasets, AutoForm yields consistent enhancements, with increases generally ranging between 3% to 5%, culminating in an overall average performance boost of 5.4%. Similarly, for Gemini Pro, AutoForm achieves an average performance enhancement of 5.7%. GPT-4 also benefits from AutoForm, with an average performance uplift of 3.3% across all tasks. These consistent improvements across various datasets demonstrate the method's model-agnostic robustness and the efficacy of utilizing alternative formats in reasoning tasks. It also suggests that alternative formats, apart from NL, can aid LLMs in task resolution. It is just that without explicit reminders, the LLMs do not explore alternative formats. For a detailed illustration of how LLM responses are influenced by AutoForm, refer to Appendix D, which presents selected examples from each LLM for every task.

Multi-Agent Communication. The outcomes of our multi-agent communication experiments, detailed in Table 2, provide valuable insights into the efficiency and effectiveness of utilizing alternative communication formats in collaborative environments. In this experiment, we experiment with different model pairings to explore this robustness more comprehensively. Since the initial speaking agent often sets the tone for the communication format, we vary the speaking order in pairings of heterogeneous models, e.g., *GPT-4 + GPT-3.5* in the table indicates GPT-4 is the initiator. Due to page limit, we place results where GPT-3.5 initiates the conversation, and where the supporting facts are split and distributed to two agents at Appendix C.

A key finding from our experiments is the notable efficiency achieved through AutoForm, as evidenced by the substantial reduction in token usage across diverse model pairings and speaking orders. This efficiency, quantified in the ΔTokens column, illustrates the capability of LLMs to move beyond their typical NL to adopt more concise and efficient communication formats. This is particularly evident in the Hotpot QA dataset with the GPT-4 and GPT-3.5 pairing, where we witness a token reduction reaching 72.7%. On most of the other datasets, AutoForm also obtains substantial token reduction. These findings imply that LLMs, though extensively trained on NL, are capable of exploring and employing alternative formats to enhance communication efficiency. Detailed case studies further elucidating the features of the chosen formats will be presented in Section 4.6, and more cases are

模型	Wiki Hop			Hotpot QA			叙事问答		
	RougeL	词元数	Δ令牌	RougeL	词元数	Δ令牌	RougeL	词元数	Δ令牌
GPT-4 + GPT-3.5	**0.53**	281.5	-	0.63	345.5	-	0.43	178.3	-
+*AutoForm*	**0.53**	**255.0**	-9.4%	**0.70**	**94.3**	-72.7%	**0.48**	**119.4**	-33.0%
GPT-4 + GPT-4	0.50	237.5	-	0.69	145.2	-	**0.43**	240.7	-
+*AutoForm*	**0.52**	**146.2**	-38.4%	**0.76**	**115.0**	-20.8%	**0.43**	**141.7**	-41.1%

表2：多智能体通信在不同问答数据集上的性能比较。表格突出了 RougeL 分数，在不同模型配对设置下表现更好的结果以粗体显示。 Δ令牌列量化了 AutoForm 方法实现的令牌减少量。

（RQs）。这些问题旨在揭示大语言模型格式选择的复杂性，以及其对各种情景下任务性能的影响：

- 研究问题1：大语言模型能否自主选择合适的格式？（第4.2节）
- 研究问题2：单大语言模型推理中会选择哪些格式？（第4.3节）
- 研究问题3：大语言模型能否根据一些任务输入为任务设计一种通用的格式？（第4.4节）

研究问题4：已决定的格式是否可以在不同的大语言模型之间迁移？（第4.5节）

- RQ5：多智能体通信中所使用的格式具有哪些特征？（第4.6节）
- RQ6：自主确定的多智能体通信格式是否与KQML（Finin等，1994年）等传统智能体通信语言相符？（第4.7节）

4.2 研究问题1：大语言模型选择合适格式的能力

单大语言模型推理。表1总结了在单大语言模型推理任务中，AutoForm方法与传统的思维链 (CoT)方法相比的比较效能。与基线思维链方法相比，我们观察到在不同数据集中采用AutoForm时，性能有了明显的提升。

对于GPT-3.5，AutoForm的实现带来了准确率的显著提升，在抛硬币数据集中尤为明显，准确率从22.2%上升至38.0%。这一大幅跃升凸显了模型在符号推理方面增强后的能力。在其他数据集中，AutoForm也带来了稳定的提升，提升幅度通常介于3%至5%之间，最终使得整体平均性能提升了5.4%。类似地，对于Gemini Pro，AutoForm实现了平均5.7%的性能增强。GPT-4同样受益于AutoForm，在所有任务中的平均性能提升了3.3%。这些在多个数据集上的一致改进证明了该方法的模型无关鲁棒性以及在推理任务中利用替代格式的有效性。它还表明，除了自然语言之外，替代格式也有助于大语言模型解决任务问题。只是在缺乏明确提示的情况下，大语言模型不会主动探索替代格式。关于大语言模型响应如何受到AutoForm影响的详细说明，请参阅附录D，其中展示了每个大语言模型在每个任务中的若干示例。

多智能体通信。我们在表2中详述的多智能体通信实验的结果，为了解在协作环境中利用替代通信格式的效率与有效性提供了宝贵的见解。在该实验中，我们尝试不同的模型配对，以更全面地探索这种鲁棒性。由于初始发言智能体通常决定了通信格式的氛围，我们在异构模型的配对中改变了发言顺序，例如，表格中的GPT-*4* + GPT-3.5表示GPT-4是发起者。由于篇幅限制，我们将GPT-3.5发起对话的结果，以及支持事实被拆分并分配给两个智能体的结果放置于附录 C。

我们实验的一项关键发现是，通过 AutoForm 实现的无表格（no-table）效率，这一点从各种模型配对和发言顺序下令牌使用量的大幅减少中得到了证实。这种效率量化在 ΔTokens 列中，说明了大语言模型有能力超越其典型的自然语言，采用更简洁、高效的通信格式。这一点在 Hotpot QA 数据集上的 GPT-4 与 GPT-3.5 配对中尤为明显，我们观察到令牌减少幅度达到 72.7%。在大多数其他数据集中，AutoForm 也实现了大幅度的令牌减少。这些发现表明，尽管大语言模型在自然语言上经过了广泛训练，但它们能够探索并采用替代格式以提升通信效率。详细阐述所选格式特征的案例研究将在第 4.6 节中呈现，更多案例将在

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

presented at Appendix D.

Furthermore, the effectiveness of the multi-agent communication facilitated by AutoForm, as gauged by RougeL scores, is found to be largely comparable to, and occasionally exceeding, that of natural language-based interactions. This is especially true when GPT-4 initiates the conversation, suggesting that more advanced LLMs possess a better ability to select communication formats that strike a balance between conciseness and clarity. Conversely, we show in Appendix C that with GPT-3.5 as the initiator, despite the similar notable efficiency in token usage, the performance is generally akin to or slightly below that achieved with natural language. These observations highlight the intricate challenge of selecting an appropriate communication format, a task that proves demanding even for sophisticated LLMs, and the importance of balancing brevity with the need to preserve the integrity of the communicative content.

4.3 RQ2: Formats Chosen in Single-LLM Reasoning

In addressing RQ1, we investigate the formats selected from LLMs when tasked with reasoning. This analysis is pivotal in understanding how LLMs, when granted the autonomy to choose, navigate away from the default NL format to potentially more efficient alternatives.

We analyze a randomly sampled set of 50 examples from each dataset, investigate the reasoning traces produced with AutoForm, and manually count the number of appeared formats. Fig. 3(a-c) display the distribution of formats chosen by Gemini Pro, GPT-3.5 and GPT-4, and Fig. 3(d) displays their combined preferences across various tasks. The data indicates a notable diversity in format selection by LLMs. A shift towards structured formats, such as lists, logical expressions, and markdown tables is observed. These formats are particularly favored in tasks that demand logical reasoning, offering clearer and more concise data representation, as illustrated in Fig. 3(d).

4.4 RQ3: Generalization of Format Selection Based on Task Inputs

An examination of Fig. 3(d) reveals a discernible variation in the LLMs' format preferences across different tasks. This variation aligns with the expectation that the optimal format would naturally differ between tasks, each with its unique requirements. In addressing RQ2, we probe whether LLMs are capable of identifying a general format suitable for a given task based on a subset of inputs, and then consistently applying this format for problem-solving. In AutoForm, as delineated in Section 3.2, LLMs typically select a format *implicitly* for each instruction on a case-by-case basis. Nonetheless, it stands to reason that certain tasks may be inherently conducive to a specific format. To investigate this hypothesis, we introduce the **two-step AutoForm**. This approach tasks an LLM with first determining the most efficient format and subsequently utilizing that format in the CoT problem-solving stage. That is, we turn the implicit format decision into an explicit step, which mathematically follows Eq. (1) instead of merging the two steps.

The two-step AutoForm experiments with two distinct settings: 1) *Instance-Based*, where the LLM selects a format for each instruction, and 2) *Task-Based*, where the LLM deduces a general format for the entire task by analyzing 5 inputs from the task. Note that, unlike few-shot prompting, the Task-Based setting does not provide answers within the inputs, and these inputs are only utilized during the format decision step.

The results are detailed in Table 3. The Task-Based setting demonstrates that both GPT-3.5 and Gemini Pro can effectively generalize a thought format from a limited set of inputs within a task, and often outperforming the Instance-Based setting. In contrast, GPT-4 shows similar performance levels in both Task-Based and Instance-Based settings, suggesting that its advanced capabilities may afford it greater flexibility in format usage. These findings indicate that LLMs, particularly less sophisticated ones like GPT-3.5, may benefit from exposure to multiple inputs from a task to better generalize an effective thought format. This ability of LLMs to generalize the format for a task makes the AutoForm approach more practical since the format can be identified only once for a specific task.

4.5 RQ4: Transferability of Format Across Different LLMs

This subsection delves into RQ3, the transferability of the format across LLMs. The concept of format transferability is crucial in understanding the universality of formats decisions made by LLMs and their applicability across various models. Instead of using the same model for format selection and problem-solving (homogeneous setting) as we have done in Section 4.4, we explore the heterogeneous setting of the two-step AutoForm, where different

附录 D 中呈现。

此外，AutoForm 所促进的多智能体通信的有效性（根据 RougeL 分数衡量），发现其与基于自然语言的交互效果大致相当，偶尔甚至更优。当 GPT-4 发起对话时，这一现象尤为明显，这表明更高级的大语言模型在选择通信格式方面具有更强的能力，能够在简洁性与清晰度之间取得平衡。相反，我们在附录 C 中展示，当 GPT-3.5 作为发起者时，尽管令牌使用效率相似且显著，但总体表现通常与使用自然语言时的效果相当或略低。这些观察结果凸显了选择合适通信格式这一复杂挑战的难度——即使是精密的大语言模型也面临这一难题——以及在简洁性与保持通信内容完整性之间取得平衡的重要性。

4.3 研究问题2：单大语言模型推理中选择的格式

在解决研究问题1时，我们研究了当大语言模型被赋予推理任务时所选择的格式。这一分析对于理解大语言模型在拥有自主选择的权利时如何偏离默认自然语言格式，转向可能更高效的选择至关重要。

我们分析了从每个数据集中随机抽样的50个示例，研究了使用AutoForm生成的推理轨迹，并手动统计了出现格式的实例数量。图 3(a-c) 展示了 Gemini Pro 、 GPT-3.5 和 GPT-4 所选择格式的分布，图 3(d) 则展示了它们在各种任务上的综合偏好。数据显示，大语言模型在格式选择上存在显著多样性。观察到向列表、逻辑表达式和 Markdown 表格等结构化格式的转变。正如 图 3(d) 所示，这些格式在需要逻辑推理的任务中尤其受到青睐，因为它们能提供更清晰、更简洁的数据表示。

4.4 研究问题3：基于任务输入的格式选择的泛化

对图 3(d) 的考察揭示了大语言模型在不同任务间的格式偏好存在可察觉的差异。这种差异符合预期，即不同任务因其各自独特的需求，最优格式自然会有所不同。在研究问题2中，我们探索大语言模型是否能够基于给定任务的部分输入识别出一种通用的格式，并随后一致地应用该格式进行问题解决。在 AutoForm 中，如 3.2 节所述，大语言模型通常针对每条指令，逐案隐式地选择一种格式。然而，可以认为某些任务可能天然适合特定格式。为检验这一假设，我们引入两步 AutoForm。该方法要求大语言模型首先确定最高效的格式，随后在思维链问题解决阶段利用该格式。也就是说，我们将隐式的格式决策转化为一个显式步骤，这在数学上遵循公式 (1)，而非将两个步骤合并。

两步 AutoForm 实验包含两种不同的设置：1) 基于实例，即大语言模型为每条指令选择一种格式；2) 基于任务，即大语言模型通过分析任务中的 5 个输入来推断整个任务的通用格式。请注意，与少样本提示不同，基于任务的设置不会在输入中提供答案，这些输入仅在格式决策步骤中使用。

结果详见表3。基于任务设置表明，GPT-3.5和Gemini Pro均能有效从任务内的有限输入集中泛化出思维格式，且通常表现优于基于实例设置。相比之下，GPT-4在基于任务和基于实例两种设置中表现出相似的性能水平，这表明其高级能力可能赋予其在格式使用上更大的灵活性。这些发现表明，大语言模型，尤其是像GPT-3.5这样功能较弱的模型，可能从接触任务中的多个输入中获益，从而更好地泛化出有效的思维格式。大语言模型将特定任务的格式进行泛化的能力，使得Auto-Form方法更具实用性，因为针对特定任务只需识别一次格式。

4.5 研究问题4：不同大型语言模型间的格式迁移性

本小节深入探讨研究问题3，即格式在大语言模型之间的迁移性。格式迁移性的概念对于理解大语言模型所作格式决策的普遍性及其在不同模型中的适用性至关重要。与我们在4.4节中为格式选择和问题解决使用同一模型（异质设置）的做法不同，我们探索了两步 AutoForm 的异质设置，其中不同的

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

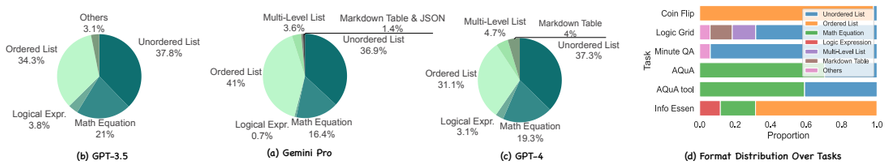


Figure 3: Format distribution chosen by Gemini Pro (a), GPT-3.5 (b) and GPT-4 (c), and the overall format distribution across tasks from both models (d).

Model	Logic Grid	Coin Flip	Info Essen	MM QA	AQuA	Average
GPT-3.5 (Instance-Based)	48.0	49.2	39.7	26.1	**66.5**	$45.9_{(+7.5\%)}$
GPT-3.5 (Task-Based)	**51.0**	**62.8**	**42.6**	**28.1**	65.0	$\mathbf{49.9}_{(+11.5\%)}$
Gemini Pro (Instance-Based)	39.5	44.8	**35.3**	27.1	**59.4**	$41.2_{(+2.5\%)}$
Gemini Pro (Task-Based)	**41.5**	**47.8**	**35.3**	**28.6**	**59.4**	$\mathbf{42.5}_{(+3.8\%)}$
GPT-4 (Instance-Based)	**71.5**	**100.0**	**76.5**	**41.4**	78.3	$\mathbf{73.5}_{(+3.2\%)}$
GPT-4 (Task-Based)	70.0	99.8	75.0	**41.4**	**79.5**	$73.1_{(+2.8\%)}$
GPT-4 ⇒ GPT-3.5 (Task-Based)	47.5	83.0	35.3	23.2	59.4	$49.7_{(+11.2\%)}$
GPT-4 ⇒ Gemini Pro (Task-Based)	40.0	74.6	25.0	25.6	47.2	$42.5_{(+3.8\%)}$
GPT-3.5 ⇒ GPT-4 (Task-Based)	65.5	98.6	72.1	42.9	79.5	$71.7_{(+1.3\%)}$

Table 3: Comparative performance of two-step AutoForm with single LLM reasoning across various datasets. The notation $model_1 \Rightarrow model_2$ denotes using $model_1$ for format selection, and $model_2$ for problem-solving. Average performance improvements over CoT results, as presented in Table 1, are denoted with a subscript.

models are employed for the two steps.

The results are presented in the last group of Table 3. Generally, the format is transferable, but may lead to slightly inferior performances compared to the homogeneous setting. For example, transferring the format decided by GPT-4 to GPT-3.5 or Gemini Pro leads to a decrease compared to the homogeneous setting for the two LLMs in most of the tasks except for Coin Flip. On the other hand, when transferring the format decided by GPT-3.5 to GPT-4, the results are generally comparable to the homogeneous setting for GPT-4. For information essentiality, the format selected by GPT-3.5 is generally less efficient, leading to incomplete problem-solving processes. We give examples in Tables 33, 36 and 39. For other tasks, the format generated by GPT-3.5 proves adequate and is easily interpreted by GPT-4, resulting in similar performances to the homogeneous setting for GPT-4.

4.6 RQ5: Features of Communication Format

In addressing RQ4, this subsection investigates the characteristics of the formats used by language agents in multi-agent communication scenarios. Our goal is to identify key attributes contributing to efficient communication by examining the formats used during their interactions.

We analyze 50 random interaction logs for each dataset and present some cases in Fig. 4 and Appendix D. Despite some retained characteristics of NL, the communication formats decided via AutoForm display distinct features:

Clarity and Structure. An important feature of the selected formats is an emphasis on clarity. LLMs consistently favor formats facilitating unambiguous and straightforward communication, which is vital in our multi-agent scenarios. In these scenarios, agents possess divergent knowledge sets, making the clear exchange of this distinct information indispensable. Structured formats, which provide an organized method of presenting information, are also prevalent. These formats enhance the comprehensibility and accessibility of the content. Contrasting this with the use of NL, we observe that the LLM-decided formats tend to be more direct and clear, effectively reducing redundancy.

Brevity and Efficiency. Another key feature is the focus on brevity, enhancing communication efficiency. Formats chosen by LLMs often omit elements like pleasantries or emotive expressions, resulting in concise exchanges. This brevity conserves computational resources and concentrates dialogue on the task at hand, optimizing the communication process for faster, more efficient information exchange. This is particularly beneficial in scenarios requiring rapid and effective decision-making.

4.7 RQ6: Alignment with Conventional Agent Communication Languages

RQ5 probes the extent to which the communication formats determined by LLMs using AutoForm

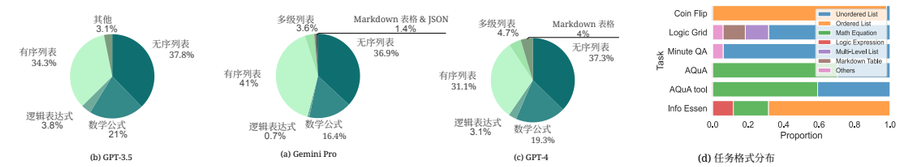


图 3：Gemini Pro (a)、GPT-3.5 (b) 和 GPT-4 (c) 选择的格式分布，以及来自这两个模型的任务总体格式分布 (d)。

模型	逻辑网格	抛硬币	Info Essen	MM QA	AQuA	平均值
GPT-3.5（基于实例）	48.0	49.2	39.7	26.1	**66.5**	$45.9_{(+7.5\%)}$
GPT-3.5（基于任务）	**51.0**	**62.8**	**42.6**	**28.1**	65.0	$\mathbf{49.9}_{(+11.5\%)}$
Gemini Pro（基于实例）	39.5	44.8	**35.3**	27.1	**59.4**	$41.2_{(+2.5\%)}$
Gemini Pro（基于任务）	**41.5**	**47.8**	**35.3**	**28.6**	**59.4**	$\mathbf{42.5}_{(+3.8\%)}$
GPT-4（基于实例）	**71.5**	**100.0**	**76.5**	**41.4**	78.3	$\mathbf{73.5}_{(+3.2\%)}$
GPT-4（基于任务）	70.0	99.8	75.0	**41.4**	**79.5**	$73.1_{(+2.8\%)}$
GPT-4 ⇒ GPT-3.5（基于任务）	47.5	83.0	35.3	23.2	59.4	$49.7_{(+11.2\%)}$
GPT-4 ⇒ Gemini Pro（基于任务）	40.0	74.6	25.0	25.6	47.2	$42.5_{(+3.8\%)}$
GPT-3.5 ⇒ GPT-4（基于任务）	65.5	98.6	72.1	42.9	79.5	$71.7_{(+1.3\%)}$

表3：两步AutoForm与单大语言模型推理在多个数据集上的对比性能。符号 $model_1 \Rightarrow model_2$ 表示使用 $model_1$ 进行格式选择，使用 $model_2$ 进行问题解决。与思维链结果的平均性能提升如表1所示，以下标形式标注。

模型被用于这两个步骤。

结果列于表 3 的最后一组中。总体而言，通信格式是可迁移的，但相较于同质性设置，其性能可能会略有下降。例如，将 GPT-4 决定的通信格式迁移至 GPT-3.5 或 Gemini Pro，在除抛硬币之外的多数任务中，会导致这两大语言模型的性能相较于同质性设置有所下降。另一方面，当将 GPT-3.5 决定的通信格式迁移至 GPT-4 时，对于 GPT-4 而言，结果通常与同质性设置相当。就信息必要性而言，GPT-3.5 选择的通信格式通常效率较低，导致问题解决过程不完整。我们在表 33、表 36 和表 39 中给出了示例。对于其他任务，GPT-3.5 生成的通信格式被证明是充分的，并且易于被 GPT-4 解读，其表现与 GPT-4 在同质性设置下的表现相似。

4.6 RQ5：通信格式 的特性

在解决研究问题4时，本小节考察了语言智能体在多智能体通信情景中所使用的格式的特征。我们的目标是通过检查它们在交互过程中使用的格式，来识别有助于高效通信的关键属性。

我们对每个数据集分析了50个随机交互日志，并在图4和附录D中展示了一些案例。尽管仍保留了自然语言的一些特征，但通过自动.....决定的通信格式形式呈现出不同的特征：

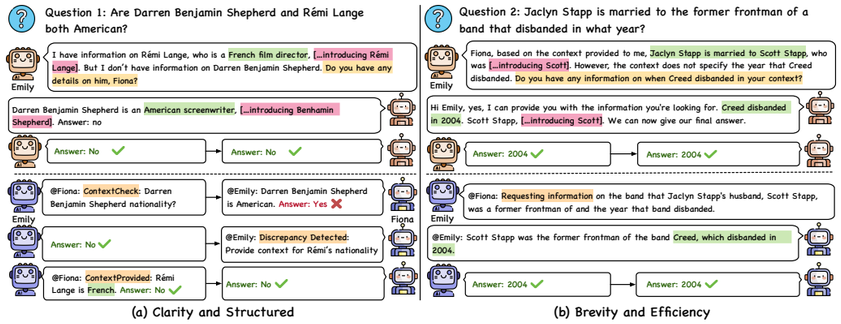
清晰度与结构。所选格式的一个重要特征是强调清晰度。大语言模型始终偏好那些促进无歧义且直接通信的格式，这在我们的多智能体场景中至关重要。在这些情景中，智能体拥有不同的知识集，使得这些不同信息的清晰交换不可或缺。结构化格式提供了一种组织化的信息呈现方法，也较为普遍。这些格式增强了内容的可理解性和可访问性。与之形成对比的是，我们观察到由大语言模型决定的格式往往更加直接和清晰，有效减少了冗余。

简洁性与效率。另一个关键特性是强调简洁性，以提升通信效率。大语言模型选择的格式通常会省略寒暄或情感表达等元素，从而实现简洁的交流。这种简洁性节省了计算资源，并使对话集中在当前任务上，优化了通信过程，以实现更快速、更高效的信息交换。这在需要快速有效决策的情景中尤为有益。

4.7 RQ6：与传统智能体通信语言的对齐

RQ5考察了大语言模型利用AutoForm确定的通信格式的范围

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next



(a) Clarity and Structured (b) Brevity and Efficiency

Figure 4: Multi-agent communication examples. The top panel illustrates a traditional natural language conversation, and the bottom panel shows a conversation using AutoForm. Necessary information related to the question is marked in green, redundant information is marked in red, and speech-act-related phrases are marked in orange.

align with traditional Agent Communication Languages (ACLs), such as KQML (Finin et al., 1994) and FIPA-ACL (FIPA, 2001). These ACLs have been instrumental in structuring communication between intelligent agents to foster cooperation and coordination. A typical KQML message, as shown below, exemplifies the structured nature of traditional ACLs:

```
(ask-one
  :sender joe
  :content (PRICE IBM ?price)
  :receiver stock-server
  :reply-with ibm-stock
  :language LPROLOG
  :ontology NYSE-TICKS)
```

Our examination of the communication patterns emerging from AutoForm reveals an interesting resemblance to these structured elements. As depicted in Fig. 4, LLMs frequently employ a structured format where "@" denotes the *receiver*, verb phrases such as "ContextCheck" indicate the *performative* (the *"ask-one"* in the above example), and a succinct text string encapsulates the *content*. This structuring mirrors the composition of ACL messages, where each part serves a specific function in the communication process.

Intrigued by this similarity, we conduct an experiment where LLMs are prompted to communicate using a format similar to KQML:

```
(performative
  :content ...
  :receiver ...)
```

The results of this experiment on two GPT-4-based agents are presented in Table 4. We have two settings, one prompts the LLMs to use the exact KQML format as presented above, and the other setting uses the JSON version of KQML format, considering that LLMs may be more adept at JSON.

	Hotpot QA		Wiki Hop		Narrative QA	
Format	RougeL	#Tokens	RougeL	#Tokens	RougeL	#Tokens
KQML	**0.76**	313.8	**0.53**	368.1	0.28	343.3
JSON	0.71	346.0	**0.53**	291.4	0.22	385.2
AutoForm	**0.76**	**115.0**	0.52	**146.2**	**0.43**	**141.7**

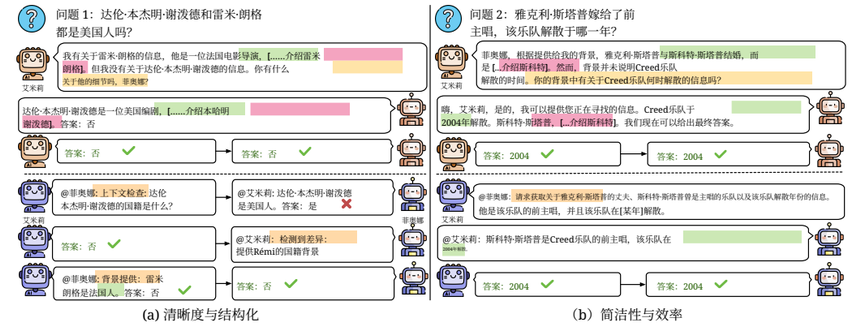
Table 4: Multi-agent communication performances using conventional ACL format.

While both settings show worse or comparable performance to AutoForm in terms of RougeL, the number of tokens consumed is larger than AutoForm. This finding suggests that while LLMs can indeed emulate the formality of traditional ACL formats, the AutoForm approach optimizes the communication by enhancing clarity and structure, yet concurrently reduces token usage.

These results highlight two key implications. First, AutoForm can generate communication patterns similar to established ACLs. Second, it efficiently distills these traditional formats into a more concise form, conserving computational resources while maintaining communicative effectiveness. This balance of clarity, structure, and brevity makes AutoForm a powerful tool for facilitating intelligent agent communication in various contexts.

5 Conclusion

In this work, we demonstrate that LLMs can autonomously determine suitable non-NL formats for reasoning and communication using the AutoForm prompting method. Our analyses address six key research questions, showing that LLMs can generalize a reasoning format from task-specific examples and transfer it across different models. Additionally, the communication formats generated by LLMs resemble traditional ACLs, offering both precision and efficiency. These insights



(a) 清晰度与结构化 （b）简洁性与效率

图4：多智能体通信示例。最高面板展示了一个传统的自然语言对话，最低面板展示了一个使用AutoForm的对话。与问题相关的必要信息以绿色标记，冗余信息以红色标记，与言语行为相关的短语以橙色标记。

与传统智能体通信语言（ACLs）保持一致，例如 KQML（Finin等，1994年）和 FIPA-ACL（FIPA，2001年）。这些 ACLs 在结构化智能体间的通信以促进合作与协调方面发挥了重要作用。如下所示的典型 KQML 消息，展示了传统 ACLs 的结构化特性：

(ask-one :sender 乔 :content (价格 IBM ?price) :receiver 股票服务器 :reply-with ibm-stock : language LPROLOG :ontology NYSE-TICKS)

我们对AutoForm中出现的通信模式的审视，揭示了其与这些结构化元素的有趣相似性。如图4所示，大语言模型（LLMs）频繁采用一种结构化格式，其中“@”表示接收者，“ ContextCheck”等动词短语表示施为行为（即上文示例中的“ask-one”），而一个简洁的文本字符串则封装了内容。这种结构化与ACL消息的构成相仿，其中每个部分都在通信过程中承担特定功能。

受此相似性的启发，我们开展了一项实验，其中大语言模型被提示使用类似于KQML的格式进行通信：

(行为式 :内容 ... : 接收者 ...)

该实验在两个基于 GPT-4 的智能体上的结果如表4所示。我们设置了两种方案：一种提示词引导大语言模型使用如上所述的精确 KQML 格式，另一种方案则使用 KQML 格式的 JSON 版本，考虑到大型语言模型可能更擅长处理 JSON。

	Hotpot QA		Wiki Hop		叙事问答	
格式	RougeL	#令牌	RougeL	#令牌	RougeL	#令牌
KQML	**0.76**	313.8	**0.53**	368.1	0.28	343.3
JSON	0.71	346.0	**0.53**	291.4	0.22	385.2
AutoForm	**0.76**	**115.0**	0.52	**146.2**	**0.43**	**141.7**

表4：使用传统 ACL 格式的多智能体通信表现。

尽管在 RougeL 指标上，两种方案的单句表现均不如 AutoForm 或与之相当，但所消耗的令牌数量多于 AutoForm。这一发现表明，尽管大语言模型确实能够模仿传统 ACL 格式的形式化特征，但 AutoForm 方法通过增强清晰度和结构来优化通信，同时减少了令牌使用。

这些结果凸显了两个关键含义。首先，AutoForm 能够生成与既定智能体通信语言相似的通信模式。其次，它高效地将这些传统格式提炼为更简洁的形式，在保持通信有效性的同时节约计算资源。这种清晰度、结构与简洁性的平衡，使 AutoForm 成为促进各种上下文中智能体通信的强大工具。

5 结论

在本研究中，我们证明大语言模型可以利用AutoForm提示方法，自主确定适合推理和智能体通信的非自然语言格式。我们的分析针对六个关键研究问题，表明大语言模型可以从特定任务的示例中归纳出一种推理格式，并将其迁移到不同的大语言模型中。此外，由大语言模型生成的通信格式类似于传统的智能体通信语言（ACLs），兼具精确性和高效性。这些见解

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

enhance our understanding of LLMs' capabilities beyond NL, improving LLM reasoning and inter-agent communication.

Limitations

Despite we have shown many kinds of formats can facilitate LLMs reasoning and communication, the scope of alternative formats explored is still not exhaustive. The potential of numerous other formats and their specific applications to various LLM architectures warrants further investigation.

Moreover, the generalization of chosen formats across tasks, while promising, shows variability in effectiveness depending on the complexity of the task and the specific LLM used. This variability highlights the nuanced nature of format suitability and its impact on task performance, suggesting that further exploration is necessary to fully harness the capabilities of alternative formats.

References

Anthropic. 2023. Introducing claude 2.1.

Daniil A. Boiko, Robert MacKnight, and Gabe Gomes. 2023. Emergent autonomous scientific research capabilities of large language models. *CoRR*, abs/2304.05332.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *CoRR*, abs/2308.07201.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chen Qian, Chi-Min Chan, Yujia Qin, Yaxi Lu, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2023. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents. *CoRR*, abs/2308.10848.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *CoRR*, abs/2211.12588.

Noam Chomsky. 2006. *Language and mind*. Cambridge University Press.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *CoRR*, abs/2305.14325.

Timothy W. Finin, Richard Fritzson, Donald P. McKay, and Robin McEntire. 1994. KQML as an agent communication language. In *Proceedings of the Third International Conference on Information and Knowledge Management (CIKM'94), Gaithersburg, Maryland, USA, November 29 - December 2, 1994*, pages 456–463. ACM.

FIPA. 2001. *FIPA ACL Message Structure Specification*. FIPA.

Jerry A Fodor. 1975. *The language of thought*, volume 5. Harvard university press.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. PAL: program-aided language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 10764–10799. PMLR.

Gemini Team Google, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, and Chenglin Wu. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *CoRR*, abs/2308.00352.

Tomás Kociský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Trans. Assoc. Comput. Linguistics*, 6:317–328.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *NeurIPS*.

George Lakoff. 2008. *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago press.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023a. CAMEL: communicative agents for "mind" exploration of large scale language model society. *CoRR*, abs/2303.17760.

Ruosen Li, Teerth Patel, and Xinya Du. 2023b. PRD: peer rank and discussion improve large language model based evaluations. *CoRR*, abs/2307.02762.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 158–167. Association for Computational Linguistics.

加深了我们对大语言模型在自然语言之外能力的理解，从而改善了大语言模型推理和智能体通信。

局限性

尽管我们已经展示了多种格式可以促进大语言模型的推理和通信，但所探索的替代格式范围仍然不够全面。众多其他格式的潜力以及它们对各大语言模型架构的具体应用仍有待进一步研究。

此外，所选格式在跨任务上的泛化能力虽然前景广阔，但其有效性会因任务的复杂度和具体使用的大语言模型的不同而存在波动。这种波动凸显了格式适用性的复杂性及其对任务性能的影响，表明需要进一步探索才能充分发挥替代格式的潜力。

参考文献

Anthropic. 2023. Introducing Claude 2.1.

Daniil A. Boiko, Robert MacKnight, and Gabe Gomes. 2023. 大语言模型的自主科学研究的涌现能力. CoRR, abs/2304.05332. Chi-Min Chan, 陈伟泽, 苏宇声, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and 刘知远. 2023. Chateval: Towards better 基于LLM的评估器 through 多智能体辩论. CoRR, abs/2308.07201. 陈伟泽, 苏宇声, Jingwei Zuo, 杨程, 袁晨飞, 钱陈, Chi-Min Chan, Yujia Qin, Yaxi Lu, 谢若冰, 刘知远, 孙茂松, and 周杰. 2023. AgentVerse: Facilitating multi- agent collaboration and exploring emergent behav- iors in 智能体. CoRR, abs/2308.10848. Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2022. 思维程序提示: Disentangling 计算 from 推理 for 数值推理任务. CoRR, abs/2211.12588. Noam Chomsky. 2006. 语言与心智. Cam- bridge University Press. Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. 通过多智能体辩论提高语言模型的事实性和推理能力. CoRR, abs/2305.14325. Timothy W. Finin, Richard Fritzson, Donald P. McKay, and Robin McEntire. 1994. KQML as an agent 通信 language. In Proceedings of the Third 国际信息与知识管理会议 (CIKM'94)，盖瑟斯堡，马里兰州，美国，1994年11月29日至12月2日，第456–463页。美国计算机协会。FIPA. 2001. FIPA ACL消息结构规范。FIPA. Jerry A Fodor. 1975. 思想的语言，第5卷。哈佛大学出版社。Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, 和 Graham Neubig. 2023. PAL: 程序辅助语言模型。收录于国际机器学习会议，ICML 2023，2023年7月23-29日，檀香山，夏威夷，美国，机器学习研究论文集，第202卷，第10764–10799页。PMLR。Google Gemini团队, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, 等. 2023. Gemini: 一个高能力多模态模型家族。arXiv预印本 arXiv:2312.11805。Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, 和 Chenglin Wu. 2023. MetaGPT: 面向多智能体协作框架的元编程。CoRR, abs/2308.00352。Tomás Kociský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, 和 Edward Grefenstette. 2018. NarrativeQA阅读理解挑战。计算语言学协会事务, 6:317–328。Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, 和 Yusuke Iwasawa. 2022. 大语言模型是零样本推理者。收录于NeurIPS。George Lakoff. 2008. 女人、火和危险的东西：范畴揭示了什么关于心智。芝加哥大学出版社。Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, 和 Bernard Ghanem. 2023a. CAMEL: 用于"心智"探索大规模语言模型社会的通信智能体。CoRR, abs/2303.17760。Ruosen Li, Teerth Patel, 和 Xinya Du. 2023b. PRD: 同伴排名和讨论改进基于大语言模型的评估。CoRR, abs/2307.02762。Wang Ling, Dani Yogata ma, Chris Dyer, 和 Phil Blunsom. 2017. 通过理由生成进行程序归纳：学习与解释代数文字问题。收录于第55届计算语言学协会年会，ACL 2017，温哥华，加拿大，7月30日至8月4日，第1卷：长文，第158–167页。计算语言学协会。

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

Tengxiao Liu, Qipeng Guo, Yuqing Yang, Xiangkun Hu, Yue Zhang, Xipeng Qiu, and Zheng Zhang. 2023. Plan, verify and switch: Integrated reasoning with diverse x-of-thoughts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2807–2822. Association for Computational Linguistics.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023. GAIA: a benchmark for general AI assistants. *CoRR*, abs/2311.12983.

OpenAI. 2023a. Assistants api.

OpenAI. 2023b. GPT-4 technical report. *CoRR*, abs/2303.08774.

Anton Osika. 2023. gpt-engineer.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST 2023, San Francisco, CA, USA, 29 October 2023- 1 November 2023*, pages 2:1–2:22. ACM.

Chau Pham, Boyi Liu, Yingxiang Yang, Zhengyu Chen, Tianyi Liu, Jianbo Yuan, Bryan A. Plummer, Zhaoran Wang, and Hongxia Yang. 2023. Let models speak ciphers: Multiagent debate through embeddings. *CoRR*, abs/2310.06272.

Steven Pinker. 2003. *The language instinct: How the mind creates language*. Penguin uK.

Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software development. *CoRR*, abs/2307.07924.

Noah Shinn, Beck Labash, and Ashwin Gopinath. 2023. Reflexion: an autonomous agent with dynamic memory and self-reflection. *CoRR*, abs/2303.11366.

Significant Gravitas. AutoGPT.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew K. Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, and et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *CoRR*, abs/2206.04615.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Trans. Assoc. Comput. Linguistics*, 6:287–302.

Benjamin Lee Whorf. 2012. *Language, thought, and reality: Selected writings of Benjamin Lee Whorf*. MIT press.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Autogen: Enabling next-gen LLM applications via multi-agent conversation framework. *CoRR*, abs/2308.08155.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2369–2380. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models. *CoRR*, abs/2305.10601.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023b. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2023. Webarena: A realistic web environment for building autonomous agents. *CoRR*, abs/2307.13854.

刘晓腾, 郭启鹏, 杨雨晴, 胡向坤, 张岳, 邱锡鹏, 和张政. 2023. 计划、验证与切换：具有多样化思维路径的集成推理。收录于2023年自然语言处理实证方法会议论文集，EMNLP 2023，新加坡，2023年12月6-10日，第2807–2822页。计算语言学协会。

格雷瓜尔·米亚隆，克莱门汀·福里尔，克雷格·斯威夫特，托马斯·沃尔夫，杨立昆，以及托马斯·斯卡洛姆。2023。GAIA：通用AI助手的基准测试。CoRR，abs/2311.12983。

OpenAI。2023a。助手API。

OpenAI。2023b。GPT-4技术报告。CoRR，abs/2303.08774。

安东·奥西卡。2023。gpt-engineer。

朴俊成，约瑟夫·C·奥布莱恩，蔡俊，梅雷迪思·林格尔·莫里斯，梁佛麟，以及迈克尔·S·伯恩斯坦。2023。生成式智能体：人类行为的交互式拟像。载于第36届ACM用户界面软件与技术研讨会论文集，2023年用户界面软件与技术研讨会，美国加利福尼亚州旧金山，2023年10月29日- 2023年11月1日，第2:1–2:22页。美国计算机协会。

范秋，刘博义，杨映翔，陈正宇，刘天懿，袁健博，布莱恩·A·普拉默，Zhao-ran Wang，以及杨红霞。2023。让模型用密码说话：通过嵌入进行多智能体辩论。CoRR，abs/2310.06272。

史蒂芬·平克。*2003*。语言本能：心智如何创造语言。英国企鹅出版社。

钱陈，从鑫，杨程，陈伟泽，苏宇声，许巨元，刘知远，以及孙茂松。2023。用于软件开发的通信智能体。CoRR，abs/2307.07924。

诺亚·辛，贝克·拉巴什，以及阿什温·戈皮纳特。2023。Reflexion：一种具有动态记忆和自我反思能力的自主智能体。CoRR，abs/2303.11366。

Significant Gravitas. AutoGPT。

阿罗希·斯里瓦斯塔瓦, 阿宾纳夫·拉斯托吉, 阿比舍克·劳, 阿布·阿瓦尔·穆罕默德·肖布, 阿布巴卡尔·阿比德, 亚当·费施, 亚当·R·布朗, 亚当·桑托罗, 阿迪提亚·古普塔, 阿德里亚·加里加-阿尔松索, 阿格涅什卡·克鲁斯卡, 艾托·莱乌科维奇, 阿克沙特·阿加瓦尔, AletheaPower, 亚历克斯·雷, 亚历克斯·沃斯塔特, 亚历山大·W·科库雷克, 阿里·萨法亚, 阿里·塔扎尔夫, 项爱丽丝, Alicia Par- rish, 倪艾伦, 阿曼·侯赛因, 阿曼达·阿斯基尔, 阿曼达·德索扎, 阿穆特·拉哈内, 阿南塔拉曼·S·艾耶, 安德斯·安德烈森, 安德烈亚·桑蒂利, 安德烈亚斯·施特洪米勒, 安德鲁 ·M· 戴 , 安德鲁 · 拉 , 安德鲁 ·K· 兰品 , 邹安迪 , 江安琪拉, 陈安琪莉卡, 阮英, 阿尼梅什·古普塔, 安娜·戈塔迪, Anto- nio Norelli, 阿努·温卡特什, 阿拉什·戈哈利达武迪, 阿尔法·塔巴苏姆, 阿鲁尔·梅内泽斯, 阿伦·基鲁巴拉詹,

Asher Mullokandov, Ashish Sabharwal, Austin Her- rick, Avia Efrat, Aykut Erdem, Ayla Karakas, and 等 2022. 超越模仿游戏：量化和预测语言模型的能力。CoRR, abs/2206.04615. Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. 思维链提示词在大语言模型中激发推理能力。In NeurIPS. Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. 构建跨文档多跳阅读理解的数据集。计算语言学协会事务, 6:287–302.

Benjamin Lee Whorf. 2012. 语言、思维与现实：*Benjamin Lee Whorf* 精选著作。麻省理工学院出版社。

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Autogen：通过多智能体对话框架启用下一代大语言模型应用。CoRR, abs/2308.08155。

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Ben- gio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa：用于多样化、可解释的多跳问答的数据集。在 2018 年自然语言处理实证方法会议论文集，布鲁塞尔，比利时，2018 年 10 月 31 日至 11 月 4 日，第 2369–2380 页。计算语言学协会。

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak S hafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. 思维树：用大语言模型进行深思熟虑的问题解决。CoRR, abs/2305.10601.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023b. React：在语言模型中协同推理与行动。In The Eleventh International Conference on Learning Representations, ICLR 2023, 基加利, 卢旺达, May 1-5, 2023. OpenReview.net.

Shuyan Zhou，Frank F. Xu，Hao Zhu，Xuhui Zhou，Robert Lo，Abishek Sridhar，Xianyi Cheng，Yonatan Bisk，Daniel Fried，Uri Alon，and Graham Neubig. 2023. WebArena：构建自主智能体的真实网络环境。CoRR, abs/2307.13854.

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

A Experimental Settings

In this section, we introduce the details of our experimental settings.

A.1 Models

For OpenAI's models, we use gpt-3.5-turbo-1106 and gpt-4-1106-preview. For Gemini pro, we use the Gemini pro 1.0 in 2024.1.

A.2 Dataset Pre-Processing

The statistics of the processed data are presented at Table 5. We now elaborate the dataset download and pre-process process.

Single-LLM Reasoning. For Logic Grid, Information Essentiality and Minute Mysteries QA that are from Big-Bench, we download the dataset from the official repo[1]. For Coin Flip, we download the dataset from https://huggingface.co/datasets/skrishna/coin_flip, and use the first 500 examples in the test set. For AQuA, we download the dataset from https://huggingface.co/datasets/aqua_rat/ and use its test set.

Multi-Agent Communication. For the Hotpot QA dataset, we adhere to the methodology outlined by Reflexion (Shinn et al., 2023), obtaining the dataset from their repository [2]. In the case of Wiki Hop, we acquire it through the Huggingface Datasets platform, from which we randomly selected 100 examples from its validation set for our study. For the Narrative QA dataset, also sourced from Huggingface Datasets, we note inconsistencies in the quality of the e-books included. To ensure higher data quality, we exclusively utilize e-books from Project Gutenberg by checking whether the e-book starts with "Project Gutenberg's". Furthermore, considering the context length limitation of 16k tokens in GPT-3.5, we exclude e-books exceeding 30k tokens. This exclusion is to enable splitting the content into two segments, each fitting within the GPT-3.5 context limit. From this refined dataset, we randomly chose 100 examples for analysis.

A.3 Metrics

For single-LLM reasoning, we report the accuracy by comparing the generated answer with the label. For multi-agent communication, we report the RougeL score of the generated answer. The RougeL is calculated using Google's implementation[3].

[1] https://github.com/google/BIG-bench

[2] https://github.com/noahshinn/reflexion/tree/main/hotpotqa_runs/data

B Prompts

We list the prompts we used in this work in Tables 9 to 14.

C Additional Experimental Results for Multi-Agent Communication

Due to constraints on the paper length, the comprehensive experimental results, including those where GPT-3.5 serves as the initiating agent, are detailed in Table 8. Our analysis reveals that the performance of AutoForm tends to be suboptimal when compared to the baseline in scenarios initiated by GPT-3.5. A closer examination of these interactions indicates that GPT-3.5, when merely prompted to employ non-NL formats without additional guidance, frequently produces overly succinct responses, or simply gives a hallucinated answer, resulting in diminished performance. This observation underscores the need for further research into optimizing prompts for less advanced LLMs to effectively utilize non-NL formats for communication, representing a promising avenue for future exploration.

Additionally, we conduct experiments with a specific setting for HotpotQA to evaluate the performance of LLMs when supporting facts are distributed between two agents. This setting, termed *separate context*, ensures that the supporting facts are divided and distributed between the agents, necessitating effective communication to aggregate the information and derive the correct answer.

The results presented in Table 7 reveal significant insights into the performance of LLMs under the *separate context* setting. The application of the AutoForm mechanism shows a distinct impact on both models. For GPT-3.5, the RougeL score slightly decreases from 0.62 to 0.53, with a corresponding 22.4% reduction in the number of tokens generated. In contrast, the GPT-4 models exhibit an increase in performance with the AutoForm mechanism, achieving a RougeL score of 0.69 and a 33.8% reduction in token usage. This improvement highlights the effectiveness of AutoForm in not only maintaining but enhancing the performance of

[3] https://github.com/google-research/google-research/tree/master/rouge

A 实验设置

在本节中，我们介绍了实验设置的详细情况。

A.1 模型

对于 OpenAI 的模型，我们使用 gpt-3.5-turbo-1106 和 gpt-4-1106-preview。对于 Gemini pro，我们使用 2024 年 1 月的 Gemini pro 1.0。

A.2 数据预处理

处理后数据的统计数据展示在表格 5 中。下面，我们详细介绍数据集的下载和预处理流程。

单大语言模型推理。对于来自大基准测试 (Big-Bench) 的逻辑网格、信息重要性和 Minute Mysteries 问答，我们从官方仓库[1]下载数据集。对于抛硬币，我们从 https://huggingface.co/ datasets/skrishna/coin_flip 下载数据集，并使用测试集中的前 500 个示例。对于 AQuA，我们从 https://huggingface.co/ datasets/aqua_rat/ 下载数据集，并使用其测试集。

多智能体通信。对于 Hotpot QA 数据集，我们遵循 Reflexion（Shinn 等人，2023）提出的方法，从其仓库 [2] 获取该数据集。对于 Wiki Hop，我们通过 Huggingface 数据集平台获取，并随机从其验证集中选取了 100 个示例用于本研究。对于叙事问答数据集，同样来源于 Huggingface 数据集，我们注意到所包含电子书的质量存在不一致性。为确保更高的数据质量，我们通过检查电子书是否以“Project Gutenberg's”开头，仅使用古腾堡计划提供的电子书。此外，考虑到 GPT-3.5 的 16k 令牌上下文限制，我们排除了超过 30k 令牌的电子书。这一排除是为了能够将内容分割成两个片段，每个片段都能适配 GPT-3.5 的上下文限制。从经过精炼的数据集中，我们随机选取了 100 个示例用于分析。

A.3 指标

对于单大语言模型推理，我们通过将生成的答案与标签进行比较来报告准确率。对于多智能体通信，我们报告

[1] https://github.com/google/BIG-bench[2] https://github.com/noahshinn/reflexion/tree/ main/hotpotqa_runs/数据

生成的答案的RougeL分数。RougeL使用谷歌的实现来计算[3]。

B 提示词

我们在此工作中使用的提示词列举于表格9至14。

C 多智能体通信的补充实验结果

由于论文长度的约束，全面的实验结果，包括 GPT-3.5 作为启动智能体的那些结果，详见表格8。我们的分析表明，在由 GPT-3.5 启动的情景中，与基线相比，AutoForm 的性能往往不尽如人意。对这些交互的进一步考察表明，当仅提示 GPT-3.5 使用非自然语言格式而没有额外指导时，它经常生成过于简略的响应，或者仅仅给出一个幻觉答案，从而导致性能下降。这一观察结果强调了对优化较不先进的大语言模型提示词的研究需求，使其能够有效地利用非自然语言格式进行通信，这代表着一个很有希望的未来探索方向。

此外，我们在 HotpotQA 的特定设置下进行实验，以评估支持事实分布在两个智能体之间时大语言模型的性能。该设置称为 separatecontext，确保支持事实被分割并分布在智能体之间，因此需要有效的通信来汇总信息并得出正确答案。

表7中展示的{v0}结果揭示了{v0}大语言模型在{v0}分离上下文设置下性能的显著洞见。{v0}AutoForm机制的应用对两个{v0}模型均产生了明显影响。对于GPT-3.5，{v0}RougeL分数从0.62略微下降至0.53，同时生成的{v0}令牌数量相应减少了22.4%。相比之下，GPT-4 {v0}模型在{v0}AutoForm机制下表现出性能提升，达到0.69的{v0}RougeL分数，且{v0}令牌使用减少了33.8%。这一改进突显了{v0}AutoForm不仅能够维持，还能提升性能方面的有效性

[3] https://github.com/google-research/google-research/tree/master/rouge

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

Dataset	# Examples	Input	Output	Category	License
Single-LLM Reasoning					
Logic Grid Puzzle	200	Clues + Question	Number	Logical Reasoning	Apache License 2.0
Information Essentiality	68	Question + Statement Options	Option Number	Logical Reasoni	Apache License 2.0
AQuA	254	Mathematical Question + Options	Option Number	Mathematical Reasoning	Apache License 2.0
Minute Mysteries QA	203	Story + Question + Options	Option Number	Causal Reasoning	Apache License 2.0
Coin Flip	500	Action Sequence	Yes / No	Symbolic Reasoning	MIT License
Multi-LLM Communication					
HotPot QA	100	Passages + Question	Free Text	Multi-Hop QA	CC BY-SA 4.0
Wiki Hop	100	Sentences + Question	Free Text	Multi-Hop QA	CC BY-SA 3.0
Narrative QA	100	Book + Question	Free Text	QA	Apache License 2.0

Table 5: The datasets we use in our experiments.

Dataset	Model	Math Equation	Unordered List	Ordered List	Markdown Table	Multi-level List	Logical Expression
Logic Grid	GPT-3.5	**50.0**	47.0	46.0	48.5	45.0	41.0
	GPT-4	58.0	**65.0**	59.0	55.0	62.0	59.0
	Gemini Pro	42.0	**50.5**	48.0	49.0	48.0	49.0
Coin Flip	GPT-3.5	71.0	39.0	**86.0**	66.0	19.2	48.0
	GPT-4	75.0	95.0	**100.0**	99.0	98.0	71.4
	Gemini Pro	47.0	61.6	60.4	**63.4**	56.4	59.6
Info Essen	GPT-3.5	22.0	**30.8**	**30.8**	29.4	**30.8**	**30.8**
	GPT-4	73.5	73.5	**76.4**	73.5	**76.4**	75.0
	Gemini Pro	**45.6**	29.4	32.4	29.4	44.1	38.2
MM QA	GPT-3.5	22.2	22.2	23.2	**27.1**	22.7	20.2
	GPT-4	39.9	41.9	37.7	36.9	**42.4**	37.9
	Gemini Pro	**29.6**	27.6	25.6	27.6	28.1	24.1
AQuA	GPT-3.5	63.4	62.2	63.0	56.7	59.8	**66.0**
	GPT-4	76.4	72.4	**80.3**	78.0	78.0	79.8
	Gemini Pro	56.7	57.5	**59.8**	47.2	55.5	56.3

Table 6: Effectiveness of various formats across different models and datasets.

more advanced models while optimizing resource utilization.

D Examples

In this section, we provide several response examples for each LLM on each task, so that readers can have a more intuitive understanding of the effects after using AutoForm. Tables 71 to 77 present the examples of using AutoForm, which corresponds to the results in Table 2. The examples for two-step AutoForm, which corresponds to Table 3, are presented in Tables 15 to 59. For the multi-agent communication experiment, we present the examples in Tables 60 to 70.

数据集	# 示例	输入	输出	类别	许可证
单大语言模型推理					
逻辑网格谜题	200	线索 +问题	数字	逻辑推理	Apache 2.0 许可证
信息必要性	68	问题 + 陈述 选项	选项编号	逻辑推理	Apache 2.0 许可证
AQuA	254	数学问题 +选项	选项编号	数学推理	Apache 2.0 许可证
Minute Mysteries 问答	203	故事 + 问题 +选项	选项编号	因果推理	Apache 2.0 许可证
抛硬币	500	动作序列	是 / 否	符号推理	MIT 许可协议
多大型语言模型通信					
Hotpot QA	100	篇章 +问题	自由文本	多跳问答	CC BY-SA 4.0 许可协议
Wiki Hop	100	句子 +问题	自由文本	多跳问答	CC BY-SA 3.0
叙事问答	100	书籍 +问题	自由文本	QA	Apache 2.0 许可证

表格 5：我们在实验中使用的数据集。

数据集	模型	Math 方程	无序 List	有序 List	Markdown 表格	多级 List	逻辑 表达式
逻辑 Grid	GPT-3.5	**50.0**	47.0	46.0	48.5	45.0	41.0
	GPT-4	58.0	**65.0**	59.0	55.0	62.0	59.0
	Gemini Pro	42.0	**50.5**	48.0	49.0	48.0	49.0
Coin Flip	GPT-3.5	71.0	39.0	**86.0**	66.0	19.2	48.0
	GPT-4	75.0	95.0	**100.0**	99.0	98.0	71.4
	Gemini Pro	47.0	61.6	60.4	**63.4**	56.4	59.6
Info 埃森	GPT-3.5	22.0	**30.8**	**30.8**	29.4	**30.8**	**30.8**
	GPT-4	73.5	73.5	**76.4**	73.5	**76.4**	75.0
	Gemini Pro	**45.6**	29.4	32.4	29.4	44.1	38.2
MM QA	GPT-3.5	22.2	22.2	23.2	**27.1**	22.7	20.2
	GPT-4	39.9	41.9	37.7	36.9	**42.4**	37.9
	Gemini Pro	**29.6**	27.6	25.6	27.6	28.1	24.1
AQuA	GPT-3.5	63.4	62.2	63.0	56.7	59.8	**66.0**
	GPT-4	76.4	72.4	**80.3**	78.0	78.0	79.8
	Gemini Pro	56.7	57.5	**59.8**	47.2	55.5	56.3

表6：不同格式在不同模型和数据集上的有效性。

更为先进的{v0}模型，同时优化{v0}资源利用。

D 示例

在本节中，我们为每个任务中的每个大语言模型提供若干响应示例，以便读者能够更直观地了解使用AutoForm后的效果。表71至77展示了使用AutoForm的示例，这些示例对应于表2中的结果。对应于表3的两步AutoForm示例展示在表15至59中。对于多智能体通信实验，我们在表60至70中展示了相关示例。

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

	Hotpot QA$_{\text{separate context}}$		
Model	**RougeL**	**# Tokens**	**ΔTokens**
GPT-3.5 + GPT-3.5	**0.62**	369.3	-
+*AutoForm*	0.53	286.6	-22.4%
GPT-4 + GPT-4	0.68	151.0	-
+*AutoForm*	**0.69**	100.0	-33.8%

Table 7: Performance on HotpotQA under the *separate context* setting. The supporting facts are distributed between two agents, necessitating inter-agent communication for aggregating information and deriving the correct answer.

	Hotpot QA$_{\text{separate context}}$		
模型	**RougeL**	**词元数**	**Δ令牌**
GPT-3.5 + GPT-3.5	**0.62**	369.3	-
+*AutoForm*	0.53	286.6	-22.4%
GPT-4 + GPT-4	0.68	151.0	-
+*AutoForm*	**0.69**	100.0	-33.8%

表 7：分离上下文设置下 HotpotQA 的性能。支持事实分布在两个智能体之间，因此需要智能体间通信来进行信息聚合并得出正确答案。

	Wiki Hop			Hotpot QA			Narrative QA		
Model	**RougeL**	**# Tokens**	**ΔTokens**	**RougeL**	**# Tokens**	**ΔTokens**	**RougeL**	**# Tokens**	**ΔTokens**
GPT-3.5 + GPT-3.5	**0.57**	192.6	-	**0.53**	499.7	-	**0.34**	140.0	-
+*AutoForm*	0.49	163.9	-14.9%	0.47	236.1	-52.8%	0.33	35.5	-74.6%
GPT-3.5 + GPT-4	0.56	246.8	-	**0.72**	333.9	-	**0.37**	208.8	-
+*AutoForm*	**0.57**	200.3	-18.8%	0.62	102.3	-69.4%	0.30	125.4	-39.9%
GPT-4 + GPT-3.5	**0.53**	281.5	-	0.63	345.5	-	0.43	178.3	-
+*AutoForm*	**0.53**	255.0	-9.4%	**0.70**	94.3	-72.7%	**0.48**	119.4	-33.0%
GPT-4 + GPT-4	0.50	237.5	-	0.69	145.2	-	**0.43**	240.7	-
+*AutoForm*	**0.52**	146.2	-38.4%	**0.76**	115.0	-20.8%	**0.43**	141.7	-41.1%

Table 8: Comparative performance in multi-agent communication across various QA datasets. The table highlights RougeL scores, with better performance in different model pairing settings indicated in bold. The ΔTokens column quantifies the token reduction achieved by the AutoForm method.

	Wiki Hop			Hotpot QA			叙事问答		
模型	**RougeL**	**词元数**	**Δ令牌**	**RougeL**	**#Tokens**	**Δ令牌**	**RougeL**	**词元数**	**Δ令牌**
GPT-3.5 + GPT-3.5	**0.57**	192.6	-	**0.53**	499.7	-	**0.34**	140.0	-
+*AutoForm*	0.49	163.9	-14.9%	0.47	236.1	-52.8%	0.33	35.5	-74.6%
GPT-3.5 + GPT-4	0.56	246.8	-	**0.72**	333.9	-	**0.37**	208.8	-
+*AutoForm*	**0.57**	200.3	-18.8%	0.62	102.3	-69.4%	0.30	125.4	-39.9%
GPT-4 +GPT-3.5	**0.53**	281.5	-	0.63	345.5	-	0.43	178.3	-
+*AutoForm*	**0.53**	255.0	-9.4%	**0.70**	94.3	-72.7%	**0.48**	119.4	-33.0%
GPT-4 +GPT-4	0.50	237.5	-	0.69	145.2	-	**0.43**	240.7	-
+*AutoForm*	**0.52**	146.2	-38.4%	**0.76**	115.0	-20.8%	**0.43**	141.7	-41.1%

表格 8：在不同问答数据集上的多智能体通信比较性能。该表格突出显示了RougeL 分数，以粗体表示在不同模型配对设置下的更优性能。 ΔTokens列量化了AutoForm方法实现的令牌减少。

PROMPT FOR COIN FLIP

CoT:
Question:
${task_description}

At the end of your response, you must give your answer in the form of "the answer is: no" or "the answer is: yes". Let's think step-by-step.

AutoForm:
Question:
${task_description}

To enhance clarity and eliminate ambiguities inherent in natural language, consider employing more structured and concise forms of communication for your step-by-step solutions. Suitable formats include code, pseudocode, JSON, markdown tables, logical operators, or mathematical equations.

At the end of your response, you must give your answer in the form of "the answer is: no" or "the answer is: yes". Remember to be concise and accurate.

Table 9: Prompt for Coin Flip

抛硬币提示

思维链：问题：$
{task_description}

在你的响应的末尾你必须以“答案是：否”或“答案是：是”的形式给出你的答案。让我们逐步思考。
逐步

AutoForm:
问题:
${task_描述}

为增强清晰度并消除自然语言中固有的歧义，建议在逐步解决方案中采用更结构化和简洁的通信形式。适用的格式包括代码、伪代码、JSON、Markdown 表格、逻辑运算符或数学方程。

在y的末尾 您的响应，您必须以“答案是：否”或“答案是：是”的形式给出您的答案。请记住
要简洁 准确的。

表 9：抛硬币提示

13

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

PROMPT FOR LOGIC GRID

CoT

—
${task_description}
—

At the end of your response, you must give your answer in the form of "the answer is: {number}", where {number} is the answer number. Now solve the problem step-by-step. Use as few words as possible.

AutoForm

—
${task_description}
—

To enhance clarity and eliminate ambiguities inherent in natural language, consider employing more structured and concise forms of communication for your step-by-step solutions. Suitable formats include code, pseudocode, JSON, markdown tables, logical operators, or mathematical equations.

At the end of your response, you must give your answer in the form of "the answer is: {number}", where {number} is the answer number. Remember to be concise and accurate.

Table 10: Prompt for Logic Grid

逻辑网格提示

CoT

—
${task_description}
—

在你的响应末尾，你必须以“答案是：{number}”的形式给出你的答案，其中 {number} 是答案的数字。现在请逐步解决这个问题。尽可能少用词。

AutoForm

— ${task_
description} —

为提升清晰度并消除自然语言固有的歧义，建议在提供逐步解决方案时，采用更结构化且简洁的通信格式。适用的格式包括代码、伪代码、JSON、Markdown 表格、逻辑运算符或数学方程。

在您的响应末尾，必须以“答案是：{number}”的形式给出答案，其中{number}为答案数字。请保持简洁且准确。

表10：逻辑网格提示

PROMPT FOR MINUTE MYSTERIES QA

CoT:

—
${task_description}
—

Now solve the problem step-by-step. At the end of your response, you must give your answer in the form of "the correct option is: number", where number is the index of the chosen option.

AutoForm:

—
${task_description}
—

To enhance clarity and eliminate ambiguities inherent in natural language, consider employing more structured and concise forms of communication for your step-by-step solutions. Suitable formats include code, pseudocode, JSON, markdown tables, logical operators, or mathematical equations.

Now solve the problem step-by-step. At the end of your response, you must give your answer in the form of "the correct option is: number", where number is the index of the chosen option.

Table 11: Prompt for Minute Mysteries QA

一分钟神秘谜题问答提示

CoT:

—
${task_description}
—

现在逐步解决该问题。在响应的末尾，你必须以“正确选项是：数字”的形式给出答案，其中“数字”为所选选项的索引。

AutoForm:

— ${task_
description} —

为了增强清晰度并消除自然语言固有的歧义，建议在您的分步解决方案中采用更结构化、更简洁的通信形式。合适的格式包括代码、伪代码、JSON、Markdown 表格、逻辑运算符或数学方程。

现在逐步解决该问题。在你的响应末尾，你必须以“正确选项为：数字”的形式给出你的答案，其中数字是所选选项的索引。

表 11：Minute Mysteries 问答提示

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

PROMPT FOR AQUA

CoT:

Solve the problem presented below:

—

${task_description}

—

RESPONSE GUIDELINES:

1. Think step by step.
2. Concluding with the Answer: End your response with "Answer: {answer}", where {answer} is the final result of your problem-solving process. The {answer} should be a single capital letter.

AutoForm:

Solve the problem presented below:

—

${task_description}

—

RESPONSE GUIDELINES:

1. Initial State Representation: Begin by providing a clear and detailed representation of the initial state or conditions of the problem.
2. Step-by-Step Solution Process: Progressively update the state representation as you work through each step of the solution. This should include all logical reasoning and calculations leading to the final answer.
3. Concluding with the Answer: End your response with "Answer: {answer}", where {answer} is the final result of your problem-solving process. The {answer} should be a single capital letter.

Table 12: Prompt for AQuA

面向AQuA的提示词

CoT:

解决下面提出的问题：

—

${task_描述}

—

响应准则：

1. 逐步思考。
2. 以答案结尾：请以“答案：{answer}”结束您的响应，其中{answer}是您问题解决过程的最终结果。{answer}应为一个单个大写字母。

AutoForm:

解决下面提出的问题：

—

${task_description}

—

响应准则：.

1 初始状态表示：首先，提供对问题初始状态或条件的清晰、详细的表示。2. 逐步求解过程：在逐步执行解决方案的每一步时，逐步更新状态表示。这应包含所有导致最终答案的逻辑推理和计算。3. 以答案结尾：在响应末尾以“答案：{answer}”结束，其中{answer}是你问题解决过程的最终结果。{answer}应为一个单个大写字母。

表12：用于 AQuA 的提示词

15

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

PROMPT FOR INFORMATION ESSENTIALITY

3.5+CoT:
Solve the problem presented below:
—
${task_description}
—

RESPONSE GUIDELINES:
1. Think step by step.
2. Your answer should be ended with "Answer: {answer}" where {answer} is the answer to the problem.

3.5+AutoForm:
Solve the problem presented below:
—
${task_description}
—

RESPONSE GUIDELINES:
1. Initial State Representation: Begin by providing a clear and detailed representation of the initial state or conditions of the problem.
2. Step-by-Step Solution Process:Progressively update the state representation as you work through each step of the solution. This should include all logical reasoning and calculations leading to the final answer.
3. To enhance clarity and eliminate ambiguities inherent in natural language, consider employing more structured and concise forms of communication for your step-by-step solutions. Suitable formats include code, pseudocode, JSON, markdown tables, logical operators, or mathematical equations.
4. Concluding with the Answer: End your response with "Answer: {answer}", where {answer} is the final result of your problem-solving process.

4+CoT:
Solve the problem presented below:
—
${task_description}
—

RESPONSE GUIDELINES:
1.You should think step by step.
2. You should consider three scenarios: using only Statement 1, using only Statement 2, and using both Statements.
3. Note (IMPORTANT): When considering Statement 1, the use of information from Statement 2 is prohibited. When considering Statement 2, the use of information from Statement 1 and the analysis derived from Statement 1 is prohibited. Both conditions can only be analyzed simultaneously during the stage where both Statements are considered together.
4. In sometime , both statement 1 and statement 2 can lead to answer alone.
5. Concluding with the Answer: End your response with "Answer: {answer}", where {answer} is the final result of your problem-solving process.

4+AutoForm:
Solve the problem presented below:
—
${task_description}
—

RESPONSE GUIDELINES:
1. Initial State Representation: Begin by providing a clear and detailed representation of the initial state or conditions of the problem.
2. Step-by-Step Solution Process:Progressively update the state representation as you work through each step of the solution. This should include all logical reasoning and calculations leading to the final answer.
3. To enhance clarity and eliminate ambiguities inherent in natural language, consider employing more structured and concise forms of communication for your step-by-step solutions. Suitable formats include code, pseudocode, JSON, markdown tables, logical operators, mathematical equations and so on.
4. You should consider three scenarios: using only Statement 1, using only Statement 2, and using both Statements.
5. Note (IMPORTANT): When considering Statement 1, the use of information from Statement 2 is prohibited. When considering Statement 2, the use of information from Statement 1 and the analysis derived from Statement 1 is prohibited. Both conditions can only be analyzed simultaneously during the stage where both Statements are considered together.
6. In sometime , both statement 1 and statement 2 can lead to answer alone.
7. Concluding with the Answer: End your response with "Answer: {answer}", where {answer} is the final result of your problem-solving process.

Table 13: prompt for Information Essentiality

信息必要性提示词

3.5+思维链：解决以下问题：—
${task_description} —

响应准则：
1. 分步骤思考。
2. 你的答案应以“Answer: {answer}”结尾，其中 {answer} 是对该问题的答案。

3.5+AutoForm：
解决以下问题：
—
${task_description}
—

响应准则: .
1 初始状态表示：通过提供问题的初始状态或条件的清晰且详细的描述开始。
2. 逐步求解过程：在逐步解决每个问题时，逐步更新状态表示。这应包括所有导致最终答案的逻辑推理和计算。
3. 为了提高清晰度并消除自然语言固有的歧义，考虑采用更结构化和简明的通信方式用于分步解决方案。合适的格式包括代码、伪代码、JSON、Markdown 表格、逻辑运算符或数学方程。
4. 以答案结尾：以"Answer: {answer}"作为你的响应的结尾，其中 {answer} 是你问题解决过程的最终结果。

4+思维链：
解决以下问题：
—
${task_描述}
—
响应准则： .
1 你应该分步骤思考
2. You
3. 注意（重要）：在考虑陈述1时，禁止使用陈述2的信息。在考虑陈述2时，禁止使用陈述1的信息以及从陈述1中得出的分析。仅在同时考虑两个陈述的阶段，才能同时分析这两个条件。 4. I
n 在一些time , both statement 1 and statement 2 can lead to 答案 a lone.
你应该考虑三种情景：仅使用陈述1、仅使用陈述2，以及同时使用两个陈述。4. 有时，陈述1和陈述2可以单独得出答案。.

5 以答案结尾：在您的响应末尾以“Answer: {answer}”结尾，其中{answer}是您问题解决过程的最终结果。

4+AutoForm:
解决下面提出的问题：
—
${task_description}
—

响应准则：。
1 初始状态表示：首先提供问题初始状态或条件的清晰且详细的表示。
2 逐步求解过程：在逐步解决每个解决方案步骤时，逐步更新状态表示 .
这应包括所有 g理性推理 g 且计算导向 g 最终答案。
3. 为了增强清晰度并消除自然语言固有的歧义，请考虑在分步解决方案中使用更结构化和简洁的通信形式。适用的格式包括代码、伪代码、JSON、Markdown 表格、逻辑运算符、数学方程等。
4. Yo u 应考虑三种情景：仅使用陈述1、仅使用陈述2，以及同时使用两个陈述 .
5. 注意（重要）：在考虑陈述1时，禁止使用陈述2中的信息。在考虑陈述2时，禁止使用陈述1中的信息以及从陈述1得出的分析。仅在同时考虑两个陈述的阶段，才能同时分析这两个条件。 6. I
n 在一些time , both statement 1 and statement 2 can lead to 答案 a lone.
7. 以答案结尾：以“Answer: {answer}”结尾，其中 {answer} 是您问题解决过程的最终结果。

表13：信息必要性提示词

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

PROMPT FOR HOTPOT QA, WIKI HOP, NARRATIVE QA

Shared portion of the prompt

You are ${agent_name}. Together with ${all_roles}, you are providing accurate answer to the user. Each of you will be provided parts of the contexts and a shared question.
EXAMPLE 1
—
Context
${example_context_1}
Question
${example_question_1}
—
${example_answer_1}.
EXAMPLE 2
—
Context
${example_context_2}
Question
${example_question_2}
—
${example_answer_2}.

Now the user gives you some contexts and the question:
—
Context
${knowledge}
Question
${task_description}
—

Baseline:

Given that each individual, including yourself, possesses unique contexts, it's essential to actively share and discuss this information with others to formulate a complete answer. Your specific context is unknown to others unless explicitly communicated. This collaborative effort is key to reaching an accurate answer based on the amalgamation of everyone's distinct contexts.

When you have reached the final answer, conclude it with "<A>xxx</A>", where "xxx" will be extracted and compared with ground truth. To end the conversation, all the players should end their responses with "<A>xxx</A>".
You are ${agent_name}. Now communicate with ${all_roles} to give the answer.

AutoForm(3.5-3.5,3.5-4,4-3.5,4-4):

Given that each individual, including yourself, possesses unique contexts, your specific context is unknown to others unless explicitly communicated.
You are ${agent_name}, collaborating with ${all_roles}, who are also intelligent assistants. Your goal is to provide a clear and concise answer to the user's question. Unlike typical communication, you will not use natural language, as it often contains ambiguities and emotional nuances. Instead, choose a more straightforward and precise communication medium, such as structured data, JSON, XML or code.
Now, start communicating with ${all_roles} using your selected non-natural language medium. Remember, clarity and brevity are key.
Once you have formulated the final answer,you must enclose it within "<A>xxx</A>", where "xxx" represents the answer phrase selected from the given choices. The conversation concludes when all participants have presented the same answer in this format. If you have different opinion, explain it to your teammates.
Don't forget to enclose your answer within "<A>xxx</A>"

Table 14: prompt for Hotpot QA, Wiki Hop, Narrative QA

HOTPO QA、WIKI HOP 和 NARRATIV 的提示词 E QA

提示词 Y 的共享部分

你是 ${agent 姓名_}. Togeth拥有 ${all_roles} 的角色，你正在为用户提供准确的答案。你们中的每一位都将获得部分上下文 以及一个共享的问题。
示例 1
—
背景
${example_background_1}
问题
${example_问题_1}
—
${example_答案_1}
示例 2
—
背景
${example_背景_2}
问题
${example_问题_2}
—
${example_答案_2}

现在用户向你提供一些上下文和问题 n:
—
背景 $
{knowledge} # 问
题 ${task_描述} —

基线:

鉴于每个人（包括你自己）都拥有独特的上下文，因此主动与他人分享并讨论这些信息以制定完整的最终答案至关重要。除非明确传达，否则他人并不知道你的具体上下文。这种协作努力是达成基于所有人独特上下文综合的准确最终答案的关键。

当你得出最终答案时，请以"<A>xxx</A>"作为结尾，其中"xxx"将被提取并与基准真值进行比较。为了结束对话，所有参与者都应以"<A>xxx</A>"作为其响应的结尾。你是${agent_姓名}。现在与${all_角色}沟通以给出答案。

AutoForm(3.5-3.5,3.5-4,4-3.5,4-4):

已知条件：包括你在内的每个人拥有独特的上下文，因此，除非明确传达，否则你的具体上下文对其他人来说未知。

你是${agent_name}，正在与${all_roles}协作，他们也是智能助手。你的目标是为用户提供清晰简洁的答案。与典型的通信不同，你不会使用自然语言，因为它通常包含歧义和情感细微差别。相反，请选择一种更直接、更精确的通信媒介，例如结构化数据、JSON、XML或代码。
现在，使用你选择的非自然语言媒介开始与${all_roles}进行通信。记住，清晰度和简洁性是关键的。

一旦你制定了最终答案，你必须将其包裹在“<A>xxx</A>”中，其中“xxx”代表从已知条件给出的选项中选择的答案短语。当所有参与者都以这种格式呈现出相同的答案时，对话即告结束。如果你有不同的意见，请向你的队友解释。别忘了将你的答案包裹在“<A>xxx</A>”中

表14：用于Hotpot QA、Wiki Hop、叙事问答的提示词

17

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

First Stage (generated by GPT-3.5):
I will use a table to organize the information and clues for each problem. This will help me keep track of the different characteristics and their corresponding houses, making it easier to eliminate possibilities and solve the problem accurately.

Second Stage (generated by GPT-4):
Let's solve the problem step by step using the clues provided:
1. The person who is wearing running shoes lives somewhere to the right of the person who is eating oranges.
Since there are only two houses, if the person wearing running shoes is to the right of the person eating oranges, the person eating oranges must be in house 1, and the person wearing running shoes must be in house 2.
2. The mystery book reader lives directly left of the cat lover.
Since the mystery book reader cannot live to the right of anyone (there are only two houses), the mystery book reader must be in house 1, and the cat lover must be in house 2.
3. The mystery book reader and the person who had stew live directly next to each other.
Since the mystery book reader is in house 1, the person who had stew must also be in house 1, as they live next to each other and there are only two houses.
With these clues, we can determine that the person who had stew lives in house 1.
The answer is: 1

Table 15: Examples of Logic Grid (GPT-3.5 ⇒ GPT-4 Task-Based)

第一阶段（由 GPT-3.5 生成）：
I will use a table to organize the information and clues for each 问题. This will help me keep track of the different 特征 and their corresponding 房屋, making it easier to eliminate 可能性 and solve the 问题 accurately.

第二阶段（由 GPT-4 生成）：
Let's solve the problem step by step using the clues provided:
1. 穿跑鞋的人住在吃橙子的人的右边。由于只有两栋房子，如果穿跑鞋的人住在吃橙子的人的右边，那么吃橙子的人必须在1号房屋，而穿跑鞋的人必须在2号房屋。2. 神秘小说读者住在爱猫者的正左边。由于神秘小说读者不能住在任何人的右边（只有两栋房子），神秘小说读者必须在1号房屋，而爱猫者必须在2号房屋。3. 神秘小说读者和吃了炖菜的人直接相邻。由于神秘小说读者在1号房屋，吃了炖菜的人也必须在1号房屋，因为他们相邻且只有两栋房子。根据这些线索，我们可以确定吃了炖菜的人住在1号房屋。答案是：1

表15：逻辑网格示例（GPT-3.5 ⇒ GPT-4任务型）

First Stage:
I will use a table to organize the information and solve the problem:

House	Fruit	Pet	Sport	Book	Shoes
1	Kiwis	?	?	?	?
2	Apples	?	?	Science Fiction	?

Using the given clues, I will fill in the table to deduce the location of the cricketer. This structured approach will help ensure that all the information is considered and that the solution is accurate.

Second Stage:
First, we know from clue 1 that the person in the first house is eating kiwis. So, we can fill in the fruit for house 1 as "Kiwis".
Next, from clue 4, we know that the science fiction book fanatic lives in the second house. So, we can fill in the book for house 2 as "Science Fiction".
Now, from clue 3, we know that the tennis player and the person eating apples live next to each other. Since the person in the first house is eating kiwis, the person in the second house must be eating apples. So, we can fill in the fruit for house 2 as "Apples" and the sport for house 2 as "Tennis".
From clue 5, we know that the dog owner does not live in the first house. So, the person in the second house must be the dog owner.
Now, from clue 2, we know that the bird keeper and the person wearing high heels live next to each other. Since the person in the second house is the dog owner, the person in the first house must be the bird keeper. Therefore, the person in the first house is wearing high heels.
So, the person in the first house is eating kiwis, is a bird keeper, plays cricket, reads mystery books, and is wearing high heels. The person in the second house is eating apples, is a dog owner, plays tennis, reads science fiction books, and is wearing sandals. Therefore, the cricketer lives in the first house.
The answer is: 1

Table 16: Examples of Logic Grid (GPT-3.5 Instance-Based)

第一阶段：
我将使用表格来整理信息并解决问题：

房子	水果	宠物	运动	书籍	鞋子
1	猕猴桃	?	?	?	?
2	苹果	?	?	科学	?

利用已知的线索，我将填写表格以推断出板球运动员的地点。这种结构化方法有助于确保考虑所有信息，并保证解决方案的准确性。

第二阶段：小说
根据线索1，我们知道第一栋房子的人员正在吃猕猴桃。因此，我们可以将1号房屋的水果填写为“猕猴桃”。接下来，根据线索4，我们知道科幻小说狂热者住在第二间房子。因此，我们可以将第二间房子的书籍填写为“科幻小说”。

现在，根据线索3，我们知道网球运动员和吃苹果的人住在一起。由于第一栋房子的人员正在吃猕猴桃，所以第二间房子的人员必须在吃苹果。因此，我们可以将第二间房子的水果填写为“苹果”，并将第二间房子的运动填写为“网球”。

从线索5，我们知道狗主人不住在第一栋房子。因此，第二间房子的人必须是狗主人。

现在，从线索2，我们知道鸟主人和穿高跟鞋的人住在相邻的位置。由于第二间房子的人是狗主人，第一栋房子的人必须是鸟主人。因此，第一栋房子的人穿着高跟鞋。

因此，第一栋房子里的人正在吃猕猴桃，是鸟主人，打板球，读推理小说，穿着高跟鞋。第二间房子里的人正在吃苹果，是狗主人，打网球，读科幻小说，穿着凉鞋。因此，板球运动员住在第一栋房子里。答案是：1

表格16：逻辑网格（GPT-3.5 基于实例）示例

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

First Stage:
I will use a table to organize the information and clues for each problem. This will help me keep track of the different characteristics and their corresponding houses, making it easier to analyze and eliminate possibilities as I work through the clues. This structured format will enhance the accuracy of my analysis by providing a clear visual representation of the relationships between the different characteristics and their respective houses.

Second Stage:
Step 1: Create a table to organize the information and clues.

House	Favorite Color	Sport	Car	Fruit
1	Yellow			Kiwis
2				
3				
4				

Step 2: Use the clues to fill in the table.
Clue 4: The person who likes yellow lives in the first house.
Clue 13: The person who is eating kiwis lives in the first house.
Clue 3: The soccer player does not live in the fourth house.
Clue 9: The person who drives a convertible lives in the fourth house.
Clue 1: The cricketer lives somewhere to the left of the person who drives a convertible.
Clue 12: The person who owns an SUV and the person who likes blue live directly next to each other.
Clue 2: The person who owns an SUV does not live in the third house.
Clue 8: The person who owns an SUV does not live in the second house.
Clue 10: The person who owns an SUV and the person who is eating a banana live directly next to each other.
Clue 11: The person who drives a minivan lives directly left of the person who likes white.
Clue 6: The swimmer lives somewhere to the left of the person who likes white.
Clue 5: The person who is eating a banana lives directly left of the person who likes white.
Clue 7: The person who is eating mangos lives somewhere to the left of the person who is eating oranges.
Clue 14: There is one house between where the swimmer lives and where the cricketer lives.

House	Favorite Color	Sport	Car	Fruit
1	Yellow	Kiwis		
2				
3				
4			Convertible	

Step 3: Analyze the clues to fill in the remaining information.
From clue 13, the person in house 1 is eating kiwis, so the person in house 2 is eating mangos.
From clue 7, the person in house 3 is eating oranges.
From clue 5, the person in house 2 likes white.
From clue 11, the person in house 1 drives a minivan.
From clue 6, the swimmer lives in house 2.
From clue 14, the cricketer lives in house 3.

House	Favorite Color	Sport	Car	Fruit
1	Yellow	Minivan		Kiwis
2	White	Swimmer		Mangos
3		Cricketer		Oranges
4			Convertible	

Step 4: Use the remaining clues to fill in the information.
From clue 12, the person who likes blue lives in house 3.
From clue 1, the cricketer lives to the left of the person who drives a convertible, so the cricketer is in house 3 and the person who drives a convertible is in house 4.
From clue 10, the person who owns an SUV and the person eating a banana live next to each other, so the person in house 2 owns an SUV and the person in house 3 is eating a banana.
From clue 2, the person who owns an SUV does not live in house 3, so the person in house 2 owns an SUV and the person in house 4 is eating a banana.
From clue 8, the person who owns an SUV does not live in house 2, so the person in house 3 owns an SUV.

House	Favorite Color	Sport	Car	Fruit
1	Yellow	Minivan		Kiwis
2	White	Swimmer	SUV	Mangos
3	Blue	Cricketer		Banana
4			Convertible	Oranges

The answer is: 4

Table 17: Examples of Logic Grid (GPT-3.5 Task-Based)

第一阶段：
我将使用一个表格来组织每个问题的信息和线索。这有助于我跟踪不同的特征及其对应的房屋，使我在分析线索并排除可能性时更加容易。这种结构化格式将通过提供不同特征及其各自房屋之间关系的清晰可视化表示，来提高我分析的准确率。

第二阶段：
第一步：创建一个表格来整理信息和线索。

房子	最喜欢的颜色	运动	汽车	水果
1	黄色			猕猴桃
2				
3				
4				

步骤2：使用线索填写表格。
线索4：喜欢黄色的人员住在第一栋房子。线索13：正在吃猕猴桃的人员住在第一栋房子。
线索3：足球运动员不住在第四栋房子 .
线索9：驾驶敞篷车的人员住在第四栋房子 .
线索1：板球运动员住在驾驶敞篷车的人员的左侧某处。
线索12：拥有运动型多用途车的人员和喜欢蓝色的人员直接住在彼此隔壁。

线索2：拥有运动型多用途车的人员不住在第三栋房子 .
线索8：拥有运动型多用途车的人员不住在第二间房子 .
线索10：拥有运动型多用途车的人员和正在吃香蕉的人员直接住在彼此隔壁
线索11：驾驶小型厢式车的人员正在吃香蕉的人员的直接左边。

线索6：游泳者住在喜欢白色的人员的某处左边。 .
线索5：正在吃香蕉的人员住在喜欢白色的人员的左边。 .
线索14：游泳者居住的地方和板球运动员居住的地方之间有一所房子。
线索7：正在吃芒果的人员住在正在吃橙子的人员的某处左边。

房子	最喜欢的颜色	运动	汽车	水果
1	黄色	猕猴桃		
2				
3				
4			敞篷车	

S步骤3：分析线索以填写剩余信息 ion.
根据线索13、4，房屋的人员正在吃猕猴桃，因此2号房屋的人员正在吃芒果
根据线索7，3号房屋的人员在吃橙子。
根据线索5，2号房屋的人员喜欢白色。
根据线索11，1号房屋的人员驾驶一辆小型厢式车。根据线索6，游泳者住在2号房屋。

根据线索14，板球运动员住在3号房屋 .

房子	最喜欢的颜色	运动	汽车	水果
1	黄色	小型厢式车		猕猴桃
2	白色	游泳者		芒果
3		板球运动员		橙子
4			敞篷车	

步骤 4：使用剩余的线索填入信息。
从线索12可知，喜欢蓝色的人住在3号房屋 .
从线索1，t 该板球运动员住在开敞篷车的人员的左边，因此板球运动员位于3号房屋，且该人员开敞篷车的人在4号房屋 .
从线索10 ，t拥有运动型多用途车的人员和吃香蕉的人员住在隔壁，因此2号房屋中的人员拥有运动型多用途车，而3号房屋中的人员正在吃香蕉。
从线索2可知，拥有运动型多用途车的人员不住在3号房屋，因此2号房屋中的人员拥有运动型多用途车，而4号房屋中的人员正在吃香蕉。

根据线索8，拥有运动型多用途车的人员不住在2号房屋，所以3号房屋中的人员拥有运动型多用途车。

房屋	最喜欢的颜色	运动	汽车	水果
1	黄色	小型厢式车		猕猴桃
2	白色	游泳者	运动型多用途车	芒果
3	蓝色	板球运动员		香蕉
4			敞篷车	橙子

答案是：4

表17：逻辑网格示例（GPT-3.5基于任务）

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

First Stage (generated by GPT-4):
For these logic puzzles, a tabular format can be particularly effective in enhancing clarity and eliminating ambiguities. Tables allow for the organization of information in a grid-like structure, which is conducive to the step-by-step process of elimination and deduction required to solve these types of problems.
Here's how a table can be used to enhance accuracy:
1. Create a table with rows representing the houses and columns representing the characteristics (e.g., book preference, sport, pet, drink, flowers, car, fruit).
2. Fill in the known information from the clues directly into the table.
3. Use process of elimination to fill in the blanks. For example, if the golfer lives in the first house, then no other person can be the golfer or live in the first house.
4. Mark impossibilities based on the clues. For instance, if the swimmer cannot live in the third house, then mark that cell as impossible.
5. Continue to cross-reference clues and fill in or eliminate possibilities until the table is complete.
The table format allows for a visual representation of the information, making it easier to see relationships and contradictions. It also provides a structured way to keep track of which possibilities have been ruled out and which remain viable.
Here's an example of what the initial setup of a table might look like for the first problem:

```
| House # | Book Preference | Sport        |
|---------|-----------------|--------------|
| 1       |                 | Golfer       |
| 2       |                 |              |
| 3       |                 |              |
| 4       |                 | Not Basketball|
```

As you fill in the table based on the clues, you can quickly see which options are left for each house, leading you to the correct solution. This method reduces the risk of overlooking a detail or making a logical error.

Second Stage (generated by GPT-3.5):
Let's use the given clues to solve the problem step by step:
1. The person who has a television lives in the fourth house.
2. The person who is wearing loafers lives in the second house.
3. The person who has a phone does not live in the second house.
4. The person who is wearing boots lives somewhere to the left of the person who is wearing loafers.
5. There are two houses between where the person who has a radio lives and where the person who is wearing sandals lives.
Using these clues, we can deduce the following:
- The person with the television lives in the 4th house.
- The person with loafers lives in the 2nd house.
- The person with the phone does not live in the 2nd house.
- The person with boots lives to the left of the person with loafers.
Based on these deductions, we can determine the following:
- The person with the gameboy must live in the 3rd house.
Therefore, the answer is: 3

Table 18: Examples of Logic Grid (GPT-4 ⇒ GPT-3.5 Task-Based)

第一阶段（由 GPT-4 生成）：
对于这些逻辑谜题，表格格式在增强清晰度并消除歧义方面尤其有效。表格允许以网格结构组织信息，这有助于解决此类问题所必需的逐步排除法与演绎推理过程。

以下是如何利用表格来提高准确率：
1. 创建一个表格，其中行代表房屋，列代表特征（例如：书籍偏好、运动、宠物、饮料、花卉、汽车、水果）。

2. 直接将线索中的已知信息填入表格。
3. 使用排除法填写空白。例如，如果高尔夫球手住在第一栋房子，那么其他人就不能是高尔夫球手，也不能住在第一栋房子。.

4. 马克·伊莫斯 可能的情况。例如，如果游泳者不能住在第三栋房子，则将相应单元格标记为不可能。
5. 继续交叉引用线索，填写或排除可能性，直到表格填写完整。
表格格式允许以可视化表示呈现信息，使其更易于发现关系和矛盾。它还提供了一种结构化方式来追踪哪些可能性已被排除，哪些仍然可行。

以下是针对第一个问题的表格初始设置的示例：

```
| 房屋编号 | 书籍偏好 | 运动          |
|---------|----------|--------------|
| 1       |          | 高尔夫球      |
| 2       |          | 手            |
| 3       |          |              |
| 4       |          | 不是篮球|
```

当你根据 线索填写表格时，你可以快速看出每个房子剩下哪些选项，从而引导你找到正确的解决方案。这种方法可以降低忽略细节或出现逻辑错误的风险。

第二阶段（由 GPT-3.5 生成）：
Let's 逐步分析vert c 拥有电视的人员住在第四栋房子。2. 穿着乐福鞋的人员住在第二间房子。3. 拥有电话的人员不住在第二间房子。4. 穿着靴子的人员住在穿着乐福鞋的人员的左侧某处。5. 拥有收音机的人员与穿着凉鞋的人员的住所之间隔着两栋房屋。利用这些线索，我们可以得出以下推论： - 拥有电视的人员住在第4栋房子。 - 拥有乐福鞋的人员住在第2间房子。 - 拥有电话的人员不住在第2间房子。 - 拥有靴子的人员住在拥有乐福鞋的人员的左侧。根据这些推论，我们可以确定以下事实： - 拥有游戏机的人员必须住在第3栋房子。因此，答案是：3

表格 18：逻辑网格示例（GPT-4 ⇒ GPT-3.5基于任务）

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

First Stage (generated by GPT-4):
For these logic puzzles, a tabular format can be particularly effective in enhancing clarity and eliminating ambiguities. Tables allow for a structured representation of the data, where each row can represent a house, and each column can represent a characteristic (e.g., pet, car, sport, etc.). This format makes it easy to visualize the constraints and relationships between different characteristics.
Here's how a table format can be used to solve these types of problems:
1. Create a table with rows corresponding to houses and columns corresponding to characteristics.
2. Fill in the known information based on the clues provided.
3. Use logical deduction to fill in the remaining cells, ensuring that each characteristic appears only once per row and column (since each person has different characteristics).
4. Update the table iteratively as new information is deduced until all cells are filled.
For example, a table for the first problem might look like this:

House	Pet	Car	Sport
1			
2	Not Cat		
3	Bird		Not Baseball
4			

Using a table format allows for a clear and organized approach to solving the problem. It helps to avoid confusion that might arise from trying to keep track of multiple pieces of information in a purely textual format. Additionally, the visual nature of tables makes it easier to spot patterns and relationships between different variables, which is crucial for solving logic puzzles.

Second Stage (generated by Gemini Pro):
Step 1: Create a table with rows corresponding to houses and columns corresponding to characteristics.

House	Sport	Flowers	Book Genre
1			
2			
3			
4			

Step 2: Fill in the known information based on the clues provided.
- Clue 1: There is one house between where the mystery book reader lives and where the swimmer lives.
- Clue 2: The tennis player lives somewhere to the right of the mystery book reader.
- Clue 3: The person who has a carnations arrangement lives in the fourth house.
- Clue 4: The fantasy book enthusiast lives directly left of the swimmer.
- Clue 5: The golfer lives somewhere to the right of the person who has a bouquet of daffodils.
- Clue 6: The person who has a rose bouquet lives in the third house.
- Clue 7: The tennis player does not live in the fourth house.
- Clue 8: There is one house between where the person who has a bouquet of daffodils lives and where the person who has a rose bouquet lives.
- Clue 9: There are two houses between where the mystery book reader lives and where the romance book lover lives.
Step 3: Use logical deduction to fill in the remaining cells, ensuring that each characteristic appears only once per row and column.
- From Clue 3, we know that the person with carnations lives in house 4.
- From Clue 6, we know that the person with roses lives in house 3.
- From Clue 8, we know that the person with daffodils lives in either house 1 or 2.
- From Clue 4, we know that the fantasy book enthusiast lives in house 2.
- From Clue 5, we know that the golfer lives in house 3 or 4.
- From Clue 7, we know that the tennis player does not live in house 4.
- From Clue 9, we know that the romance book lover lives in either house 1 or 2.
Step 4: Update the table iteratively as new information is deduced until all cells are filled.

House	Sport	Flowers	Book Genre
1	Tennis	Daffodils	Romance
2	Fantasy	Carnations	Mystery
3	Golf	Roses	History
4	Swimming	Lilies	Not Provided

Answer: The answer is: 1

Table 19: Examples of Logic Grid (GPT-4 ⇒ Gemini Pro Task-Based)

第一阶段（由GPT-4生成）：
对于这些逻辑谜题，表格格式可以特别有效地增强清晰度并消除歧义。表格允许对数据进行结构化表示，其中每一行可以代表一个房子，每一列可以代表一个特征（例如，宠物、汽车、运动等）。这种格式使得直观地可视化不同特征之间的约束和关系变得更容易。
表格格式可用于解决此类问题，具体方法如下：
1. 创建一个表格，行对应房屋，列对应特征。
2 Fi n the known inf$_{\text{ormat}}$ ion based on the cl$^{\text{ues prov}}$ ided.
3. 使用逻辑推理填写剩余的单元格，确保每个特征在每一行和每一列中只出现一次（因为每个人员都有不同的特征）。

4. 随着新信息的推断，迭代更新表格，直到所有单元格都被填写完毕。
例如，第一个问题的表格可能如下所示：

房子	宠物	汽车	运动
1			
2	Not Cat		
3	鸟		非棒球
4			

使用表格格式可以提供一种清晰且有组织的方法来解决该问题。它有助于避免因试图在纯文本格式中跟踪多个信息片段而产生的混乱。此外，表格的可视化特性使得更容易发现不同变量之间的模式和关系，这对于解决逻辑谜题至关重要。

第二阶段（由 Gemini Pro 生成）：
第一步：创建一个表格，行对应房屋，列对应特征。

房子	运动	花卉	书籍类型
1			
2			
3			
4			

步骤2：根据提供的线索填入已知信息。
- 线索1：神秘小说读者的住所与游泳者的住所之间有一栋房子。
- 线索2：网球运动员住在神秘小说读者的某个右边位置。
- 线索3：拥有康乃馨花束的人住在第四栋房子。
- 线索4：奇幻书爱好者住在游泳者的正左边。
- 线索4：高尔夫球手住在拥有水仙花束的人的某个右边位置。
- 线索6：拥有玫瑰花束的人住在第三栋房子。
- 线索7：网球运动员不住在第四栋房子 .
- 线索 8：持有水仙花束的人员居住的房子里和持有玫瑰花束的人员居住的房子里之间，有一栋房子。

- Clue 9: Th$^{\text{ere are two}}$ h$_{\text{ouses}}$ b$^{\text{etween w}}$ h$^{\text{ere t}}$ he mystery book reader li$^{\text{ves an}}$ d wh$^{\text{ere t}}$ he 言情 book lover lives.
步骤 3：使用逻辑推理填充其余的单元格，确保每个特征在每一行和每一列中只出现一次。

- 根据线索3，我们知道养康乃馨的人住在4号房屋。- 根据线索6，我们知道养玫瑰的人住在3号房屋。- 根据线索8，我们知道养水仙花的人住在1号房屋或2号房屋。- 根据线索4，我们知道奇幻书爱好者住在2号房屋。- 根据线索5，我们知道高尔夫球手住在3号房屋或4号房屋。- 根据线索7，我们知道网球运动员不住在4号房屋。- 根据线索9，我们知道言情小说爱好者住在1号房屋或2号房屋。**步骤 4：随着新信息的推论，迭代更新表格，直到所有单元格都被填满。**

房子	运动	花卉	书籍类型
1	网球	水仙花	言情
2	奇幻	康乃馨	神秘
3	高尔夫	玫瑰	历史
4	游泳	百合	未提供

答案： 答案是：1

表19：逻辑网格（GPT-4 ⇒Gemini Pro 基于任务）示例

21

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

First Stage:
To enhance the accuracy of the analysis for this logic puzzle, I would choose to use a markdown table format. Markdown tables provide a clear and structured way to represent the information given and the deductions made from the clues. This format allows for easy cross-referencing of clues and characteristics, which is essential for solving logic puzzles.
Here's how the markdown table format can be used to enhance clarity:
1. Create a table with headers for Houses (1, 2, 3), Lunch (Spaghetti, Pizza, Fried Rice), and Shoes (Loafers, Running Shoes, Boots).
2. Fill in the table with the information given directly by the clues.
3. Use process of elimination to fill in the rest of the table based on the clues and the constraints they impose.
The table would initially look like this:

```
| House # | Lunch       | Shoes         |
|---------|-------------|---------------|
| 1       | Spaghetti   | ?             |
| 2       | ?           | ?             |
| 3       | ?           | ?             |
```

Then, as you apply each clue, you update the table accordingly. For example, after applying clue 3, you know that the person in house 1 ate spaghetti. After applying clue 4, you know that the person in house 1 is not wearing loafers.
The table format allows you to visualize the problem and systematically apply each clue, reducing the chance of making a mistake and increasing the accuracy of the solution. It also makes it easier to communicate the reasoning process to others.

Second Stage:
Let's solve the problem step by step using the clues provided:
1. The person who ate fried rice does not live in the second house.
2. The person who is wearing boots lives directly left of the person who ate fried rice.
3. The person who ate spaghetti lives in the first house.
4. The person who is wearing loafers does not live in the first house.
Starting with clue 3, we can fill in the lunch for house 1:

```
| House # | Lunch       | Shoes         |
|---------|-------------|---------------|
| 1       | Spaghetti   | ?             |
| 2       | ?           | ?             |
| 3       | ?           | ?             |
```

Next, clue 4 tells us that the person in house 1 is not wearing loafers:

```
| House # | Lunch       | Shoes         |
|---------|-------------|---------------|
| 1       | Spaghetti   | Not Loafers   |
| 2       | ?           | ?             |
| 3       | ?           | ?             |
```

Clue 1 tells us that the person who ate fried rice does not live in house 2, so fried rice must be in house 3:

```
| House # | Lunch       | Shoes         |
|---------|-------------|---------------|
| 1       | Spaghetti   | Not Loafers   |
| 2       | ?           | ?             |
| 3       | Fried Rice  | ?             |
```

Clue 2 tells us that the person wearing boots lives directly left of the person who ate fried rice, which means the person in house 2 must be wearing boots:

```
| House # | Lunch       | Shoes         |
|---------|-------------|---------------|
| 1       | Spaghetti   | Not Loafers   |
| 2       | ?           | Boots         |
| 3       | Fried Rice  | ?             |
```

Since the person in house 1 is not wearing loafers and the person in house 2 is wearing boots, the person in house 3 must be wearing loafers:

```
| House # | Lunch       | Shoes         |
|---------|-------------|---------------|
| 1       | Spaghetti   | Not Loafers   |
| 2       | ?           | Boots         |
| 3       | Fried Rice  | Loafers       |
```

Now we know that the person in house 3 is wearing loafers. Therefore, the answer is: 3.

Table 20: Examples of Logic Grid (GPT-4 Instance-Based)

第一阶段：
为了提高这个逻辑谜题分析的准确率，我会选择使用 Markdown 表格格式。Markdown 表格提供了一种清晰、结构化的方式来表示已知信息和从线索中做出的推论。这种格式便于对线索和特征进行交叉引用，对于解决逻辑谜题至关重要。以下是如何使用 Markdown 表格格式来提高清晰度的示例：
1. 创建一个表格，表头分别为房屋（1、2、3）、午餐（意大利面、披萨、炒饭）和鞋子（乐福鞋、跑鞋、靴子）。

2. 根据线索直接提供的信息填写表格。
3. 使用排除法，根据线索及其施加的约束来填写表格的其余部分 .
该表格最初看起来会是这样：

```
|房屋编号 |  午餐        | 鞋子          |
|--------|------------|-------------|
| 1      |  意大利面    | ?           |
| 2      | ?          | ?           |
| 3      | ?          | ?           |
```

然后，当你应用 每条线索时，你相应地更新表格。例如，在应用线索3后，你知道1号房屋的人员吃了意大ghetti。应用线索4后，你知道1号房屋里的人没有穿乐福鞋。
表格格式让你能够可视化问题并系统地应用每一条线索，从而降低出错几率并提高解决方案的准确率。它还有助于向他人清晰地传达推理过程。

第二阶段：
让我们根据提供的线索分步骤解决问题：1. 吃炒饭的人不住在第二间房子。2. 穿靴子的人住在吃炒饭的人的左边。3. 吃意大利面的人住在第一栋房子。4. 穿乐福鞋的人不住在第一栋房子。

从线索3开始，我们可以填写1号房屋的午餐：

```
|房屋编号 | 午餐         | 鞋子          |
|--------|------------|-------------|
| 1      |  意大利面    | ?           |
| 2      | ?          | ?           |
| 3      | ?          | ?           |
```

接下来，线索4告诉我们，1号房屋的人员不穿乐福鞋：

```
|房屋编号 | 午餐         | 鞋子          |
|--------|------------|-------------|
| 1      | 意大利面     | 不是乐福鞋    |
| 2      | ?          | ?           |
| 3      | ?          | ?           |
```

线索1告诉我们，吃炒饭的人员不住在2号房屋，所以炒饭必须在3号房屋：

```
|房屋编号 | 午餐         | 鞋子          |
|--------|------------|-------------|
| 1      | 意大利面     | 不是乐福鞋    |
| 2      | ?          | ?           |
| 3      | 油炸  Rice   | ?           |
```

线索2 告诉我们，穿靴子的人员住在吃炒饭的人员的左侧，这意味着 2号房屋 的人员必须穿着靴子：

```
|房屋编号 | 午餐         | 鞋子          |
|--------|------------|-------------|
| 1      | 意大利面     | 不是乐福鞋    |
| 2      | ?          | 靴子          |
| 3      | 油炸  Rice   | ?           |
```

因为1号房屋的人员没有穿乐福鞋，2号房屋的人员穿着靴子，所以3号房屋的人员必须穿乐福鞋：

```
|房屋编号 | 午餐         | 鞋子          |
|--------|------------|-------------|
| 1      | 意大利面     | 不是乐福鞋    |
| 2      | ?          | 靴子          |
| 3      | 炒饭         | 乐福鞋        |
```

现在我们知道，3号房屋的人员穿着乐福鞋。因此，答案是：3。

表 20：逻辑网格示例（GPT-4 基于实例）

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

First Stage:
Communication Form: Logical Operators and Markdown Tables
Explanation:
Logical operators and markdown tables provide a structured way to represent the given information and clues, which can help in systematically deducing the solution to each problem. Logical operators such as "AND", "OR", "NOT", "=>"(implies), and "=="(equals) can be used to translate the clues into logical statements. Markdown tables can be used to visualize the possible arrangements of characteristics for each house.
For example, the clues for the first problem can be translated into logical statements as follows:
1. Tennis == House3
2. GrilledCheese == House5
3. Golfer == House4
4. FriedRice != Spaghetti AND (FriedRice == HouseN-1 OR FriedRice == HouseN+1) WHERE N is the house number of Spaghetti
5. Pizza == HouseM AND M < CricketerHouseNumber
6. Stew == HouseP AND Spaghetti == HouseP+1
7. TennisHouseNumber < SoccerHouseNumber
Using markdown tables, we can create a table with five columns (one for each house) and rows for each characteristic (sport, lunch). We can fill in the known information from the clues and use the logical operators to deduce the remaining characteristics.
Here's an example of how the table might look after filling in some known information:

```
| House 1 | House 2 | House 3 | House 4 | House 5 |
|---------|---------|---------|---------|---------|
|         |         | Tennis  | Golfer  |         |
|         |         |         |         | GrilledCheese |
```

The use of logical operators and markdown tables enhances accuracy by providing a clear framework for organizing information and making deductions. It reduces the cognitive load of keeping track of multiple pieces of information and allows for easier identification of contradictions and inferences. This structured approach is particularly useful for complex problems with multiple variables and constraints.

Second Stage:
Let's solve the problem step by step using logical operators and a markdown table.
We have three houses (1, 2, 3) and three characteristics for the residents (favorite drink, type of shoes). We'll use the clues to fill in the table.
Clues:
1. HighHeels == House2
2. CoffeeDrinker != House2
3. MilkDrinker != House1
4. CoffeeDrinker != House1
5. RunningShoes == House3
Let's create a table with the information we have:

```
| House 1 | House 2       | House 3       |
|---------|---------------|---------------|
|         | HighHeels     | RunningShoes  |
|         |               |               |
|         |               |               |
```

From clues 2 and 4, we know the CoffeeDrinker is not in House1 or House2, so they must be in House3.

```
| House 1 | House 2       | House 3           |
|---------|---------------|-------------------|
|         | HighHeels     | RunningShoes      |
|         |               | CoffeeDrinker     |
|         |               |                   |
```

Since the CoffeeDrinker is in House3, and the MilkDrinker is not in House1 (clue 3), the MilkDrinker must be in House2.

```
| House 1 | House 2           | House 3           |
|---------|-------------------|-------------------|
|         | HighHeels         | RunningShoes      |
|         | MilkDrinker       | CoffeeDrinker     |
|         |                   |                   |
```

The only remaining house for the WaterDrinker is House1, and the only remaining type of shoes for House1 is Loafers.

```
| House 1       | House 2           | House 3           |
|---------------|-------------------|-------------------|
| Loafers       | HighHeels         | RunningShoes      |
| WaterDrinker  | MilkDrinker       | CoffeeDrinker     |
|               |                   |                   |
```

Now we have all the information filled in, and we can see that the person who is wearing high heels lives in House2.
The answer is: 2

Table 21: Examples of Logic Grid (GPT-4 Task-Based)

第一阶段：
通信形式：逻辑运算符和 Markdown 表格 说明：

逻辑运算符和Markdown 表格提供了一种结构化的方式来表示给定的已知条件和线索，这有助于系统地推导每个问题的解决方案。像"AND"、"OR"、"NOT"、"=>"（蕴含）和"=="（等于）这样的逻辑运算符可以用于将线索转化为逻辑语句。Markdown 表格可用于可视化每座房子的特征的可能排列方式。
例如，第一个问题的线索可以转化为如下逻辑语句：
1. 网球 ==3号房子 2. 烤奶酪三明
治 == 5号房子 3. 高尔夫球手 ==4
号房子
4. FriedRice != Spaghetti AND (FriedRice ==HouseN-1 OR FriedRice == HouseN+1) WHERE N 是 Spaghetti 的房屋编号

5. 披萨 == 房子正面 与 正面 <板球手房屋编号
6. 炖菜 == 房子P 与 意大利面 == 房子P+1
7. 网球房子数字 < 足球房子数字
使用 Markdown 表格，我们可以创建一个包含五列（每栋房子一列）和针对每个特征（运动、午餐）的行数的表格。我们可以根据线索填入已知的信息，并使用逻辑运算符推导其余的特征。以下是填入部分已知信息后表格的可能样子：

```
| 房子 1 | 房子 2 |房子 3 | 房子 4 | 房子 5 |
|---------|---------|---------|---------|---------|
|         |         |         | 网球    | 高尔夫球 |         |
|         |         |         |         | 手      | 烤奶酪三明治 |
```

使用逻辑运算符和 Markdown 表格通过提供清晰的框架来组织信息并进行推论，从而提高准确率。它减少了跟踪多条信息所带来的认知负荷，并使识别矛盾和推断变得更加容易。这种结构化方法对于包含多个变量和约束的复杂问题尤其有用。

第二阶段：
让我们使用逻辑运算符和 Markdown 表格分步骤解决这个问题。
我们有三个 房屋（1、2、3）以及居民的三个特征（最喜欢的饮料、鞋子类型）。我们将使用线索来填写表格。
线索：
1 == 2 . 高跟鞋 房
2 ==
子 . 咖啡饮用者 ! 2号房子 3. 牛奶
饮用者!=1号房子 4. 咖啡饮用者!=
1号房子 5. 跑鞋 ==3号房子

让我们根据已有信息创建一个表格：

```
| 房子 1 | 2号房屋        |3号房子         |
|---------|---------------|---------------|
|         | 高跟鞋        | 跑鞋          |
|         |               |               |
|         |               |               |
```

根据线索2和4，我们知道咖啡饮用者不在1号房子或2号房子，所以一定在3号房子。

```
| 房子 1 | 房子 2        |房子 3          |--------
|---------|---------------|               | 高跟鞋 |跑鞋 |
|         |               |咖啡饮用者 |
|         |               |                |
```

由于C 由于咖啡饮用者在3号房子，并且牛奶饮用者不在1号房子（线索3），所以牛奶饮用者必须在2号房子。

```
| 房子 1 | 2号房子          | 3号房子          |
|---------|---------------|---------------|
|         | 高跟鞋        | 跑鞋          |
|         | 牛奶饮用者    | 咖啡饮用者    |
|         |               |               |
```

水饮用者唯一剩下的房子是1号房子，而1号房子唯一剩下的鞋子类型是乐福鞋。

```
| 1号房屋        | 2号房屋          | 3号房屋          |
|---------------|---------------|---------------|
| 乐福鞋        | 高跟鞋        | 跑鞋          |
| 水饮用者      | 牛奶饮用者    | 咖啡饮用者    |
|               |               |               |
```

现在所有信息都已填写完毕，我们可以看到，穿高跟鞋的人住在2号房子。答案是：2

表21：逻辑网格示例（GPT-4任务型）

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

First Stage:
To enhance the clarity and eliminate ambiguities inherent in natural language, a structured and concise form of communication can be employed for the step-by-step solution. One suitable format is a JSON object, which can be used to represent the information provided in the problem statement in a structured manner. The JSON object can include keys for each of the characteristics mentioned in the problem statement, such as "device" and "lunch", and the values for each key can be arrays containing the different options for that characteristic. For example, the "device" key could have an array containing the values ["phone", "gameboy", "computer", "radio", "television"].
Using a JSON object to represent the information in the problem statement can enhance the accuracy of the analysis by providing a structured and organized way to store and manipulate the data. This can make it easier to identify patterns and relationships between the different characteristics, and to identify any inconsistencies or contradictions in the information provided. Additionally, the JSON object can be easily converted into other formats, such as a table or a graph, which can be used to visualize the data and identify additional insights.

Second Stage:
```
{
  "devices": ["phone", "gameboy", "computer", "radio", "television"],
  "lunch": ["spaghetti", "pizza", "stew", "cheese", "fried rice"],
  "clues": [
    "The person who has a gameboy and the person who has a computer live directly next to each other.",
    "The person who has a television lives somewhere to the left of the person who has a phone.",
    "There is one house between where the person who ate cheese lives and where the person who has a radio lives.",
    "There is one house between where the person who has a gameboy lives and where the person who ate fried rice lives.",
    "The person who ate cheese lives directly left of the person who ate spaghetti.",
    "The person who has a radio does not live in the fourth house.",
    "The person who had pizza lives in the fourth house.",
    "The person who has a computer does not live in the fourth house."
  ]
}
```

1. From clue 7, we know that the person who had pizza lives in the fourth house.
2. From clue 8, we know that the person who has a computer does not live in the fourth house.
3. From clue 1, we know that the person who has a gameboy and the person who has a computer live directly next to each other.
4. From clue 2, we know that the person who has a television lives somewhere to the left of the person who has a phone.
5. From clue 3, we know that there is one house between where the person who ate cheese lives and where the person who has a radio lives.
6. From clue 4, we know that there is one house between where the person who has a gameboy lives and where the person who ate fried rice lives.
7. From clue 5, we know that the person who ate cheese lives directly left of the person who ate spaghetti.
8. From clue 6, we know that the person who has a radio does not live in the fourth house.
Based on the above analysis, we can deduce the following:
```
House 1: Gameboy, Fried rice
House 2: Computer, Spaghetti
House 3: Television, Cheese
House 4: Pizza, ?
House 5: Phone, Stew
```
Therefore, the answer is: 2

Table 22: Examples of Logic Grid (Gemini Pro Instance-Based)

第一阶段：
为了增强清晰度并消除自然语言固有的歧义，可以采用结构化且简洁的通信形式来进行逐步解决方案。一种合适的格式是JSON对象，它可用于以结构化方式表示问题陈述中提供的信息。该JSON对象可以包含问题陈述中提到的每个特征的键，例如“设备”和“午餐”，并且每个键的数值可以是包含该特征不同选项的数组。例如，“设备”键可以包含一个数组，其中包含数值 ["电话", "游戏机", "电脑", "收音机", "电视"]。

使用JSON对象来表示问题陈述中的信息，可以通过提供结构化且有序的方式来存储和操作数据，从而提高分析的准确率。这有助于更轻松地识别不同特征之间的模式和关系，并发现所提供信息中的任何不一致之处或矛盾。此外，JSON对象可以轻松地转换为其他格式，例如表格或图形，用于可视化数据并识别额外的见解。

第二阶段：
```
{ "设备": ["电话", "游戏机", "电脑", "收音机", "电视"],"午餐": ["意大利面", "披萨", "炖菜", "奶酪", "炒饭"],"线索": ["拥有游戏机的人员与拥有电脑的人员相邻居住。", "拥有电视的人员居住在拥有电话的人员的左侧某处。", "食用奶酪的人员与拥有收音机的人员之间隔着一栋房子。", "拥有游戏机的人员与食用炒饭的人员之间隔着一栋房子。", "食用奶酪的人员直接居住在食用意大利面的人员的左侧。", "拥有收音机的人员不住在第四栋房子。", "食用披萨的人员住在第四栋房子。", "拥有电脑的人员不住在第四栋房子。"]}、、、
```

1. 根据线索 7，我们知道吃披萨的人住在第四栋房子 .
2. 从线索8，我们知道拥有电脑的人不住在第四栋房子。 3. 从线索1，我们知道拥有游戏机的人和拥有电脑的人直接相邻。 4. 从线索2，我们知道拥有电视的人住在拥有电话的人的某个左侧位置。 5. 从线索3，我们知道吃了奶酪的人的住所和拥有收音机的人的住所之间隔了一栋房子。 6. 从线索4，我们知道拥有游戏机的人的住所和吃了炒饭的人的住所之间隔了一栋房子。 7. 从线索5，我们知道吃了奶酪的人直接住在吃了意大利面的人的左边。 8. 从线索6，我们知道拥有收音机的人不住在第四栋房子。 基于以上分析，我们可以得出以下结论：、、、1号房屋：游戏机，炒饭 2号房屋：电脑，意大利面 3号房屋：电视，奶酪 4号房屋：披萨，？第五栋房子：电话，炖菜、、、因此，答案是：2

表22: 逻辑网格 (Gemini Pro 基于实例) 示例

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

First Stage:
Using a structured communication form, such as a table, can enhance the accuracy of the analysis by providing a clear and organized representation of the information. Each row of the table can represent a person, and each column can represent a characteristic. This allows for easy identification of patterns and relationships between the characteristics. Additionally, the use of logical operators, such as "and" and "or," can help to express complex relationships between the characteristics. For example, the clue "The romance book lover lives somewhere to the left of the fantasy book enthusiast" can be represented as "RomanceBookLover < FantasyBookEnthusiast." This structured approach helps to ensure that all of the information is accurately captured and analyzed, leading to a more accurate solution.

Second Stage:
Step 1: Create a structured communication form.

```
| House Number | Lunch | Favorite Color | Shoes |
|---|---|---|---|
| 1 | ? | White | ? |
| 2 | ? | Green | ? |
| 3 | ? | ? | ? |
```

Step 2: Analyze the clues and fill in the table.

```
- Clue 1: The person who is wearing high heels does not live in the third house.
  - This means that the person wearing high heels lives in either house 1 or house 2.

- Clue 2: The person who likes white lives in the first house.
  - This means that the person in house 1 likes white.

- Clue 3: The person who ate spaghetti does not live in the second house.
  - This means that the person in house 2 did not eat spaghetti.

- Clue 4: The person who likes green lives in the second house.
  - This means that the person in house 2 likes green.

- Clue 5: The person who is wearing boots lives directly left of the person who is wearing high heels.
  - Since the person wearing high heels lives in either house 1 or house 2, the person wearing boots must
  live in house 1.

- Clue 6: The person who is wearing boots lives somewhere to the left of the person who had stew.
  - This means that the person who had stew lives in house 3.

- Clue 7: The person who had stew does not live in the second house.
  - This confirms that the person who had stew lives in house 3.
```

Step 3: Complete the table.

```
| House Number | Lunch | Favorite Color | Shoes |
|---|---|---|---|
| 1 | ? | White | Boots |
| 2 | Fried Rice | Green | Running Shoes |
| 3 | Stew | ? | High Heels |
```

Answer:
The answer is: 2

Table 23: Examples of Logic Grid (Gemini Pro Task-Based)

第一阶段：
使用结构化沟通形式，例如表格，可以通过提供信息的清晰、有条理表示来增强分析的准确率。表格的每一行可以代表一个人员，每一列可以代表一个特征。这使得能够轻松地识别特征之间的模式和关系。此外，使用诸如“与”和“或”的逻辑运算符，有助于表达特征之间的复杂关系。例如，线索“言情小说爱好者住在奇幻书爱好者的某个左侧位置”可以表示为“RomanceBookLover < FantasyBookEnthusiast”。这种结构化方法有助于确保所有信息都被准确记录和分析，从而得出更准确的解决方案。

第二阶段：
第一步：创建一个结构化沟通形式。
| 房屋编号 |午餐 |最喜欢的颜色 | 鞋子 |---|---|---|---|
| 1 | ? | 白色 |? | 2 | ? | 绿色 |? | 3 | ? | ? |?|

步骤2：分析线索并填写表格。

- 线索1：穿高跟鞋的人员不住在第三栋房子。 - 这意味着穿高跟鞋的人员住在1号房屋或2号房屋。 - 线索2：喜欢白色的人员住在第一栋房子。 - 这意味着1号房屋的人员喜欢白色。 - 线索3：吃意大利面的人员不住在第二间房子。 - 这意味着 2 号房屋的人员没有吃意大利面。 - 线索 4：喜欢绿色的人员住在第二间房子。 - 这意味着 2 号房屋的人员喜欢绿色。 - 线索5：穿靴子的人员住在穿高跟鞋的人员的正左边。 - 由于穿高跟鞋的人员住在1号房屋或2号房屋，穿靴子的人员必须住在1号房屋。 - 线索6：穿靴子的人员住在吃炖菜的人员的左边某处。 - 这意味着吃炖菜的人员住在3号房屋。 - 线索7：吃炖菜的人员不住在第二间房子。 - 这证实了吃炖菜的人员住在3号房屋。

步骤 3：完成表格 .
| 房屋编号 |午餐 |最喜欢的颜色 | 鞋子 |---|---|---|---|
| 1 | ? | 白色 |靴子 | 2 | 炒饭 |绿色 |跑鞋 |
| 3 | 炖菜 |? | 高跟鞋 |**答案：** Theansweris:
2

表23：基于任务的逻辑网格示例（Gemini Pro 任务）

25

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

First Stage (generated by GPT-3.5):
Coin state:
- Initial state: heads up
- List of people who flip the coin: [kolbin, jaleeya, lacey, dimitricus, aaliyah-mae, kaytor, lantigua, tjeu, harout, halynn]
- List of people who do not flip the coin: [woodsr, eliborio, selars, korianna, domanico, xzaveon, tonnica, naomh, eldwon, samon]

Second Stage (generated by GPT-4):
Initial State Representation:
- Initial state of the coin: heads up
- People who flip the coin: [niral, bořivoj]
- People who do not flip the coin: [jaylo, tao]
Step-by-Step Solution Process:
1. The coin starts heads up.
2. Niral flips the coin. This changes the state from heads up to tails up.
3. Bořivoj flips the coin. This changes the state from tails up back to heads up.
4. Jaylo does not flip the coin, so the state remains unchanged.
5. Tao does not flip the coin, so the state remains unchanged.
Concluding with the Answer:
Answer: yes

Table 24: Examples of Coin Flip (GPT-3.5 ⇒ GPT-4 Task-Based)

第一阶段（由 GPT-3.5 生成）：
Coi n状态: - 初始状态: 正面朝上 - 翻硬币的人员列表: [科尔宾, 贾丽雅, 蕾西, 迪米特里库斯, 阿里娅·梅, 凯特, 兰蒂瓜, 特尤, 哈鲁特, 哈利恩]- 不翻硬币的人员列表: [伍兹, 埃利博里奥, 塞勒斯, 科里安娜, 多曼尼科, 泽维恩, 托尼卡, 诺姆, 埃尔多恩, 萨蒙]

第二阶段（由 GPT-4 生成）：我
nitial State R$_{\text{epresentat}}$ ion:
- 硬币的初始状态：正面朝上
- 翻硬币的人： [尼拉尔, 博日沃伊]
- 不翻硬币的人： [杰洛, 涛]
逐步求解过程：
1. 硬币开始为正面朝上。
2. 尼拉尔翻转硬币。状态从正面朝上变为反面朝上。
3. 博日沃伊翻转硬币。状态从反面朝上变回正面朝上。
4. 杰洛没有翻转硬币，因此状态保持不变。
5. 涛没有翻转硬币，因此状态保持不变。
以答案结尾：
答案：是

表24：抛硬币示例（GPT-3.5 ⇒ GPT-4任务型）

First Stage:
Initial State Representation:
- Coin is heads up.
- guardacosta does not flip the coin.
- suy flips the coin.
- venterea flips the coin.
- leshawn flips the coin.

Second Stage:
Initial State Representation:
- Coin is heads up.
- guardacosta does not flip the coin.
- suy flips the coin.
- venterea flips the coin.
- leshawn flips the coin.
Step-by-Step Solution:
1. The initial state is that the coin is heads up.
2. suy flips the coin, which could result in the coin being either heads or tails.
3. venterea flips the coin, which could result in the coin being either heads or tails.
4. leshawn flips the coin, which could result in the coin being either heads or tails.
Conclusion:
The coin could be either heads or tails after the series of flips, so the final state is uncertain.
Answer: No

Table 25: Examples of Coin Flip (GPT-3.5 Instance-Based)

第一阶段：
初始状态表示： - 硬币正面朝上。
- guardacosta 不翻硬币。 -
suy 翻硬币。 - venterea 翻硬币。
- leshawn 翻硬币。

第二阶段：
初始状态表示：- 硬币正面朝上。 - guardacosta 不翻硬币。 - suy 翻硬币。 - venterea 翻硬币。 - leshawn 翻硬币。 逐步求解过程：1. 初始状态是硬币正面朝上。 2. suy 翻硬币，可能导致硬币变为正面或反面。 3. venterea 翻硬币，可能导致硬币变为正面或反面。 4. leshawn 翻硬币，可能导致硬币变为正面或反面。 结论：经过一系列翻转后，硬币可能是正面也可能是反面，因此最终状态不确定。 答案：否

表格 25：抛硬币示例（GPT-3.5 基于实例）

26

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

First Stage:
The unified representation form for the given problems could be a logical statement:
Coin is heads up if and only if (n1 does not flip the coin) AND (n2 flips the coin) AND (n3 does not flip the coin) AND (n4 does not flip the coin)
Where n1, n2, n3, and n4 represent the individuals involved in flipping or not flipping the coin.

Second Stage:
Initial State Representation:
Coin is heads up if and only if (yiğitcan flips the coin) AND (treymon does not flip the coin) AND (kindig flips the coin) AND (letcho flips the coin)
Step-by-Step Solution Process:
1. Initially, the coin is heads up.
   Coin is heads up if and only if (yiğitcan flips the coin) AND (treymon does not flip the coin) AND (kindig flips the coin) AND (letcho flips the coin)
2. yiğitcan flips the coin.
   Coin is heads up if and only if (treymon does not flip the coin) AND (kindig flips the coin) AND (letcho flips the coin)
3. treymon does not flip the coin.
   Coin is heads up if and only if (kindig flips the coin) AND (letcho flips the coin)
4. kindig flips the coin.
   Coin is not heads up.
Answer: No

Table 26: Examples of Coin Flip (GPT-3.5 Task-Based)

第一阶段：给定问题的统一表示形式可以是一个逻辑语句：硬币正面朝上当且仅当（n1 不翻硬币）与（n2 翻硬币）与（n3 不翻硬币）与（n4 不翻硬币）。 其中 n1、n2、n3 和 n4 代表参与翻或不翻硬币的个体。 第二阶段：初始状态表示：硬币正面朝上当且仅当（yi˘gitcan 翻硬币）与（treymon 不翻硬币）与（kindig 翻硬币）与（letcho 翻硬币）。 逐步求解过程：1. 初始时，硬币正面朝上。硬币正面朝上当且仅当（yi˘gitcan 翻硬币）与（treymon 不翻硬币）与（kindig 翻硬币）与（letcho 翻硬币）。 2. yi˘gitcan 翻硬币。硬币正面朝上当且仅当（treymon 不翻硬币）与（kindig 翻硬币）与（letcho 翻硬币）。 3. treymon 不翻硬币。硬币正面朝上当且仅当（kindig 翻硬币）与（letcho 翻硬币）。 4. kindig 翻硬币。硬币未正面朝上。 答案：否

表26：抛硬币示例（GPT-3.5基于任务）

First Stage (generated by GPT-4):
JSON Representation:

```json
{
  "initialState": "heads",
  "actions": [
    {"person": "Person1", "action": "flip"},
    {"person": "Person2", "action": "flip"},
    {"person": "Person3", "action": "no_flip"},
    {"person": "Person4", "action": "no_flip"}
  ]
}
```

Second Stage (generated by GPT-3.5):

```json
{
  "initialState": "heads",
  "actions": [
    {"person": "vickers", "action": "flip"},
    {"person": "encarna", "action": "no_flip"},
    {"person": "devan", "action": "flip"},
    {"person": "steingerður", "action": "flip"}
  ]
}
```

The coin was initially heads up. Vickers flipped the coin, so it is now tails up. Therefore, the coin is not still heads up.
Answer: no

Table 27: Examples of Coin Flip (GPT-4 ⇒ GPT-3.5 Task-Based)

第一阶段（由GPT-4生成）：
JSON表示：

```json
{
  "initialState": "正面",
  "动作": [
    {"人员": "Person1", "动作": "翻转"},
    {"人员": "Person2", "动作": "翻转"},
    {"人员": "Person3", "动作": "否_翻转"},
    {"人员": "人物4", "动作": "否_翻转"}
  ]
}
```

第二阶段（由 GPT-3.5 生成）：

```json
{
  "initialState": "正面",
  "动作": [
    {"人员": "维克斯", "动作": "翻转"},
    {"人员": "恩卡尔纳", "动作": "no_flip"},
    {"人员": "德凡", "动作": "翻转"},
    {"人员": "斯泰因格尔杜尔", "动作": "翻转"}
  ]
}
```

硬币最初是正面朝上。维克斯翻转了硬币，因此现在是反面朝上。所以，硬币不再处于正面朝上的状态。
答案：否

表27：抛硬币示例（GPT-4 ⇒ GPT-3.5基于任务）

27

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

First Stage (generated by GPT-4):
Logical Operators Representation:
Let H represent the state of the coin being heads up, and T represent the state of the coin being tails up. Let F represent the action of flipping the coin. The initial state of the coin is H.
For each person, we assign a binary variable that can be either 0 (does not flip the coin) or 1 (flips the coin). We can represent the action of flipping the coin as a function that takes the current state of the coin and the binary variable associated with a person's action:
flip(state, action) = (state AND NOT action) OR (NOT state AND action)
This function uses logical AND, OR, and NOT operators. The state of the coin after a sequence of actions can be determined by applying the flip function iteratively for each person's action.

Second Stage (generated by Gemini Pro):
Initial State Representation:
```
State: H
```

Step-by-Step Solution Process:
1. Andraya does not flip the coin. This means that the state of the coin remains the same.
```
State: H
```

2. Jerdy flips the coin. We apply the flip function:
```
flip(H, 1) = (H AND NOT 1) OR (NOT H AND 1)
flip(H, 1) = (H AND 0) OR (NOT H AND 1)
flip(H, 1) = 0 OR 1
flip(H, 1) = 1
```

Therefore, the state of the coin changes to tails up.
```
State: T
```

3. Kylila flips the coin. We apply the flip function again:
```
flip(T, 1) = (T AND NOT 1) OR (NOT T AND 1)
flip(T, 1) = (T AND 0) OR (NOT T AND 1)
flip(T, 1) = 0 OR 1
flip(T, 1) = 1
```

Therefore, the state of the coin changes back to heads up.
```
State: H
```

4. Degges flips the coin. We apply the flip function one more time:
```
flip(H, 1) = (H AND NOT 1) OR (NOT H AND 1)
flip(H, 1) = (H AND 0) OR (NOT H AND 1)
flip(H, 1) = 0 OR 1
flip(H, 1) = 1
```

Therefore, the state of the coin changes to tails up.
```
State: T
```

Answer: no

Table 28: Examples of Coin Flip (GPT-4 ⇒ Gemini Pro Task-Based)

第一阶段（由GPT-4生成）：
逻辑运算符表示法：
设H表示硬币正面朝上的状态，T表示硬币反面朝上的状态。设F表示翻转硬币的动作。硬币的初始状态为H。

对于每个人员，我们分配一个二元变量，该变量可以取0（不翻转硬币）或1（翻转硬币）。我们可以将翻转硬币的动作表示为一个函数，该函数接收硬币的当前状态和与人员的动作相关联的二元变量：

翻转函数(state, action) = (state 与 非 action) 或 (非 state 与 action)
该函数使用逻辑与、逻辑或和逻辑非运算符。硬币在一系列动作后的状态可以通过对每个人员的动作迭代应用翻转函数来确定。

第二阶段（由 Gemini Pro 生成）：
Initial State Representation:
```
状态：正面
```

逐步求解过程：
1. 安德莱亚不翻转硬币。这意味着硬币的状态保持不变。
```
状态：正面
```

2. 杰迪翻转硬币。我们应用翻转函数：
```
flip(正面, 1) = (正面 与 非 1)或 (非 正面 与 1)
flip(正面, 1) = (正面 与 0)或 (非 正面 与 1)
flip(正面, 1) = 0或 1 flip(正面, 1) = 1
```

因此，硬币的状态变为反面朝上。
```
状态：反面
```

3. 基拉拉翻转硬币。我们再次应用翻转函数：
```
flip(反面, 1) = (反面 与 非 1)或 (非 反面 与 1)
flip(反面, 1) = (反面 与 0)或 (非 反面 与 1)
flip(反面, 1) = 0或 1 flip(反面, 1) = 1
```

因此，硬币的状态变回正面朝上。
```
状态：正面
```

4. 德格斯翻转硬币。我们再次应用翻转函数：
```
flip(正面, 1) = (正面 与 非 1)或 (非 正面 与 1)
flip(正面, 1) = (正面 与 0)或 (非 正面 与 1)
flip(正面, 1) = 0或 1 flip(正面, 1) = 1
```

因此，硬币的状态变为反面朝上。
```
状态：反面
```

答案：否

表28: 抛硬币示例（GPT-4 ⇒ Gemini Pro 任务-基于）

28

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

First Stage:

```json
{
  "initialState": "heads",
  "flips": [
    {"person": "jerelyn", "action": "flip"},
    {"person": "shilap", "action": "flip"},
    {"person": "alireza", "action": "flip"},
    {"person": "fanette", "action": "no_flip"}
  ]
}
```

Second Stage:

```json
{
  "initialState": "heads",
  "flips": [
    {"person": "jerelyn", "action": "flip", "resultingState": "tails"},
    {"person": "shilap", "action": "flip", "resultingState": "heads"},
    {"person": "alireza", "action": "flip", "resultingState": "tails"},
    {"person": "fanette", "action": "no_flip", "resultingState": "tails"}
  ]
}
```

Answer: no

Table 29: Examples of Coin Flip (GPT-4 Instance-Based)

First Stage:
Logical Operators Representation:
Let H represent the state of the coin being heads up.
Let F represent the action of flipping the coin.
Let N represent the action of not flipping the coin.
The state of the coin after a series of actions can be represented using the XOR (exclusive or) logical operator, which we will denote as $\oplus$. Each flip of the coin changes its state, while not flipping leaves the state unchanged.
The generic representation for a sequence of actions on the coin is:
$H \oplus F1 \oplus F2 \oplus ... \oplus Fn$
Where:
- H is the initial state of the coin (heads up).
- F1 through Fn represent the actions taken by individuals, with F being a flip (value 1) and N being no flip (value 0).
This representation can be used to determine the final state of the coin after any number of flips and non-flips by evaluating the XOR operations in sequence.

Second Stage:
1. Initial State Representation:
Let's represent the initial state of the coin as H (heads up). Each flip of the coin will be represented as F, and not flipping the coin will be represented as N.
Initial state: H
2. Step-by-Step Solution Process:
- Chantasia flips the coin: $H \oplus F$
- Hagg flips the coin: $(H \oplus F) \oplus F$
- Jacob-James flips the coin: $((H \oplus F) \oplus F) \oplus F$
- Cheatham does not flip the coin: $(((H \oplus F) \oplus F) \oplus F) \oplus N$
Now, let's evaluate the XOR operations:
- $H \oplus F$ changes the state to tails (T), since flipping changes the state.
- $T \oplus F$ changes the state back to heads (H), since flipping again changes the state.
- $H \oplus F$ changes the state back to tails (T), since flipping changes the state.
- $T \oplus N$ leaves the state as tails (T), since not flipping does not change the state.
3. Concluding with the Answer:
Answer: no

Table 30: Examples of Coin Flip (GPT-4 Task-Based)

第一阶段：

```json { "initialState": "正面", "flips": [{"
person": "杰里林", "action": "翻转"}, {"person": "
希拉普", "action": "翻转"}, {"person": "阿里雷扎", "
action": "翻转"}, {"person": "法内特", "action": "不
_翻转"}]}、、、
```

第二阶段：

```json { "initialState": "正面", "flips": [{"person": "杰里林", "action": "翻转", "
resultingState": "反面"}, {"person": "希拉普", "action": "翻转", "resultingState":
"正面"}, {"person": "阿里雷扎", "action": "翻转", "resultingState": "反面"}, {"
person": "法内特", "action": "no_翻转", "resultingState": "反面"}]}、、、
```

答案：否

表29：抛硬币示例（GPT-4 基于实例）

第一阶段：
逻辑运算符表示法：令 H 表示硬币正面朝上的状态。令 F 表示翻转硬币的动作。令 N 表示不翻转硬币的动作。经过一系列动作后硬币的状态可以使用异或（exclusive or）逻辑运算符来表示，我们将其记为 $\oplus$。每次翻转硬币都会改变其状态，而不翻转则使状态保持不变。硬币上一系列动作的通用表示为：$H \oplus F1 \oplus F2 \oplus ... \oplus Fn$ 其中：- H 是硬币的初始状态（正面朝上）。- F1 到 Fn 表示个体所采取的动作，其中 F 表示翻转（值为 1），N 表示不翻转（值为 0）。该表示法可用于通过依次计算异或运算，确定经过任意次数翻转与不翻转后硬币的最终状态。

第二阶段：.
1
初始状态表示：让我们将硬币的初始状态表示为 H（正面朝上）。每次翻转硬币将表示为 F，不翻转硬币表示为 N。初始状态：H 2. 逐步求解过程：- 昌塔西亚翻转硬币：$H \oplus F$ - 哈格翻转硬币：$(H \oplus F) \oplus F$ - 雅各布-詹姆斯翻转硬币：$((H \oplus F) \oplus F) \oplus F$ - 希萨姆没有翻转硬币：$(((H \oplus F) \oplus F) \oplus F) \oplus N$ 现在，让我们评估异或操作：- $H \oplus F$ 将状态变为反面（T），因为翻转会改变状态。- $T \oplus F$ 将状态变回正面（H），因为再次翻转会改变状态。- $H \oplus F$ 将状态变回反面（T），因为翻转会改变状态。- $T \oplus N$ 使状态保持为反面（T），因为不翻转不会改变状态。3. 以答案结尾：答案：否

表30：抛硬币示例（GPT-4任务型）

29

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

First Stage:
```
[coin_state, garen, leontae, neveda, prabhjot]
```

Second Stage:
Initial State Representation:
```
[coin_state, garen, leontae, neveda, prabhjot]
```
Step 1:
* Garen does not flip the coin, so the coin state remains the same.
Step 2:
* Leontae flips the coin, so the coin state changes from heads up to tails up or vice versa.
Step 3:
* Nevada does not flip the coin, so the coin state remains the same.
Step 4:
* Prabhjot does not flip the coin, so the coin state remains the same.
Concluding with the Answer:
Answer: no

Table 31: Examples of Coin Flip (Gemini Pro Instance-Based)

第一阶段：

、、[硬币_状态, garen, leontae, neveda, prabhjot]、、第二阶段：**初始状态表示：**、、[硬币_状态, garen, leontae, neveda, prabhjot]、、**步骤 1：** * 加伦没有翻转硬币，因此硬币状态保持不变。 **步骤 2：** * 莱昂泰翻转硬币，因此硬币状态从正面朝上变为反面朝上，反之亦然。 **步骤 3：** * 内华达没有翻转硬币，因此硬币状态保持不变。 **步骤 4：** * 普拉布乔特没有翻转硬币，因此硬币状态保持不变。 **以答案结尾：** 答案：否

表31：抛硬币示例（Gemini Pro 基于实例）

First Stage:
```
coin_state = heads
person_1_flips = False
person_2_flips = False
person_3_flips = False
person_4_flips = False
```

Second Stage:
Initial State Representation:
```
coin_state = heads
person_1_flips = False
person_2_flips = False
person_3_flips = False
person_4_flips = True
```
Step-by-Step Solution Process:
1. mcgirth does not flip the coin, so the coin state remains the same.
2. wilem does not flip the coin, so the coin state remains the same.
3. bunney does not flip the coin, so the coin state remains the same.
4. adam-jorg flips the coin, so the coin state changes from heads to tails.
Answer: no

Table 32: Examples of Coin Flip (Gemini Pro Task-Based)

第一阶段：
```
硬币_状态 =正面 人员
_1_翻转 =假 人员_2_翻
转 =假 人员_3_翻转 =
假 人员_4_翻转 =假、、、
```

第二阶段：
初始状态表示：
```
coin_状态 =正面
人员_1_翻转 =非
人员_2_翻转 =非
人员_3_翻转 =非
人员_4_翻转 =真
```
逐步求解过程： .
1 麦基思没有翻硬币，因此硬币状态保持不变。
2. 怀勒姆没有翻硬币，因此硬币状态保持不变。
3. 巴尼没有翻硬币，因此硬币状态保持不变。
4. 亚当-乔格翻硬币，因此硬币状态从正面变为反面。
答案：否

表32：基于任务的抛硬币示例（Gemini Pro）

30

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

First Stage (generated by GPT-3.5):
I will use a table to present the information in a structured and concise manner. This will help to clearly outline the sufficiency of each statement and make it easier to compare and analyze the information provided.

Problem	Statement 1	Statement 2	Sufficiency
Elephants vs Zebras	Elephants are Alice's favourite animals.	Zebras are the animals that Alice likes the least.	Either statement 1 or statement 2 is sufficient
Q and R	R is the niece of P.	S has one sister R.	Statement 2 alone is sufficient while statement 1 alone is insufficient
PQRs and XYZs	Some XYZs are ABCs.	Some XYZs are PQRs.	Statement 2 alone is sufficient while statement 1 alone is insufficient
Tony's birthday	Natasha correctly states that Tony's birthday is after 16th and before 19th March.	Brenna incorrectly states that Tony's birthday is on 20th March.	Statement 1 alone is sufficient while statement 2 alone is insufficient
Cole's height	Cole is under 7 feet tall.	Every player on Cole's varsity basketball team is above 6'1”.	Statement 2 alone is sufficient while statement 1 alone is insufficient

Using this table format allows for a clear and concise comparison of the sufficiency of each statement for the given problems. It enhances the accuracy of the analysis by providing a structured framework for evaluating the information provided in each problem.

Second Stage (generated by GPT-4):
Step 1: Analyze Statement 1
Karan said that he had one sister.
This statement tells us that Karan has a sister, but it does not provide any information about Jhanvi. Without additional information, we cannot determine if Jhanvi is Karan’s sister or not.
Step 2: Analyze Statement 2
Jhanvi said that she had one brother.
This statement tells us that Jhanvi has a brother, but similarly to Statement 1, it does not provide any information about Karan. Without additional information, we cannot determine if Karan is Jhanvi’s brother or not.
Step 3: Combine Statement 1 and Statement 2
If we take both statements together, we can deduce that Karan has one sister and Jhanvi has one brother. If Karan and Jhanvi are siblings and each has only one sibling of the opposite gender, then it is logical to conclude that Karan is Jhanvi’s brother and Jhanvi is Karan’s sister.
Answer: Statement 1 and statement 2 taken together are sufficient.

Table 33: Examples of Information Essentiality (GPT-3.5 ⇒ GPT-4 Task-Based)

First Stage:
Initial State Representation:

- We need to determine if x is a prime number.
- We have two statements to consider:
  1. x is divisible by the factors of 10.
  2. x is greater than 13.

Based on the given statements, let’s analyze each statement separately:

Second Stage:
Statement 1: x is divisible by the factors of 10.
- This statement alone is not sufficient to determine if x is a prime number. For example, if x is 10, it is not a prime number, but if x is 17, it is a prime number.
Statement 2: x is greater than 13.
- This statement alone is also not sufficient to determine if x is a prime number. For example, if x is 14, it is not a prime number, but if x is 17, it is a primenumber.
Considering both statements together:
- Even when considering both statements together, we cannot definitively determine if x is a prime number. For example, if x is 15, it is not a prime number, but ifx is 17, it is a prime number.
Therefore, neither statement 1 nor statement 2 nor statements 1 and 2 taken together is sufficient to determine if x is a prime number.
Answer: Neither statement 1 nor statement 2 nor statements 1 and 2 taken together is sufficient.

Table 34: Examples of Information Essentiality (GPT-3.5 Instance-Based)

第一阶段（由GPT-3.5生成）：我
will use a table to present the information in a structured and concise manner. This will help to clearly outline the sufficiency of each statement and make it easier to compare and analyze the information provided.
以结构化且简洁的方式呈现信息。这将有助于清晰地概述每个陈述的充分性，并使所提供的信息更容易比较和分析。

问题	陈述 1	陈述 2	充分性
大象与斑马	大象是爱丽丝最喜欢的动物。	斑马是爱丽丝最不喜欢的动物。	陈述1或陈述2单独均不充分
Q 和 R	R 是 P 的侄女。	S 有一个姐妹 R。	仅陈述2充分，而仅陈述1不充分
PQRs 和 XYZs	一些 XYZs 是 ABCs。	一些 XYZs 是 PQRs。	仅陈述2充分，而仅陈述1不充分
托尼的生日	娜塔莎正确地陈述，托尼的生日在3月16日之后且3月19日之前。	布伦娜错误地陈述，托尼的生日在 3 月 20 日。	仅陈述 1 充分，而仅陈述 2 不充分
科尔的身高	科尔身高低于7英尺高。	科尔的大学篮球队中每一位球员身高都高于6英尺1英寸。	仅陈述2充分，而仅陈述1不充分

使用该表格格式可以清晰、简洁地比较每个陈述在给定问题中的充分性。

它通过为评估每个中提供的信息提供一个结构化框架，从而提高分析的准确率
问题。

第二阶段 (由 GPT-4 生成): S
tep 1 Analyze Statement 1
陈述 卡兰说，他有一个姐妹。这个陈述告诉我们卡兰有一个姐妹，但没有提供关于珍薇的任何信息。在没有额外信息的情况下，我们无法确定珍薇是否是卡兰的姐妹。步骤2：分析陈述2 珍薇说，她有一个兄弟。这个陈述告诉我们珍薇有一个兄弟，但与陈述1类似，它没有提供关于卡兰的任何信息。在没有额外信息的情况下，我们无法确定卡兰是否是珍薇的兄弟。步骤 3：将陈述1和陈述2结合起来 如果我们把两个陈述放在一起，我们可以推断出卡兰有一个姐妹，珍薇有一个兄弟。如果卡兰和珍薇是兄弟姐妹，并且各自只有一个异性兄弟姐妹，那么可以合乎逻辑地推断出卡兰是珍薇的兄弟，珍薇是卡兰的姐妹。答案：陈述1和陈述2合在一起是充分的。

表33：信息必要性示例（GPT-3.5 ⇒ GPT-4任务型）

第一阶段：
初始状态表示：
- 我们需要确定x是否为质数。 - 我们需要考虑两个
陈述：1. x可被10的因数整除。2. x大于13。

根据已知条件的陈述，让我们单独分析每个陈述：

第二阶段：
陈述1：x可以被10的因数整除。- 仅凭此陈述不足以判断x是否为质数。例如，如果x是10，则它不是质数；但如果x是17，则它是质数。陈述2：x大于13。- 仅凭此陈述也不足以判断x是否为质数。例如，如果x是14，则它不是质数；但如果x是17，则它是质数。综合考虑两个陈述：- 即使将两个陈述综合考虑，我们仍然无法明确判断x是否为质数。例如，如果x是15，则它不是质数；但如果x是17，则它是质数。因此，陈述1、陈述2以及陈述1和2的组合都不足以判断x是否为质数。答案：陈述1、陈述2以及陈述1和2的组合都不充分。

表34：信息必要性示例（GPT-3.5 基于实例）

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

First Stage:
I will use a table to organize the information and determine which statement(s) are sufficient to answer the question.

Problem	Statement 1	Statement 2	Conclusion
1	Sister	Brother	Neither statement 1 nor statement 2 nor statements 1 and 2 taken together is sufficient
2	7.5 hours	Twice as fast	Statement 1 alone is sufficient while statement 2 alone is insufficient
3	Even date	Prime number	Either statement 1 or statement 2 is sufficient
4	U got highest marks, difference between T and U is 11	Difference between S and T is 15, difference between U and V is 9	Statement 1 and statement 2 taken together are sufficient
5	Justin's sister's father's son is Tim	Justin is Tim's brother's father-in-law's only daughter's brother-in-law	Statement 1 alone is sufficient while statement 2 alone is insufficient

Using this table format enhances the accuracy of the analysis by clearly organizing the information and allowing for easy comparison of the sufficiency of each statement. It also provides a structured approach to evaluating the statements for each problem.

Second Stage:
To solve this problem, we need to determine the rate at which the water pump fills the tank. Let's analyze the given statements:
Statement 1: In half an hour, seven-twelfths of the tank is filled.
This statement provides information about the rate at which the tank is being filled. If we know the rate, we can calculate the time required to fill the entire tank.
Statement 2: The pump is switched on at 9:00 a.m.
This statement provides the starting time of the pump, but it does not give any information about the rate at which the tank is being filled.
Analyzing the statements:
- Statement 1 alone provides information about the rate of filling the tank, but it doesn't give the starting time.
- Statement 2 alone does not provide any information about the rate of filling the tank. Conclusion:
From the analysis, we can conclude that Statement 1 alone is sufficient while Statement 2 alone is insufficient to answer the question.
Answer: Statement 1 alone is sufficient while statement 2 alone is insufficient.

Table 35: Examples of Information Essentiality (GPT-3.5 Task-Based)

第一阶段：
我将使用表格来组织信息，并判断哪些陈述足以回答该问题。| 问题 | 陈述 1 |陈述 2 | 结论 |---|---|---|---| 1 | 姐妹 |兄弟 |陈述1、陈述2以及陈述1和2的组合都不充分 | 2 | 7.5 小时| 速度是两倍 | 陈述1单独是充分的，而陈述2单独是不充分的 | 3 | 偶数日期| 质数 | 陈述1或陈述2任一都是充分的 | | 4 | U获得了最高分，T和U之间的差值为 11 | S和T之间的差值为15，U和V之间的差值为 9 | 陈述1和陈述2组合起来是充分的 | 5 | 贾斯汀的姐妹的父亲的儿子是蒂姆 |贾斯汀是蒂姆的兄弟的岳父的唯一女儿的兄弟 | 陈述1单独是充分的，而陈述2单独是不充分的 |使用该表格格式通过清晰地组织信息并便于比较每个陈述的充分性，从而提高了分析的准确率。它还提供了一种结构化方法，用于评估每个问题的陈述。

第二阶段：
为了解决这个问题，我们需要确定水泵向水箱注水的速率。让我们分析给定的陈述：陈述1：在半小时之内，水箱被填满了七分之十二。该陈述提供了关于水箱填充速率的信息。如果我们知道速率，就可以计算填满整个水箱所需的时间。陈述2：水泵在上午9:00启动。该陈述提供了水泵的启动时间，但没有提供关于水箱填充速率的任何信息。分析这些陈述：- 仅陈述1提供了关于水箱填充速率的信息，但没有给出启动时间。- 仅陈述2没有提供关于水箱填充速率的任何信息。结论：通过分析，我们可以得出结论，仅陈述1是充分的，而仅陈述2是不充分的，不足以回答该问题。答案：仅陈述1是充分的，而仅陈述2是不充分的。

表35：信息必要性示例（GPT-3.5基于任务）

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

First Stage (generated by GPT-4):
To enhance clarity and eliminate ambiguities in solving these types of problems, I would suggest using a structured approach that combines logical operators and tabular representation. This method allows for a clear and concise presentation of the information provided by each statement and how they relate to the question at hand.
Structured Approach Description:
1. **Logical Operators**: These are used to represent the relationships and conditions given in the statements. For example, "AND" (&), "OR" (|), "NOT" (!), "IMPLIES" (->), and "EQUIVALENT" (<->) can be used to connect different pieces of information.
Tabular Representation: A table can be used to organize the information and the sufficiency of each statement. The table would have columns for each statement and rows for the conditions or relationships they establish. An additional column can be used to indicate whether the combination of statements is sufficient to answer the question.
Table Format:

```
| Condition/Relationship | Statement 1 | Statement 2 | Combined Statements |
|------------------------|-------------|-------------|---------------------|
| Condition A            | True/False  | True/False  | True/False          |
| Condition B            | True/False  | True/False  | True/False          |
| ...                    | ...         | ...         | ...                 |
| Sufficiency          | Sufficient/Insufficient | Sufficient/Insufficient | Sufficient/Insufficient |
```

How It Enhances Accuracy:
- **Clarity**: The tabular format provides a clear visual representation of the information, making it easier to see which conditions are met by each statement.
- **Organization**: It organizes the logical flow of information, helping to avoid confusion when multiple conditions or relationships are involved.
- **Comparison**: It allows for easy comparison between the sufficiency of individual statements and their combination.
- **Consistency**: Using a consistent method for all problems ensures that the same criteria are applied to determine the sufficiency of the information provided.
By employing this structured approach, we can systematically analyze each problem, ensuring that all necessary conditions are considered and that the conclusion regarding the sufficiency of the statements is reached with precision.

Second Stage (generated by GPT-3.5):
Initial State Representation:
Let's represent the initial state of the problem using a table:

```
| Condition/Relationship | Statement 1 | Statement 2 | Combined Statements |
|------------------------|-------------|-------------|---------------------|
| Circle A smaller than Circle B | ? | ? | ? |
```

Step-by-Step Solution Process:
1. We need to determine if circle A is smaller than circle B. Let's consider each statement individually and then combine them to see if they are individually or collectively sufficient to answer the question.
2. Statement 1: The area of circle B is two times the area of circle A.
- Let's denote the area of circle A as A and the area of circle B as B. According to the statement, B = 2A.
- This statement alone does not provide information about the radii of the circles, so it is insufficient to answer the question.
Updated State Representation:

```
| Condition/Relationship | Statement 1 | Statement 2 | Combined Statements |
|------------------------|-------------|-------------|---------------------|
| Circle A smaller than Circle B | Insufficient | ? | ? |
```

3. Statement 2: The radius of circle B is less than the radius of circle A.
- This statement directly addresses the comparison of the radii of the circles. If the radius of circle B is less than the radius of circle A, then circle A is smaller than circle B.
- This statement alone is sufficient to answer the question.
Updated State Representation:

```
| Condition/Relationship | Statement 1 | Statement 2 | Combined Statements |
|------------------------|-------------|-------------|---------------------|
| Circle A smaller than Circle B | Insufficient | Sufficient | ? |
```

4. Combined Statements:
- When we combine the statements, we can see that statement 2 alone is sufficient to answer the question. Therefore, we don't need to consider the combined sufficiency.
Updated State Representation:

```
| Condition/Relationship | Statement 1 | Statement 2 | Combined Statements |
|------------------------|-------------|-------------|---------------------|
| Circle A smaller than Circle B | Insufficient | Sufficient | N/A |
```

Concluding with the Answer:
Answer: Statement 2 alone is sufficient while statement 1 alone is insufficient.

Table 36: Examples of Information Essentiality (GPT-4 ⇒ GPT-3.5 Task-Based)

第一阶段（由GPT-4生成）：
为了增强清晰度并消除解决此类问题时的歧义，我建议采用一种结构化方法，该方法结合逻辑运算符和表格表示。这种方法允许以清晰简洁的方式呈现每个陈述所提供的信息，以及它们与当前问题的关联。

结构化 A pproach 描述p:
1. **逻辑运算符**：这些用于表示陈述中给出的关系和条件。例如，可以使用“与”（&）、“或”（|）、“非”（!）、“蕴含”（->）和“等价”（<->）来连接不同的信息片段。

表格表示：可以使用表格来组织信息以及每个陈述的充分性。表格将包含表示每个陈述的列，以及表示它们所建立的条件或关系的行。还可以使用额外的列来指示陈述组合是否足以回答问题。

表格 bl格式:

```
| 条件/关系 | 陈述 1 | 陈述 2 | 合并陈述 |
|----------------------|-----------|-----------|-------------------|
| 条件 A                | 真/假      | 真/假      | 真/假              |
| 条件 B                | 真/假      | 真/假      | 真/假              |
| ...                  | ...       | ...       | ...               |
| 充分性                | 充分/不充分 |充分/不充分 | 充分/不充分 |
```

它如何提升准确率:
- **Clarity**: The tabular format 提供了信息的清晰可视化表示，便于查看哪些条件由每条陈述满足。
- **组织**：它整理了信息的逻辑流程，有助于在涉及多个条件或关系时避免混淆。

- **Compar iso n**：它便于对各个陈述的充分性及其组合进行比较。
- **Consistenc y**：对所有问题使用一致的方法，可确保在判定sufficiency of t he information provided。
通过采用这种结构化方法，我们可以系统地分析每个问题，确保所有必要的条件都被考虑，并且关于陈述的充分性的结论能够精确得出。

第二阶段（由GPT-3.5生成）：
初始状态表示:
让我们用表格来表示问题的初始状态:

```
| 条件/关系 | 陈述 1 | 陈述 2 | 合并陈述 |----------------------|-----------|-----------|-------------------
| 圆A小于圆B | ? | ? | ? |
```

逐步求解过程:
1. 我们需要确定圆A是否小于圆B。让我们单独考虑每个陈述，然后将它们合并，以查看它们是否单独或 collectively 足以回答问题。

2. 陈述1：圆B的面积是圆A面积的两倍。
- 令圆A的面积为A，圆B的面积为B。根据陈述，B = 2A。
- 仅凭此陈述无法提供关于圆半径的信息，因此不足以回答问题。更新状态表示:

```
| 条件/关系 |陈述 1 | 陈述 2 | 合并陈述 |
|----------------------|-----------|-----------|-------------------|
| 圆A 小于 圆B | 不充分 | ? | ? |
```

3. 陈述2：圆B的半径小于圆A的半径。
- 该陈述 直接针对两个圆的半径进行比较 . 如果圆B的半径小于圆A的半径，则圆 圆A小于圆B。
- Thi陈述 a lone is suffici答案 t h问题 ion.
更新的状态表示:

```
| 条件/关系 |陈述 1 | 陈述 2 | 合并陈述 |
|----------------------|-----------|-----------|-------------------|
| 圆A小于圆B | 不充分 | 充分 | ? |
```

4. 合并陈述:
- 当我们合并这些陈述时，我们可以看到陈述2本身足以回答问题。因此，我们不需要考虑联合充分性。

更新的状态表示:

```
| 条件/关系 |陈述 1 | 陈述 2 | 合并陈述 |
|----------------------|-----------|-----------|-------------------|
| 圆A 小于 圆B | 不充分 | 充分 | 不适用 |
```

以答案结尾：答案：仅陈述2充分，而仅陈述1不充分。

表36：信息必要性示例（GPT-4 ⇒ GPT-3.5基于任务）

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

First Stage (generated by GPT-4):
For the purpose of enhancing clarity and eliminating ambiguities in the analysis of the given problems, I will describe the use of logical operators and structured conditions in a tabular format. This approach allows for a clear and concise representation of the information provided and the conditions necessary to answer the questions. The tabular format is particularly suitable for these problems as it can systematically organize the given statements and the conditions they need to satisfy to be considered sufficient.
Here's how the tabular format with logical operators can be used:
1. Define the question as a variable that needs to be resolved.
2. List each statement as a separate row in the table.
3. For each statement, define the conditions that must be met for the statement to be sufficient to answer the question.
4. Use logical operators such as AND (&&), OR (||), and NOT (!) to combine conditions when necessary.
5. Determine if each statement alone is sufficient, if they are sufficient when combined, or if they are insufficient even when combined.
An example table structure would look like this:

```
| Statement | Condition(s) to Satisfy | Sufficient Alone (Y/N) | Combined Sufficiency (Y/N) |
|-----------|-------------------------|------------------------|----------------------------|
| Statement 1 | Condition A             | Y/N                    | Y/N                        |
| Statement 2 | Condition B             | Y/N                    | Y/N                        |
```

The "Condition(s) to Satisfy" column would detail what needs to be true for the statement to answer the question. The "Sufficient Alone" column indicates whether the statement by itself is enough to answer the question. The "Combined Sufficiency" column is used to determine if the combination of both statements provides enough information to answer the question.
This structured approach enhances accuracy by:
- Clearly defining what is needed to answer the question.
- Separating each piece of information for individual assessment.
- Allowing for a logical combination of statements to assess their collective sufficiency.
- Providing a visual representation of the logical structure of the problem.
By using this method, one can systematically approach each problem, assess the sufficiency of the information provided, and arrive at a clear conclusion. This format is versatile and can be applied to a wide range of logical and analytical problems.

Second Stage (generated by Gemini Pro):
Initial State Representation:
‘‘‘
Question: On which date of the month was Jayant born in December 2012?
Statement 1: Jayant's birthday is on an even date of the month.
Statement 2: The day of the month on which Jayant was born is a prime number.
Possible Dates: 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31
‘‘‘
Step-by-Step Solution Process:

```
**Step 1:**
* Analyze Statement 1:
    * It provides a condition that Jayant's birthday is on an even date.
    * Even dates in December 2012: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30.
**Step 2:**
* Analyze Statement 2:
    * It provides a condition that Jayant's birthday is on a prime date.
    * Prime dates in December 2012: 3, 5, 7, 11, 13, 17, 19, 23, 29, 31.
**Step 3:**
* Combine the conditions from both statements:
    * Even dates that are also prime: 2, 3, 5, 11, 13, 17, 19, 23, 29, 31.
```

Concluding with the Answer:
* There are 10 possible dates that satisfy both conditions: 2, 3, 5, 11, 13, 17, 19, 23, 29, 31.
Answer: Insufficient information to determine the exact date of Jayant's birth.

Table 37: Examples of Information Essentiality (GPT-4 ⇒ Gemini Pro Task-Based)

第一阶段（由 GPT-4 生成）：
为了增强已知条件问题分析的清晰度并消除歧义，我将以表格格式描述逻辑运算符和结构化条件的使用。该方法能够清晰、简洁地表示所提供的信息以及回答问题所需的条件。表格格式特别适用于这些问题，因为它可以系统地组织已知条件的陈述以及它们需满足的、被视为充分的条件。以下是使用逻辑运算符的表格格式的使用方法：

1. 将问题定义为需要解决的变量。
2. 将每个陈述列为表格中的一行。
3. 对于每个陈述，定义使其足以回答问题的条件。
4. 必要时，使用逻辑运算符如 AND（&&）、OR（||）和 NOT（!）来组合条件。
5. 判断每个陈述单独是否充分，或者当它们组合使用时是否充分，或者即使.....也不充分结合。
例如，表格结构如下所示：

陈述	条件	单独充分（Y/N）	联合充分性（Y/N）
陈述 1	条件 A	Y/N	Y/N
陈述 2	条件 B	Y/N	Y/N

“需满足的条件”列将详细说明为使陈述能够回答问题，哪些内容必须成立。“单独充分”列用于指示该陈述本身是否足以回答问题。“联合充分性”列用于判断两个陈述的组合是否提供了足够的信息来回答问题。

这种结构化方法通过以下方式提高准确率：- 明确界定回答该问题所需的内容。- 将每条信息分开以进行独立评估。- 允许对陈述进行逻辑组合，以评估其整体充分性。- 提供该问题逻辑结构的可视化表示。

通过使用这种方法，可以系统地处理每个问题，评估所提供信息的充分性，并得出明确的结论。该格式具有通用性，可应用于广泛的逻辑与分析问题。

第二阶段（由 Gemini Pro 生成）：
Initial 状态表示：
‘‘‘
问题：2012年12月，贾扬特的生日是几号？
陈述1：贾扬特的生日是当月的偶数日期。
陈述2：贾扬特出生的那天的日期是一个质数。
可能的日期：2，3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31
‘‘‘
逐步求解过程：
第一步：
* 分析陈述1： * 它提供了一个条件，即贾扬特的生日是在一个偶数日期。 * 2012年12月的偶数日期：2、4、6、8、10、12、14、16、18、20、22、24、26、28、30。 **步骤 2：** * 分析陈述2： * 它提供了一个条件，即贾扬特的生日是在一个质数日期。 * 2012年12月的质数日期：3、5、7、11、13、17、19、23、29、31。 **步骤 3：** * 结合两个陈述中的条件： * 既是偶数又是质数的日期：2、3、5、11、13、17、19、23、29、31。 **得出结论答案：** * 有10个可能的日期同时满足这两个条件：2、3、5、11、13、17、19、23、29、31。 **答案：信息不充分，无法确定贾扬特的确切出生日期。**

表37：信息必要性示例（GPT-4 ⇒ Gemini Pro 基于任务）

34
翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

First Stage:
For this type of probability problem, mathematical equations are a suitable form of structured communication. Equations are precise, unambiguous, and universally understood in the context of mathematics, which makes them ideal for conveying the relationships between probabilities of different events. They can be used to solve a wide range of probability problems, including the one presented.
Using mathematical equations, we can represent the information given and the relationships between events A and B. The probability of an event is typically denoted by P(event), and the probability of either event A or event B or both occurring can be represented using the formula for the union of two events:
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Where:
- $P(A \cup B)$ is the probability of either event A or event B or both occurring.
- $P(A)$ is the probability of event A occurring.
- $P(B)$ is the probability of event B occurring.
- $P(A \cap B)$ is the probability of both events A and B occurring.
This equation can be rearranged to solve for $P(B)$ if the other probabilities are known. However, without knowing $P(A \cap B)$, we cannot find the exact value of $P(B)$ using only one of the statements.
The use of mathematical equations in this context enhances accuracy by providing a clear, step-by-step framework for determining the necessary and sufficient conditions to solve for the probability of event B. This method can be applied to any similar problem involving the calculation of probabilities for individual events or their combinations.

Second Stage:
Step 1: Identify the known probabilities from the statements.
- From Statement 1: $P(A) = 0.55$
- From Statement 2: $P(A \cup B) = 0.65$
Step 2: Write down the formula for the probability of the union of two events.
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Step 3: Determine if the information provided is sufficient to find $P(B)$.
- Statement 1 alone gives us $P(A)$, but it does not provide any information about $P(B)$ or $P(A \cap B)$, so it is insufficient.
- Statement 2 alone gives us $P(A \cup B)$, but without additional information about $P(A \cap B)$, we cannot find $P(B)$ using the formula from Step 2, so it is insufficient.
Step 4: Combine the information from both statements to see if they are sufficient together.
- Using Statement 1, we have $P(A) = 0.55$.
- Using Statement 2, we have $P(A \cup B) = 0.65$.
- We can plug these values into the formula from Step 2 to get:
$0.65 = 0.55 + P(B) - P(A \cap B)$
Step 5: Recognize that we still do not have the value for $P(A \cap B)$, which is necessary to solve for $P(B)$.
- Without the probability of the intersection of A and B, we cannot isolate $P(B)$ on one side of the equation.
Step 6: Conclude whether the statements are sufficient to answer the question.
- Since we do not have enough information to calculate $P(B)$, neither Statement 1 nor Statement 2, nor both statements taken together, are sufficient to determine the probability of event B occurring.
Answer: Neither statement 1 nor statement 2 nor statements 1 and 2 taken together is sufficient.

Table 38: Examples of Information Essentiality (GPT-4 Instance-Based)

第一阶段：
对于这类概率问题，数学方程是结构化通信的一种合适形式。方程精确、无歧义，并且在数学背景下被普遍理解，这使它们非常适合传达不同事件概率之间的关系。它们可用于解决各种概率问题，包括这里提出的问题。使用数学方程，我们可以表示给定的信息以及事件A和事件B之间的关系。事件发生的概率通常用 P(事件) 表示，事件A或事件B或两者都发生的概率可以用两个事件的并的公式表示：$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ 其中：
- $P(A \cup B)$ 是事件A或事件B或两者都发生的概率。
- $P(A)$ 是事件A发生的概率。
- $P(B)$ 是事件B发生的概率。
- $P(A \cap B)$ 是事件A和事件B同时发生的概率。

如果其他概率已知，该方程可以重新排列以求解 $P(B)$。然而，如果不知道 $P(A \cap B)$，我们无法仅使用其中一个陈述找到 $P(B)$ 的精确值。在这种背景下使用数学方程，通过提供清晰的逐步框架来确定求解事件B概率的必要充分条件，提高了准确率。此方法可应用于任何涉及单独事件或其组合的概率计算的类似问题。

第二阶段：
第一步：从陈述中识别已知的概率。
- 从陈述1：$P(A) = 0.55$
- 从陈述2：$P(A \cup B) = 0.65$

步骤2：写下两个事件的并的概率的公式。$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ 步骤 3：判断所提供的信息是否充分以找到$P(B)$。
- 仅陈述1给我们$P(A)$，但没有提供有关$P(B)$或$P(A \cap B)$的任何信息，因此它是不充分的。
- 仅陈述2给我们$P(A \cup B)$，但没有关于$P(A \cap B)$的额外信息，我们无法使用步骤2的公式找到$P(B)$，因此它是不充分的。

步骤 4：结合两个陈述的信息，看看它们一起是否充分。
- 使用陈述1，我们有$P(A) = 0.55$。
- 使用陈述2，我们有$P(A \cup B) = 0.65$。
- 我们可以将这些数值代入步骤2的公式，得到：$0.65 = 0.55 + P(B) - P(A \cap B)$

第5步：认识到我们仍然没有$P(A \cap B)$的值，这对于求解$P(B)$是必需的。
- 在没有A和B的交的概率的情况下，我们不可能将$P(B)$单独隔离到方程的一侧。

第6步：得出结论，陈述是否足以回答该问题。
- 由于我们没有足够的信息来计算$P(B)$，陈述1或陈述2，或者将两个陈述合在一起，都不足以确定事件B发生的概率。

答案：陈述1或陈述2，或陈述1和2合在一起都不充分。

表38：信息必要性示例（GPT-4 基于实例）

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

First Stage:
To enhance clarity and eliminate ambiguities in solving these types of problems, I would suggest using a structured approach that combines logical operators and tabular representation. This approach allows for a clear and concise evaluation of the sufficiency of each statement and their combinations.
Structured Approach Description:
1. **Logical Operators**: These are used to represent the relationships between different statements and the information they provide. Logical operators such as AND, OR, and NOT can help determine whether the combination of statements is sufficient to answer the question.
2. **Tabular Representation**: A table can be used to organize the information provided by each statement and to assess their sufficiency. The table will have columns for each statement and rows for the possible outcomes (sufficient/insufficient). This visual representation helps in comparing and combining the information from different statements.
How the Structured Approach Enhances Accuracy:
- **Elimination of Ambiguity**: By using logical operators, the approach clearly defines the relationships between statements, leaving no room for ambiguous interpretations.
- **Visual Clarity**: The tabular format allows for a quick visual assessment of which statements are sufficient on their own or in combination.
- **Consistency**: Applying the same structured approach to each problem ensures that the analysis is consistent across different scenarios.
- **Focus on Relevant Information**: The approach helps to focus only on the information that is relevant to the sufficiency of the statements, avoiding extraneous details.
Example of Structured Approach (Not Solving the Problems):

Problem	Statement 1	Statement 2	S1 Alone	S2 Alone	S1 AND S2	S1 OR S2	Conclusion
Problem 1	Information A	Information B	Sufficient/Insufficient	Sufficient/Insufficient	Sufficient/Insufficient	Sufficient/Insufficient	Final Assessment
Problem 2	Information C	Information D	Sufficient/Insufficient	Sufficient/Insufficient	Sufficient/Insufficient	Sufficient/Insufficient	Final Assessment
...	...	...	...	...	...	...	...

- **S1 Alone**: Evaluate if Statement 1 alone is sufficient to answer the question.
- **S2 Alone**: Evaluate if Statement 2 alone is sufficient to answer the question.
- **S1 AND S2**: Evaluate if the combination of Statement 1 and Statement 2 is necessary and sufficient to answer the question.
- **S1 OR S2**: Evaluate if either Statement 1 or Statement 2 is sufficient to answer the question.
- **Conclusion**: Based on the evaluations, determine the final assessment of sufficiency.
By applying this structured approach to each problem, we can systematically determine the sufficiency of the information provided and arrive at a clear and accurate conclusion.

Second Stage:
Let’s analyze the information provided by each statement to determine if they are sufficient to answer the question: "How many fans went to the stadium yesterday?"
Statement 1: In total, 2343 single-person tickets were bought.
This statement gives us the exact number of single-person tickets that were bought. Assuming each ticket corresponds to one fan, this statement alone tells us that at least 2343 fans went to the stadium.
Statement 2: The stadium can seat 2600 fans.
This statement tells us the maximum capacity of the stadium. However, it does not provide information about the actual number of fans who attended the event yesterday.
Combining Statement 1 and Statement 2:
When we combine both statements, we know that there were 2343 tickets bought and that the stadium can hold up to 2600 fans. This means that the number of fans who could have attended is at least 2343 but no more than 2600. However, we do not have information about whether the stadium was filled to capacity or not.
Conclusion:
- **Statement 1 Alone**: Sufficient to determine the minimum number of fans.
- **Statement 2 Alone**: Insufficient as it does not provide actual attendance.
- **Statement 1 AND Statement 2**: Not necessary to combine since Statement 1 already provides a specific number of fans.
- **Statement 1 OR Statement 2**: Statement 1 is sufficient on its own.
Answer: Statement 1 alone is sufficient while statement 2 alone is insufficient.

Table 39: Examples of Information Essentiality (GPT-4 Task-Based)

第一阶段：
为了提高清晰度并消除解决这些问题类型的歧义，我建议采用一种结合逻辑运算符和表格表示的结构化方法。这种方法允许对每个陈述及其组合的充分性进行清晰而简洁的评估。
结构化A pproach Description:
1. **逻辑运算符**：用于表示不同陈述之间的关系以及它们所提供的信息。诸如“与”、“或”和“非”等逻辑运算符有助于判断陈述的组合是否足以回答该问题。
2. **表格表示**：可以使用表格来组织每个陈述提供的信息，并评估其充分性。该表格将为每个陈述设有列，为可能的结果（充分/不充分）设有行。这种可视化表示有助于比较和整合来自不同陈述的信息。 ### 正面 How the Structured Approach Enhances Accuracy:
- **消除歧义**：通过使用逻辑运算符，该方法清晰地定义了陈述之间的关系，不留任何余地产生歧义解释。
- **视觉清晰度**：表格格式允许快速直观地评估哪些陈述单独或组合起来是充分的。
- **一致性**：对每个问题应用相同的结构化方法，可以确保分析在不同情景下保持一致。
- **聚焦相关信息**：该方法帮助聚焦于与陈述充分性相关的信息，避免无关细节。

结构化方法示例（非解题）：

问题	陈述 1	陈述 2	仅陈述1	仅陈述2	陈述1 和 S2	陈述1 或 S2	结论
问题 1	信息 A	信息 B	充分/不充分	充分/不充分	充分/不充分	充分/不充分	充分/不充分
问题 2	信息 C	信息 D	充分/不充分	充分/不充分	充分/不充分	充分/不充分	充分/不充分
...	...	...	...	...	...	...	...

- **仅陈述1**：评估仅陈述1是否足以回答该问题。
- **仅陈述2**：评估仅陈述2是否足以回答该问题。
- **陈述1 和 陈述2**：评估陈述1和陈述2的组合是否对回答该问题既必要又充分。
- **陈述1或陈述2**：评估陈述1或陈述2是否足以回答该问题。
- **结论**：基于评估结果，确定充分性的最终评估。
通过将这种结构化方法应用于每个问题，我们可以系统地判断所提供信息的充分性，并得出明确而准确的结论。

第二阶段：
让我们分析每个陈述所提供的信息，以确定它们是否足以回答以下问题：“昨天有多少球迷去了体育场？” ### 陈述1：总共购买了2343张单人票。这个陈述提供了所购买的单人票的确切数字。假设每张门票对应一名球迷，这一陈述告诉我们至少有2343名球迷去了体育场。
陈述2：该体育场可容纳2600名球迷。这一陈述告诉我们体育场的最大容量。然而，它没有提供关于实际数球迷昨天结束了该事件。
结合陈述1和陈述2：
当我们将这两项陈述结合起来时，我们知道总共售出了2343张门票，且体育场最多可容纳2600名球迷。这意味着可能有至少2343名、但不超过2600名球迷能够到场观看。然而，我们并不掌握体育场是否满员的信息。

结论：
- **陈述1单独**：足以确定球迷的最小数量。
- **陈述2 单独**：不充分，因为它没有提供实际出席人数。- **陈述1 与 陈述2**：由于陈述1已经提供了具体的球迷人数，无需将两者结合。- **陈述1 或 陈述2**：陈述1本身就足够。答案：陈述1单独足够，而陈述2单独不充分。

表39：信息必要性示例（GPT-4任务型）

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

First Stage:
1. Initial State Representation:

```
- Let x be the total number of people in Maria's office.
- Let y be the number of people in Maria's office who read the newspaper.

Statement 1:
- Half the people in the office don't read the newspaper.
- This can be represented as: x/2 = x - y.

Statement 2:
- 10% of the people in Maria's office read the newspaper.
- This can be represented as: y = 0.1x.
```

Second Stage:

```
**Step 1:**

From Statement 1:
x/2 = x - y

Solving for y:
y = x - x/2
y = x/2

**Step 2:**

From Statement 2:
y = 0.1x

**Step 3:**

Equating the two expressions for y:
x/2 = 0.1x

Solving for x:
x/2 = 0.1x
x = 0.1x * 2
x = 0.2x

**Step 4:**

Solving for x:
x - 0.2x = 0
0.8x = 0
x = 0
```

Conclusion:
Since x (the total number of people in Maria's office) is 0, it is not possible to determine the number of people who read the newspaper.
Answer: Neither statement 1 nor statement 2 nor statements 1 and 2 taken together is sufficient.

Table 40: Examples of Information Essentiality (Gemini Pro Instance-Based)

第一阶段：.
1 初始状态表示:
- 设 x 为玛丽亚办公室的人数总数。- 设 y 为玛丽亚办公室中看报纸的人数。

陈述1：- 办公室里有一半的人不看报纸。- 这可以表示为：x/2 = x - y。

陈述2：- 玛丽亚办公室中10%的人阅读报纸。 - 这可以表示为：y = 0.1x。

第二阶段：

```
**第一步：**

从陈述1：
x/2 =x - y

解出 y：
y =x - x/2
y =x/2

**步骤2：**

从陈述2：
y = 0.1x

**步骤 3：**

令 y 的两个表达式相等：
x/2 = 0.1x

求解 x：
x/2 = 0.1x
x = 0.1x * 2
x = 0.2x

**步骤 4：**

求解 x：
x - 0.2x = 0
0.8x = 0
x = 0

**结论：**
由于 x（玛丽亚办公室的总人数）为 0，无法确定阅读该报纸。
.
答案：陈述1、陈述2或陈述1与陈述2一起均不充分。
```

表40：信息必要性示例（Gemini Pro 基于实例）

37

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

First Stage:
To enhance clarity and eliminate ambiguities inherent in natural language, consider employing more structured and concise forms of communication for your step-by-step solutions. Suitable formats include code, pseudocode, JSON, markdown tables, logical operators, or mathematical equations.
One suitable format is a truth table. A truth table is a table that shows the truth value of a compound statement for all possible combinations of truth values of its component statements. Truth tables can be used to determine whether a statement is a tautology, a contradiction, or neither.
To use a truth table to solve the given problems, we first need to identify the component statements and their truth values. Then, we can construct a truth table to determine the truth value of the compound statement.
For example, to solve the first problem, we can identify the following component statements:
* P: Alexia is a doctor.
* Q: Jenny is a doctor.
We can then construct the following truth table:

P	Q	P ∨ Q
T	T	T
T	F	T
F	T	T
F	F	F

From the truth table, we can see that the compound statement "Either Alexia or Jenny is a doctor" is true in all cases. Therefore, either statement 1 or statement 2 is sufficient to answer the question.
Truth tables can be used to solve all of the given problems. They are a structured and concise way to represent and analyze logical statements.

Second Stage:
Initial State Representation:
```
Classroom: {
  Students: {
    Total: Unknown,
    Failed: 5,
    Ranked: {
      Top: 29,
      Bottom: 16
    }
  }
}
```
Step 1:
Statement 1 tells us that 5 students failed the exam. However, this information alone is not sufficient to determine the total number of students in the classroom.
Step 2:
Statement 2 tells us that there is one student who is 29th from the top and 16th from the bottom. This means that there are 28 students above this student and 15 students below this student. Therefore, the total number of students in the classroom is 28 + 1 + 15 = 44.
Answer:
Statement 2 alone is sufficient while statement 1 alone is insufficient.

Table 41: Examples of Information Essentiality (Gemini Pro Task-Based)

第一阶段：
为增强清晰度并消除自然语言固有的歧义，请在您的分步解决方案中采用更结构化、简洁的通信形式。适用的格式包括代码、伪代码、JSON、Markdown 表格、逻辑运算符或数学方程。
一种合适的格式是真值表。真值表是一种表格，用于展示复合语句在其成分命题的所有可能真值组合下的真值。真值表可用于判断一个陈述是重言式、矛盾式还是两者都不是。
为了使用真值表解决已知条件中的问题，我们首先需要识别成分命题及其真值。然后，我们可以构造一个真值表来确定复合语句的真值。

例如，为了解决第一个问题，我们可以识别出以下成分命题： * P
: Alexia is a d
octor .
* Q: 珍妮是一名医生。
随后，我们可以构造如下真值表：
| P | Q | P ∨
Q |--|--|--| T |
T | T | T |F |
T | F |T | T |
F |F | F |
从真值表中可以看出，复合语句“要么亚历克西亚，要么珍妮是医生”在所有情况下都为真。因此，陈述1或陈述2足以回答该问题。

真值表可以 可用于解决所有已知问题。它们是一种结构化和简洁的方式来表示和分析
逻辑状态 nts.

第二阶段：
初始状态表示：
```教室：{ 学生：{
总数：未知，不及格：
5，排名：{ 最高：
29，最低：16 } } }
```

第一步： 陈述1告诉我们，有5名学生考试不及格。然而，仅此信息不足以确定教室中学生的总数。**步骤2：** 陈述2告诉我们，有一名学生从最高排名是第29名，从最低排名是第16名。这意味着该学生上方有28名学生，下方有
15 28 + 1
+ 15 = 44
名学生。因此，教室中学生的总数为。**答案：** 仅陈述2是充分的，而仅陈述1是不充分的。

表41：信息必要性示例（Gemini Pro 基于任务）

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

First Stage (generated by GPT-3.5):
Unified Representation Form:

```
{
  "Problem": "The problem description",
  "Options": [
    {
      "Name": "Option 1",
      "Description": "Description of the option"
    }, {
      "Name": "Option 2",
      "Description": "Description of the option"
    }, {
      "Name": "Option 3",
      "Description": "Description of the option"
    },
    ...
  ]
}
```

Explanation:
This unified representation form provides a structured and standardized way to represent the problem descriptions and the options for each problem. It allows for clear and consistent documentation of the details of each problem and the available options. By using this representation form, the analysis of the problems can be more accurate as it eliminates the potential for misinterpretation or misunderstanding of the problem descriptions and options. It also allows for easy comparison and evaluation of the options for each problem.

Second Stage (generated by GPT-4):
Initial State Representation:

```
{
  "Problem": "The Straw Hat Theater is experiencing financial discrepancies despite a successful season.
  Bills have not been paid in full, and there is a carry-over balance of $400 from the previous month's
  bills. The Resident Director, Arthur Glendon, is trying to determine the cause of the shortfall and
  who might be responsible.",
  "Options": [
    {
      "Name": "Basil Carmody",
      "Description": "Accountant to the Straw Hat Theater Company, responsible for collecting and
      depositing the theater's receipts."
    }, {
      "Name": "John Franklin",
      "Description": "Treasurer of the Straw Hat Theater Company, responsible for writing checks and
      paying bills."
    }, {
      "Name": "Lawrence Blake",
      "Description": "Head Teller at Vermont Enterprise Bank, handles the theater's account deposits."
    }, {
      "Name": "Martha Gilmont",
      "Description": "Cashier at the Straw Hat Theater Company, responsible for ticket sales and
      record-keeping."
    }
  ]
}
```

Step-by-Step Solution Process:
1. Analyze the information provided about each individual's role and responsibilities in the theater's financial procedures.
2. Basil Carmody collects and deposits the money, and there is no indication of discrepancies in the amounts he deposits.
3. John Franklin writes checks for bills after ensuring there are sufficient funds, and he has reported a lack of funds to pay bills in full.
4. Lawrence Blake processes the deposits and has not reported any irregularities.
5. Martha Gilmont is responsible for ticket sales and record-keeping. She stays at the box office until halfway through the first act and then completes the performance records after the show. She also mentioned that she double and triple checks everything due to her eyesight not being what it used to be.
6. Arthur Glendon notices that the discount coupons are photocopies, not originals, which suggests that someone is using copied coupons to pocket the difference in ticket sales.
7. Since Martha Gilmont is responsible for the box office and the only one who handles the coupons and ticket sales, the evidence points to her as the person responsible for the missing $400.
Answer: 3.Martha Gilmont

Table 42: Examples of Minute Mysteries QA (GPT-3.5 ⇒ GPT-4 Task-Based)

第一阶段（由GPT-3.5生成）：统一表示形式：{ "Problem": "问题描述", "Options": [{ "Name": "选项1", "Description": "选项的描述" }, { "Name": "选项2", "Description": " 选项的描述 " }, { "Name": " 选项 3", "Description": "选项的描述" }, ...]}

解释：
This unified 表 tation f示 id一 tructured and standardized way t个表 t th示 blem descripti ons an d th 每个问题的选项。它允许清晰且一致地记录每个问题的细节以及可用选项。通过使用这种表示形式，对问题的分析可以更加准确，因为它消除了对问题描述和选项的潜在误解或误读。它还便于对每个问题的选项进行比较和评估。第二阶段（由GPT-4生成）：初始状态表示：{ "问题": "草帽剧院尽管度过了一个成功的演出季，但正面临财务差异。账单未全额支付，且从上个月账单中结转了400美元的余额。驻场总监亚瑟·格伦顿试图确定资金短缺的原因以及谁可能对此负责。", "选项": [{ "姓名": "巴斯尔·卡莫迪", "描述": "草帽剧院公司的会计，负责收集和存入剧院的收入。" }, { "姓名": "约翰·富兰克林", "描述": "草帽剧院公司的财务主管，负责开具支票和支付账单。" }, { "姓名": "劳伦斯·布莱克", "描述": "佛蒙特企业银行的首席柜员，处理剧院的账户存款。" }, { "姓名": "玛莎·吉尔蒙特", "描述": "草帽剧院公司的收银员，负责售票和记录保存。" }]} 逐步求解过程：1. 分析有关每个个人在剧院财务程序中的角色和职责所提供的信息。2. 巴斯尔·卡莫迪收集并存储款项，没有迹象表明他所存的款项存在差异。3. 约翰·富兰克林在确保有足够的资金后为账单开具支票，并报告资金不足，无法全额支付账单。4. 劳伦斯·布莱克处理存款，没有报告任何异常。5. 玛莎·吉尔蒙特负责售票和记录保存。她会在第一幕进行到一半时留在售票处，并在演出结束后完成演出记录。她还提到，由于她的视力不如从前，她会反复两次和三次检查所有内容。6. 亚瑟·格伦顿注意到折扣券是复印件，而不是原件，这表明有人在用复制的券来窃取售票差额。7. 由于玛莎·吉尔蒙特负责售票处，并且是唯一处理优惠券和售票的人，证据指向她是对缺失的400美元负责的人。答案：3.玛莎·吉尔蒙特

表42：Minute Mysteries 问答示例（GPT-3.5 ⇒ GPT-4任务型）

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

First Stage:
I will use a structured approach to analyze the alibis and evidence to determine the thief. I will create a table with the alibis and evidence for each suspect, and then use logical operators to narrow down the possibilities.

```
| Suspect | Alibi | Evidence |
|---------|-------|----------|
| Denise  | Broke her leg skiing in Aspen, Colorado | N/A |
| Jim     | Visited the International UFO Museum in Roswell, New Mexico | Returned from Long Beach,
California |
| Juan    | Spent the weekend in Long Beach, California learning to surf | N/A |
| Skye    | Scuba diving in Cozumel, Mexico, and had a delayed flight back | N/A |
```

Using the evidence of the sand from a different location and the timing of the theft, we can eliminate Jim and Skye as suspects, as their alibis and evidence do not match the evidence from the crime scene. Denise's alibi of breaking her leg skiing in Aspen, Colorado also matches the evidence of the sand from a different location, so she can be eliminated as a suspect.
This leaves Juan as the only remaining suspect, as his alibi of spending the weekend in Long Beach, California matches the evidence of the sand from a different location. Therefore, Juan is the likely thief.

Second Stage:
Initial State Representation:
The suspects are Denise, Jim, Juan, and Skye. The evidence includes the theft occurring when the alarm was deactivated and reactivated, the presence of sand from a different location, and the timing of the theft. The alibis of the suspects are related to their weekend activities.
Step-by-Step Solution Process:

```
1. Evaluate the alibis and evidence for each suspect:
   - Denise: Broke her leg skiing in Aspen, Colorado.
   - Jim: Visited the International UFO Museum in Roswell, New Mexico and returned from Long Beach,
   California.
   - Juan: Spent the weekend in Long Beach, California learning to surf.
   - Skye: Scuba diving in Cozumel, Mexico, and had a delayed flight back.

2. Use the evidence of the sand from a different location and the timing of the theft to eliminate
suspects:
   - Jim and Skye's alibis and evidence do not match the evidence from the crime scene, so they can
   be eliminated as suspects.
   - Denise's alibi of breaking her leg skiing in Aspen, Colorado matches the evidence of the sand
   from a different location, so she can be eliminated as a suspect.

3. Based on the remaining evidence and alibis, Juan is the only remaining suspect whose alibi matches
the evidence of the sand from a different location.
```

Concluding with the Answer:
Answer: 2. Juan Carde

Table 43: Examples of Minute Mysteries QA (GPT-3.5 Instance-Based)

第一阶段：
I will采用结构化方法来分析不在场证明和证据，以确定小偷。我将创建一张表格，列出每个嫌疑人的不在场证明和证据，然后使用逻辑运算符来缩小可能性。

| 嫌疑人 | 不在场证明 |证据 |---|---|---|---| 丹尼斯| 在科罗拉多州阿斯彭滑雪时摔断了腿 | 不适用 | 吉姆| 访问了新墨西哥州罗斯威尔的国际UFO博物馆 | 从加利福尼亚州长滩返回 | 胡安| 在加利福尼亚州长滩度过周末学习冲浪 | 不适用 | 斯凯| 在墨西哥科苏梅尔潜水，返程航班延误 | 不适用 |

利用来自不同地点的沙土这一证据以及盗窃的时间，我们可以排除吉姆和斯凯作为嫌疑人，因为他们的不在场证明和证据与来自犯罪现场的证据不匹配。丹尼斯在科罗拉多州阿斯彭滑雪时摔断腿的不在场证明也与来自不同地点的沙土证据相符，因此她也可以被排除。这样就只剩下胡安作为唯一的嫌疑人，因为他在加利福尼亚州长滩度过周末的不在场证明与来自不同地点的沙土证据相符。因此，胡安很可能是小偷。

第二阶段：
初始状态表示:
嫌疑人包括丹尼斯、吉姆、胡安和斯凯。证据包括警报器被停用后重新激活期间发生的盗窃、来自不同地点的沙土的存在，以及盗窃的时间。嫌疑人的不在场证明与其周末活动有关。
逐步求解过程:
1. 评估每位嫌疑人的不在场证明和证据：- 丹尼斯：在科罗拉多州阿斯彭滑雪时摔断了腿。- 吉姆：访问了新墨西哥州罗斯威尔的国际UFO博物馆，并从加利福尼亚州长滩返回。- 胡安：周末在加利福尼亚州长滩学习冲浪。- 斯凯：在墨西哥科苏梅尔潜水，回程航班延误。

2. 利用来自不同地点的沙土证据和盗窃的时间来排除嫌疑人：

- 吉姆和斯凯的不在场证明与证据同犯罪现场的证据不匹配，因此可以排除他们作为嫌疑人的嫌疑。 - 丹尼斯关于在科罗拉多州阿斯彭滑雪时摔断腿的不在场证明与来自不同地点的沙土这一证据相符，因此可以排除她作为嫌疑人的嫌疑。

3. 根据剩余的不在场证明和证据，胡安是唯一一名其不在场证明与来自不同地点的沙土这一证据相符的剩余嫌疑人。

以答案结尾：答案：2. 胡安·卡
德

表43: Minute Mysteries 问答示例（GPT-3.5 基于实例）

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

First Stage:
Unified Representation Form:

```
{
  "Problem": "The problem description",
  "Suspects": [
    {
      "Name": "Name of the suspect",
      "Motive": "Suspect's possible motive for the crime",
      "Alibi": "Suspect's alibi or whereabouts at the time of the crime"
    }, {
      "Name": "Name of the suspect",
      "Motive": "Suspect's possible motive for the crime",
      "Alibi": "Suspect's alibi or whereabouts at the time of the crime"
    },
    ...
  ]
}
```

Explanation:
This unified representation form captures the essential elements of each problem, including the problem description and the details of the suspects involved. By structuring the information in this way, it allows for a clear and organized analysis of the problems. It also enables easy comparison of the motives and alibis of the suspects across different scenarios, enhancing the accuracy of the analysis. This form provides a standardized way to represent the key details of each problem, making it easier to identify patterns and make logical deductions.

Second Stage:
Initial State Representation:

```
{
  "Problem": "Mick, a homeless man, is at a bus stop with the narrator, Mr. Storimon, and Danny. Mr.
  Storimon's wallet falls out of his pocket, and when he realizes it's missing, he accuses Mick of
  stealing it. Later, Mr. Storimon discovers that his $100 bills are missing from the wallet.",
  "Suspects": [
    {
      "Name": "Mick",
      "Motive": "Possible motive for stealing the wallet and money",
      "Alibi": "His whereabouts and actions during the incident"
    }, {
      "Name": "Mr. Storimon",
      "Motive": "Possible motive for falsely accusing Mick and hiding the truth about the money",
      "Alibi": "His actions and whereabouts during the incident"
    }, {
      "Name": "Danny",
      "Motive": "Possible motive for stealing the money",
      "Alibi": "His actions and whereabouts during the incident"
    }, {
      "Name": "Policeman",
      "Motive": "Possible motive for stealing the money",
      "Alibi": "His actions and whereabouts during the incident"
    }
  ]
}
```

Step-by-Step Solution Process:
1. Mick's motive for stealing the wallet and money is not clear. His alibi is that he was at the bus stop and did not steal the wallet.
2. Mr. Storimon's motive for falsely accusing Mick and hiding the truth about the money could be to shift the blame. His alibi is that he was at the bus stop and lost his wallet.
3. Danny's motive for stealing the money is not clear. His alibi is that he arrived at the bus stop on his skateboard and did not steal the money.
4. The policeman's motive for stealing the money is not clear. His alibi is that he was at the bus stop and interacted with the other characters.
Based on the information provided, it is evident that Mick, Mr. Storimon, and Danny were at the bus stop when the incident occurred. The policeman also interacted with the characters. Mick's alibi seems to be the most solid, as he was present at the bus stop and did not steal the wallet. Mr. Storimon's motive for falsely accusing Mick and hiding the truth about the money raises suspicion. Danny's motive is unclear, but his alibi seems plausible. The policeman's motive is also unclear, but his alibi is that he was at the bus stop.
Concluding with the Answer:
Answer: 2. Mr. Storimon

Table 44: Examples of Minute Mysteries QA (GPT-3.5 Task-Based)

第一阶段：统一表示形式：{ "问题": "问题描述", "嫌疑人": [{ "姓名": "嫌疑人姓名", "动机": "嫌疑人可能的犯罪动机", "不在场证明": "嫌疑人案发时的不在场证明或行踪" }, { "姓名": "嫌疑人姓名", "动机": "嫌疑人可能的犯罪动机", "不在场证明": " 嫌疑人案发时的不在场证明或行踪 " }, ...]}

解释：
这种统一表示形式捕捉了每个问题的核心要素，包括问题描述以及所涉嫌疑人的细节。通过以这种方式对信息进行结构化，它能够实现对问题清晰且有条理的分析。它还便于在不同情景中比较嫌疑人的动机和不在场证明，从而提升分析的准确率。该形式提供了一种标准化的方式来表示每个问题的关键细节，使识别模式和进行逻辑推理变得更加容易。

第二阶段：
初始状态表示：{ "问题": "米克，一个无家可归的人，与叙述者、斯托里蒙先生和丹尼一起在一个公交车站。斯托里蒙先生的钱包从口袋里掉了出来，当他意识到钱包不见了时，他指控米克偷了钱包。后来，斯托里蒙先生发现他钱包里的100美元纸币不见了。", "嫌疑人": [{ "姓名": "米克", "动机": "偷窃钱包和钱的可能动机", "不在场证明": "事件发生期间他的行踪和动作" }, { "姓名": "斯托里蒙先生", "动机": "错误指控米克并对钱款真相隐瞒的可能动机", "不在场证明": "事件发生期间他的动作和行踪" }, { "姓名": "丹尼", "动机": "偷钱的可能动机", "不在场证明": "事件发生期间他的动作和行踪" }, { "姓名": "警察", "动机": "偷钱的可能动机", "不在场证明": "事件发生期间他的动作和行踪" }]} 逐步求解过程：1. 米克偷窃钱包和钱的动机尚不明确。他的不在场证明是他当时在公交车站，并没有偷钱包。2. 斯托里蒙先生错误指控米克并对钱款真相隐瞒的动机可能是为了转移责任。他的不在场证明是他当时在公交车站并且丢失了钱包。3. 丹尼偷钱的动机尚不明确。他的不在场证明是他滑着滑板来到公交车站，并且没有偷钱。4. 警察偷钱的动机尚不明确。他的不在场证明是他当时在公交车站并与其他角色互动。根据所提供的信息，显然在事件发生时，米克、斯托里蒙先生和丹尼都在公交车站。警察也与这些角色有过互动。米克的不在场证明似乎最坚实，因为他当时在公交车站，且没有偷钱包。斯托里蒙先生错误指控米克并对钱款真相隐瞒的动机令人怀疑。丹尼的动机尚不明确，但他的不在场证明似乎可信。警察的动机也不明确，但他的不在场证明是他当时在公交车站。以答案结尾：答案：2. 斯托里蒙先生

表44：Minute Mysteries 问答示例（GPT-3.5基于任务）

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

First Stage (generated by GPT-4):
Unified Representation Form:

```json
{
  "scenario_description": "String",
  "suspects": [
    {
      "name": "String",
      "alibi": "String",
      "evidence_against": "String",
      "unique_detail": "String"
    }
  ],
  "crime_details": {
    "time": "String",
    "location": "String",
    "object_stolen": "String",
    "additional_clues": "String"
  },
  "investigator_notes": "String",
  "solution": {
    "guilty_suspect": "String",
    "justification": "String"
  }
}
```

Explanation:
This JSON representation form enhances the accuracy of analysis by providing a structured and standardized way to capture all relevant details of a mystery or crime scenario. Each problem is broken down into key components such as the scenario description, a list of suspects with their alibis and unique details, specific crime details, notes from the investigator, and the solution with justification for the guilty suspect.
- The 'scenario_description' field provides a brief overview of the context or setting of the problem.
- The 'suspects' array contains objects for each suspect, detailing their name, alibi, any evidence against them, and any unique detail that might be relevant to the case.
- The 'crime_details' object captures the specifics of the crime, including the time it occurred, the location, what was stolen, and any additional clues that were found.
- The 'investigator_notes' field allows for any additional observations or insights from the investigator that may not fit neatly into the other categories.
- The 'solution' object provides a space to name the guilty suspect and the justification for their guilt, based on the analysis of the information provided.
By using this form, an analyst can ensure that no critical information is overlooked and that each aspect of the problem is considered. The structured nature of JSON allows for easy comparison between different problems and can be parsed programmatically, which could be useful for automated reasoning systems or databases. It also ensures consistency in the way information is recorded, which is crucial for accurate analysis and comparison across multiple similar problems.

Second Stage (generated by GPT-3.5):
Unified Representation Form:

```json
{
    ......
}
```

Solution Process:
1. John, John's wife, Bo, and Nathan Drew are the suspects.
2. Each suspect had the opportunity to leave the gate untied.
3. John's wife confirmed seeing several knots at the gate.
4. Bo has a badly crippled left hand, making it difficult for him to guard the gate effectively.
5. Nathan Drew didn't check the knots during his watch.
6. The spy is suspected to have left the gate unlocked, potentially endangering the fort.
Answer: 3. Nathan Drew

Table 45: Examples of Minute Mysteries QA (GPT-4 ⇒ GPT-3.5 Task-Based)

第一阶段（由 GPT-4 生成）：
统一表示形式：```json { "场景_描述": "字符串", "嫌疑人": [{ "姓名": "字符串", "不在场证明": "字符串", "证据_不利": " 字符串 ", " 独特 _ 细节 ": " 字符串 " }],"犯罪_细节": { "时间": "字符串", "地点": "字符串", "物品_被窃": "字符串", "额外_线索": "字符串" }, "调查员_笔记": "字符串", "解决方案": { "有罪_嫌疑人": "字符串", "理由": "字符串" } }、、、

解释：
这种JSON表示形式通过提供一种结构化和标准化的方式来捕捉神秘或犯罪情景的所有相关细节，从而提升了分析的准确率。每个问题被分解为关键组成部分，例如情景描述、包含各嫌疑人不在场证明及独特细节的嫌疑人列表、具体的犯罪细节、调查员的笔记，以及针对有罪嫌疑人的解决方案及其理由。
"_场景描述"字段简要概述了问题背景或设置。
- "suspects"数组包含每个嫌疑人的对象，详细说明其姓名、不在场证明、针对他们的任何证据，以及与案件可能相关的任何独特细节。

- "crime_details"对象记录了犯罪的具体细节，包括发生时间、地点、被盗物品以及发现的其他线索。

- "investigator_notes"字段用于记录调查员可能无法归入其他类别的额外观察或见解。

- The 'solution' 该对象提供一个空间，用于根据分析结果命名有罪嫌疑人及其有罪的理由
信息 提供。
通过使用这种格式，分析师可以确保不会遗漏任何关键信息，并且问题中的每个方面都得到考虑。JSON的结构化特性使得不同问题之间的比较更加便捷，并且可以被程序解析，这对于自动化推理系统或数据库可能非常有用。它还确保了信息记录方式的一致性，这对于多个相似问题的准确分析和比较至关重要。

第二阶段（由 GPT-3.5 生成）： U
nified R epresentat ion Form:
```json { .....
}、、、

解决方案过程：
1
约翰、约翰的妻子、博和纳森·德鲁是嫌疑人。2. 每个嫌疑人都有机会让大门保持未系状态。3. 约翰的妻子证实曾看到大门上有好几个绳结。4. 博的左手严重残废，使他难以有效地看守大门。5. 纳森·德鲁在值班期间没有检查绳结。6. 该间谍被怀疑将大门留作未上锁状态，可能危及堡垒的安全。答案：3。纳森·德鲁

表45：Minute Mysteries 问答示例（GPT-4 ⇒ GPT-3.5基于任务）



翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next
```

First Stage (generated by GPT-4):
Unified Representation Form: JSON
Explanation:
JSON (JavaScript Object Notation) is a lightweight data-interchange format that is easy for humans to read and write, and easy for machines to parse and generate. It is based on a subset of the JavaScript Programming Language and is a text format that is completely language independent. JSON is ideal for representing the state or conditions of the problems provided because it allows for a structured representation of the problem elements, such as characters, locations, events, and clues, in a way that is both human-readable and machine-parsable.
Using JSON to represent the problems enhances the accuracy of analysis in several ways:
1. **Structured Data**: JSON provides a clear structure with key-value pairs, making it easy to organize the information in a consistent manner. This helps in avoiding ambiguity and ensures that each aspect of the problem is addressed.
2. **Scalability**: JSON can represent simple to complex structures, which means it can handle the varying complexity of the problems without requiring changes to the representation format.
3. **Interoperability**: JSON is widely supported across different programming languages and platforms, making it easy to use the same representation to interface with various analytical tools or algorithms.
4. **Flexibility**: JSON objects can contain nested arrays and objects, allowing for a hierarchical representation of information that can mirror the logical relationships within the problem.
5. **Clarity**: The use of JSON can help separate the data (problem conditions) from the logic used to solve the problem, leading to clearer reasoning paths and more accurate conclusions.
Here is an example of how a problem might be represented in JSON without solving it or including specific details:

```json
{
  "characters": [
    {"name": "Character1", "attributes": ["attribute1", "attribute2"]},
    {"name": "Character2", "attributes": ["attribute1", "attribute2"]}
  ],
  "locations": [
    {"name": "Location1", "attributes": ["attribute1", "attribute2"]},
    {"name": "Location2", "attributes": ["attribute1", "attribute2"]}
  ],
  "events": [
    {"name": "Event1", "attributes": ["attribute1", "attribute2"]},
    {"name": "Event2", "attributes": ["attribute1", "attribute2"]}
  ],
  "clues": [
    {"description": "Clue1", "location": "Location1", "relatedTo": ["Character1"]},
    {"description": "Clue2", "location": "Location2", "relatedTo": ["Character2"]}
  ],
  "questions": [
    {"text": "Question1", "type": "multipleChoice", "options": ["Option1", "Option2"]}
  ]
}
```

In this representation, "attribute1", "attribute2", "Character1", "Location1", "Event1", "Clue1", "Question1", and "Option1" are placeholders for actual data that would be filled in for each specific problem. The JSON structure remains the same, providing a unified form for representing any of the problems.

Second Stage (generated by Gemini Pro):
Initial State Representation (JSON):

```json
{  ......  }
```

Step-by-Step Solution Process:
1. **Analyze the Clues:**
  - The clues suggest that everyone in the room is a liar, as they make statements that contradict their actions or appearances.
  - The focus of the mystery is on the stolen dice, which have sentimental value to Debra's father.
2. **Consider the Statements Made by the Characters:**
  - Debra's father initially denies knowing the narrator is a mystery solver, but later confronts him about it, indicating that he lied.
  - Olivia and Luke make contradictory statements about their weight and health, suggesting they are lying.
3. **Examine the Behavior of the Characters:**
  - The butler's careful movements and lack of noise suggest he is trying to avoid attention.
  - The fact that the dice disappear during a power outage raises suspicion.
4. **Identify the Most Likely Thief:**
  - The butler is the most likely thief because he had the opportunity to take the dice during the power outage.
  - The other characters' lies are likely attempts to distract attention from the butler.
Concluding with the Answer:
Answer: 3. The Butler

Table 46: Examples of Minute Mysteries QA (GPT-4 ⇒ Gemini Pro Task-Based)

第一阶段（由 GPT-4 生成）：
统一表示形式：JSON 说明：

JSON（JavaScript 对象表示法）是一种轻量级数据交换格式，易于人类阅读和编写，也便于机器解析和生成。它基于 JavaScript 编程语言的一个子集，是一种完全语言无关的文本格式。JSON 非常适合表示所提供问题的状态或条件，因为它允许以人类可读且机器可解析的结构化表示形式来描述问题的各个元素，例如角色、地点、事件和线索。

使用 JSON 来表示问题，在以下几个方面提高了分析的准确率：
1. **结构化数据**：JSON 提供带有键值对的清晰结构，便于以一致的方式组织信息。这有助于避免歧义，并确保问题的每个方面都得到处理。
2. **可扩展性**：JSON 可以表示从简单到复杂的结构，这意味着它能够应对问题的不同复杂程度，而无需更改表示形式。

3. **互操作性**：JSON在不同编程语言和平台上得到广泛支持，使得使用相同的表示方式来与各种分析工具或算法进行接口交互变得轻而易举。
4. **灵活性* *：JSON对象可以包含嵌套数组和对象，从而允许对信息（informatio）进行层次化表示，使其能够反映问题的逻辑关系。
5. **清晰度**：使用JSON有助于将数据（问题条件）与用于解决问题的逻辑分离开来，从而带来更清晰的推理路径和更准确的结论。

Here 是一个示例，说明如何在未解决问题或不包含具体细节的情况下使用 JSON 表示问题：

```json { "characters": [{"name": "角色1", "attributes": ["属性1", "属性2"]}, {"name": "角色2", "attributes": ["属性1", "属性2"]},"locations": [{"name": "地点1", "attributes": ["属性1", "属性2"]}, {"name": "地点2", "attributes": ["属性1", "属性2"]},"events": [{"name": "事件1", "attributes": ["属性1", "属性2"]}, {"name": "事件2", "attributes": ["属性1", "属性2"]},"clues": [{"description": "线索1", "location": "地点1", "relatedTo": ["角色1"]}, {"description": "线索2", "location": "地点2", "relatedTo": ["角色2"]},"questions": [{"text": "问题1", "type": "multipleChoice", "options": ["选项1", "选项2"]}]}、、、
```

在该表示中，“attribute1”、“attribute2”、“Character1”、“Location1”、“Event1”、“Clue1”、“Question1”和“Option1”是占位符，用于在解决每个具体问题时填入实际数据。JSON结构保持不变，为表示任何一个问题提供统一的格式。

第二阶段（由 Gemini Pro 生成）：
初始状态表示（JSON）：

```json { ..... }
```

逐步求解过程：
1. **分析线索：** - 线索表明，房间里的每个人都在说谎，因为他们的陈述与他们的动作或外表相矛盾。 - 这个神秘事件的焦点是那些被盗骰子，它们对黛布拉的父亲具有情感价值。 2. **考虑角色们做出的陈述：** - 黛布拉的父亲最初否认知道叙述者是一个解谜者，但后来就此事与他当面对质，表明他撒了谎。 - 奥利维亚和卢克就他们的体重和健康状况做出了相互矛盾的陈述，表明他们在撒谎。 3. **审视角色们的行为：** - 男仆小心翼翼的动作和没有发出噪音，表明他试图避免引起注意。 - 骰子在停电期间消失这一事实引发了怀疑。 4. **确定最有可能的小偷：** - 男仆是最有可能的小偷，因为他在停电期间有机会拿走骰子。 - 其他角色的谎言很可能是为了转移人们对男仆的注意力。 **以答案结尾：** 答案：3. 男仆

表 46：Minute Mysteries 问答示例（GPT-4 ⇒ Gemini Pro 基于任务）

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

First Stage:
To enhance clarity and eliminate ambiguities in solving the problem of who stole the chess pieces, I will describe the use of logical operators and a markdown table. These tools can help structure the information given in the narrative and allow for a systematic analysis of the facts to deduce the culprit.
Logical Operators:
Logical operators such as AND, OR, and NOT can be used to combine different conditions that must be met for a suspect to be considered the thief. For example, if a suspect had the opportunity to steal the pieces (Opportunity = True) AND had a motive to stop the chess boxing fights (Motive = True), then they could be considered a potential thief (Suspect = True).
Markdown Table:
A markdown table can be used to organize information about each character in a structured way. Each row represents a character, and each column represents a key piece of information relevant to the mystery (e.g., Opportunity, Motive, Alibi, Actions). By filling in the table with information from the narrative, patterns may emerge that point to the likely thief.
Here's an example of how the markdown table might be structured:

```
| Character | Opportunity | Motive | Alibi | Actions | Suspect |
|-----------|-------------|--------|-------|---------|---------|
| Ernie     | ?           | ?      | ?     | ?       | ?       |
| Gordon    | ?           | ?      | ?     | ?       | ?       |
| Jesse     | ?           | ?      | ?     | ?       | ?       |
| Mac       | ?           | ?      | ?     | ?       | ?       |
```

By filling in the table with details from the story, we can systematically evaluate each character's likelihood of being the thief. The markdown table provides a clear visual representation of the information, making it easier to compare suspects and identify inconsistencies or patterns that could lead to a solution.
Using these tools, the analysis can be conducted in a step-by-step manner, ensuring that each piece of evidence is considered and that the conclusion is based on a logical deduction rather than speculation or assumption.

Second Stage:
Initial State Representation:
Let's organize the information we have about each character in the markdown table format:

```
| Character | Opportunity | Motive | Alibi | Actions | Suspect |
|-----------|-------------|--------|-------|---------|---------|
| Ernie     | Yes         | Yes    | No    | Studying chess openings | ? |
| Gordon    | Yes         | No     | Yes   | Joking, training Ernie | ? |
| Jesse     | Yes         | No     | No    | Studying chess openings | ? |
| Mac       | Yes         | No     | No    | Boxing, not talking | ? |
```

Opportunity: All characters were present at the event and could have had the opportunity to steal the pawns.
Motive: Ernie has a motive to avoid boxing Mac, but it's not clear if others have a motive.
Alibi: Gordon was watching the door, so he has an alibi.
Actions: Ernie and Jesse were studying chess openings, Mac was boxing, and Gordon was joking and training Ernie.
Suspect: We need to determine who is most likely to have stolen the pawns based on the information provided.
Step-by-Step Solution Process:
1. The referee was sure that Gordon did not enter the room with the chessboard through the door, which gives Gordon an alibi for the time the pawns were stolen.
2. Ernie was busy studying chess openings and preparing for his match, which would make it unlikely for him to risk stealing the pawns and jeopardizing his chance to compete.
3. Mac, the real boxer, was involved in the boxing match and had just finished a fight, making it unlikely for him to have the opportunity to steal the pawns without being noticed.
4. Jesse was the one who discovered the pawns were missing, which could either mean he is innocent and just reporting the fact, or he is using the discovery as a way to deflect suspicion from himself.
Considering the above points, Jesse had the opportunity to steal the pawns when he went to get the chessboard. He had no alibi since he was not being watched at that moment. The other characters either had an alibi or were engaged in activities that made it unlikely for them to steal the pawns.
Concluding with the Answer:
Answer: 2.Jesse

Table 47: Examples of Minute Mysteries QA (GPT-4 Instance-Based)

第一阶段：
为增强清晰度并消除"谁偷了棋子"这一问题的歧义，我将描述如何使用逻辑运算符和 Markdown 表格。这些工具有助于对叙事中给出的信息进行结构化，并允许对事实进行系统性分析，以推断出嫌疑人。

Logical Op算子：
可以使用"与"、"或"、"非"等逻辑运算符来组合必须满足的不同条件，以判断某位嫌疑人是否为小偷。例如，如果某位嫌疑人有机会偷走棋子（机会 = True），并且有动机阻止象棋搏击比赛（动机 = True），那么他们就可能被视为潜在的小偷（嫌疑人 = True）。Markdown 表格：
Markdown 表格可以用于以结构化的方式组织每个角色的信息。每一行代表一个角色，每一列代表与神秘案件相关的关键信息（例如，机会、动机、不在场证明、动作）。通过用叙事中的信息填充表格，可能会显现出指向最可能的小偷的模式。 H

ere'作为一个例子 le of how th标记 kdown ta ble might be结构 d:

```
| 角色   | 机会 |动机  | 不在场证明 | 动作  | 嫌疑人 |
|---------|-----------|-------|------|-------|--------|
| 厄尼    | ?         | ?     | ?    | ?     | ?      |
| 戈登    | ?         | ?     | ?    | ?     | ?      |
| 杰西    | ?         | ?     | ?    | ?     | ?      |
| Mac     | ?         | ?     | ?    | ?     | ?      |
```

通过用故事中的细节填充表格，我们可以系统地评估每个角色是小偷的可能性。Markdown 表格提供了信息的清晰可视化表示，使得比较嫌疑人并识别可能导致解决方案的不一致之处或模式变得更加容易。

使用这些工具，分析可以以逐步的方式进行，确保考虑每一份证据，并确保结论基于逻辑推理而非猜测或假设。

第二阶段：
初始状态表示：
让我们整理这些信 mation 关于 Markdown 表格中每个角色的信息 格式：

```
| 角色   | 机会 |动机  | 不在场证明 | 动作  | 嫌疑人 |
|---------|---|-------|------|-------|---|
| 厄尼    | 是 | 是的  | No   | 是的  | 研究国际象棋开局 | ? |
| 戈登    | 是的 | No  | 是的 | 开玩笑，训练 厄尼 | ? |
| 杰西    | 是的 | No  | No   | 研究国际象棋开局 | ? |
| 麦克    | 是 | No  | No   | 拳击， 不是谈话 | ? |
```

机会：所有角色 都出席了该事件，并且可能有机 会偷走兵。
动机：厄尼有动机避免与麦克进行拳击，但不清楚其他人是否有动机。
不在场证明：戈登在看门，因此他有不在场证明 .
Actions：厄尼和杰西在研究国际象棋开局，麦克在拳击，戈登在开玩笑并训练厄尼 e.
嫌疑人：我们需要根据所提供的信息确定谁最有可能偷走了兵。S
tep-by-Step Solution Process:
1. 裁判确信戈登没有通过带棋盘的门口进入房间，这为戈登在兵被偷的时间段提供了不在场证明。

2. 厄尼正忙着研究国际象棋开局并为他的比赛做准备，这使得他不太可能会冒险偷走兵，从而危及自己参赛的机会。

3 麦克，这位真正的拳 击手，参加了拳击比赛并刚刚打完一场，这使得他不太可能有
opp会 y to 在不被发现的情况下偷走兵。
4. 杰西是发现兵丢失的人，这可能意味着他是无辜的，只是在陈述事实；或者他是在利用这一发现来转移对自身的怀疑。

鉴于上述几点，杰西在他去取棋盘时有机会偷走兵。由于当时无人监视他，他没有不在场证明。其他角色要么有不在场证明，要么正在从事不太可能偷走兵的活动。

以答案结尾：
答案：2. 杰西

表47：Minute Mysteries 问答示例（GPT-4 基于实例）

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

First Stage:
Unified Representation Form: JSON (JavaScript Object Notation)
Explanation:
JSON is a lightweight data-interchange format that is easy for humans to read and write, and easy for machines to parse and generate. It is based on a subset of the JavaScript Programming Language and is a text format that is completely language independent. JSON is ideal for representing the state or conditions of the problems described because it can structure data in a hierarchical manner, which is particularly useful for representing complex scenarios with multiple entities and attributes.
Using JSON to represent the problems enhances accuracy in the following ways:
- **Standardization**: JSON provides a consistent format for representing the various elements of each problem, such as characters, locations, items, and events. This standardization reduces ambiguity and ensures that each aspect of the problem is clearly defined.
- **Hierarchy and Nesting**: JSON's ability to nest objects within other objects allows for a clear representation of the relationships between different elements of the problem, such as the location hierarchy (house > room > closet) or the sequence of events.
- **Key-Value Pairs**: JSON's key-value pairs enable precise labeling of each piece of information, which helps in accurately identifying and retrieving data for analysis.
- **Flexibility**: JSON can represent a wide range of data types, including strings, numbers, arrays, and booleans, which is useful for capturing the diverse information presented in the problems.
- **Machine Readability**: JSON is easily parsed by most programming languages, which allows for automated processing and analysis of the problem data, potentially leading to more accurate and efficient problem-solving.
Here is an example of how one of the problems might be represented in JSON:

```json
{
  "problem": {
    "title": "Mystery of the Hidden Brother",
    "characters": [
      {"name": "Andy", "age": 5.5, "attributes": ["expressive", "knows Matthew's location"]},
      {"name": "Matthew", "age": 3, "attributes": ["hiding", "younger brother"]},
      ......
    ],
    "locations": [
      {"name": "Matthew's room", "attributes": ["no door on closet", "crib with books underneath"]},
      {"name": "Andy's room", "attributes": ["water spill at entrance"]},
      ......
    ],
    "events": [
      {"description": "Andy tells Dad that Matthew is hiding", "location": "home", "time": "evening"},
      {"description": "Andy is excited for Dad to find Matthew", "location": "den"},
      ......
    ],
    "hints": [
      {"description": "Grass stains on Andy's feet", "implication": "recently outside"},
      {"description": "Chocolate remnants around Andy's lips", "implication": "ate chocolate"},
      ......
    ],
    "question": "Where is Matthew?"
  }
}
```

This JSON representation captures the essential details of the problem, organizing them into categories such as characters, locations, events, and hints. Each category is an array of objects, where each object represents a specific element with its own set of attributes. This structured format allows for a clear and detailed representation of the problem, which can be easily translated into a solution algorithm.

Second Stage:
1. Initial State Representation:

```json
{  ......  }
```

2. Step-by-Step Solution Process:
- The grandfather clock stopped at 10:10 p.m., which is shortly after Ed left the shop. It is unlikely that the thief broke in immediately after Ed left, as it would be too risky.
- .....
Given that the mantle clock was left on the edge of the counter by Ed and was found on the floor, it is likely that the intruder knocked it over, causing it to stop at 10:59 p.m. This suggests that the break-in occurred around this time.
3. Concluding with the Answer:
Answer: 1. The mantle clock (stopped at 10:59 p.m.)

Table 48: Examples of Minute Mysteries QA (GPT-4 Task-Based)

第一阶段：
统一表示形式：JSON（JavaScript 对象表示法）说明：

JSON 是一种轻量级的数据交换格式，易于人类阅读和编写，也易于机器解析和生成。它基于 JavaScript 编程语言的一个子集，是一种完全语言无关的文本格式。JSON 非常适合表示上述问题中涉及的状态或条件，因为它可以以层级方式组织数据，这对于表示包含多个实体和属性的复杂情景尤为有用。
使用 JSON 表示问题可从以下方面提高准确率：
- **标准化**：JSON 为表示每个问题的各个要素（如角色、地点、物品和事件）提供了一致的格式。这种标准化减少了歧义，并确保了问题的每个方面都定义明确。
- **层级与嵌套**：JSON 允许在一个对象中嵌套其他对象，从而可以清晰地表示问题不同要素之间的关系，例如地点层级（房子 > 房间 > 壁橱）或事件序列。
- **键值对**：JSON 的键值对能够对每条信息进行精确标注，从而有助于准确地识别和检索数据以进行分析。
- **灵活性**：JSON 可以表示多种数据类型，包括字符串、数字、数组和布尔值，这对于捕获问题中呈现的多样化信息非常有用。
- **Machine Readability**：JSON 易于被大多数编程语言解析，从而实现自动化处理和对问题数据的分析，可能带来更准确、更高效的问题解决。
以下是其中一个问题如何以 JSON 表示的示例：

```json { "problem": { "title": "隐藏兄弟之谜", "characters": [{"name": "安迪", "age": 5.5, "attributes": ["富有表现力", "知道马修的地点"]}, {"name": "马修", "age": 3, "attributes": ["躲藏", "弟弟"]}, .....],"locations": [{"name": "马修的房间", "attributes": ["壁橱没有门", "书在摇篮下面"]}, {"name": "安迪的房间", "attributes": ["入口处有溅水"]}, .....],"events": [{"description":"安迪告诉爸爸马修在躲藏", "location": "家", "time": "傍晚"}, {"description": "安迪对爸爸找到马修感到兴奋", "location": "阁楼"}, .....],"hints": [{"description": "安迪脚上的草渍", "implication": "最近在外面"}, {"description": "安迪嘴唇周围的巧克力残留", "implication": "吃了巧克力"}, .....],"question": "马修在哪里？" } }、、、
```

此JSON表示捕捉了问题的关键细节，并将其组织成诸如角色、地点、事件和线索等类别。每个类别都是一个对象数组，其中每个对象表示一个特定的元素，并拥有一组自己的属性。这种结构化格式允许对问题进行清晰而详细的表示，并且可以轻松地转换为求解算法。

第二阶段：。
1 初始状态表示：

```json { ..... }
```

2. 逐步求解过程：
- 落地钟在晚上10:10停止，这正是在埃德离开商店后不久。小偷不太可能在埃德离开后立即闯入，因为那样风险太大。
- .....
已知条件：壁炉架钟被埃德留在了台面的边缘，之后在地板上被发现，因此入室者很可能将其碰倒，导致它停在晚上10:59。这表明入室盗窃发生在大约这个时间。
3. 以答案结尾：
答案：1. 壁炉架钟（停在晚上10:59）

表48：Minute Mysteries 问答示例（GPT-4任务型）

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

First Stage:
To enhance the clarity and accuracy of the analysis, a structured and concise form of communication can be employed. One suitable format is a markdown table, which can be used to organize and present the relevant information in a clear and concise manner. The table can include columns for each of the five friends, their departure times, and any other relevant information. This format would allow for easy comparison and identification of the missing person.
For example, the table could be structured as follows:

```
| Friend | Departure Time | Additional Information |
|---|---|---|
| Billy Friend | 9:50 PM | Drove Susan Allen home |
| Diana Scott | 11:00 PM | Drove home in her father's Cadillac |
| Harrell Garner | 9:30 PM | Called parents to inform them he was on his way home |
| Susan Allen | 9:50 PM | Rode with Billy Friend to her dorm |
| Steve Collier | N/A | Host of the study session |
```

By organizing the information in this way, it becomes easier to identify the missing person. Since all of the friends are accounted for except for one, it is clear that the missing person is the one who left last, which is Diana Scott.
Using a structured and concise form of communication, such as a markdown table, can help to improve the accuracy of the analysis by providing a clear and organized way to present the relevant information. This can make it easier to identify patterns and inconsistencies, and to draw accurate conclusions.

Second Stage:
Initial State Representation:
* Five friends (Billy, Diana, Harrell, Susan, and Steve) gathered at Steve's apartment on Thursday night, November 21, 1963, to study for upcoming exams.
* Harrell left first at 9:30 PM and drove home, which usually took about 20 minutes.
* Billy and Susan left next at 9:50 PM. Billy drove Susan to her dorm and then headed to his own home.
* Diana left last at 11:00 PM and drove home in her father's Cadillac.
* Steve stayed at his apartment and went to bed.
Step-by-Step Solution Process:
1. We know that one of the friends is missing, but we don't know who it is.
2. We can eliminate Harrell as a possibility because he arrived home safely.
3. We can also eliminate Billy and Susan because they were together and there is no indication that anything happened to them.
4. This leaves Diana as the only remaining friend who has not been accounted for.
Concluding with the Answer:
Answer: 1. Diana Scott

Table 49: Examples of Minute Mysteries QA (Gemini Pro Instance-Based)

第一阶段：T
o enh为了增强分析的清晰度和准确率，可以采用结构化且简洁的通信形式。一种合适的格式是 Markdown 表格，可用于以清晰简洁的方式组织和呈现相关信息。表格可以包含各列，分别对应五位朋友、他们的出发时间以及其他相关信息。此格式便于进行比较并识别缺失的人员。例如，表格可以结构化为以下形式：

好友	离开时间	附加信息
比利·弗兰德	9:50 PM	开车送苏珊·艾伦回家
戴安娜·斯科特	11:00 PM	乘坐她父亲的凯迪拉克回家
哈雷尔·加纳	9:30 PM	打电话给父母告知他们他正在回家的路上
苏珊·艾伦	9:50 PM	与比利·弗兰德同车前往她的宿舍
史蒂夫·科利尔	N/A	学习会的主办人

通过以这种方式组织信息，更容易找到失踪的人。既然除了一个人之外所有好友都已确认，很明显失踪者就是最后离开的那个人，即戴安娜·斯科特。
使用结构化且简洁的通信形式，例如 Markdown 表格，可以通过提供一种清晰且有条理的方式来呈现相关信息，从而提高分析的准确率。这使得更容易发现模式和不一致之处，并得出准确的结论。

第二阶段：
初始状态表示：
* 五位好友（比利、戴安娜、哈雷尔、苏珊和史蒂夫）于1963年11月21日星期四晚上聚集在史蒂夫的公寓里，为即将进行的考试复习。 * 哈雷尔最先于晚上9:30离开并开车回家，通常需要大约20分钟。 * 比利和苏珊接下来于晚上9:50离开。比利开车送苏珊到她的宿舍，然后返回自己的家。 * 戴安娜最后于晚上11:00离开，开着她父亲的凯迪拉克回家。 * 史蒂夫留在他的公寓里并上床睡觉。 **逐步求解过程：** 1. 我们知道其中一位朋友失踪了，但我们不知道是谁。 2. 我们可以排除哈雷尔，因为他已安全到家。 3. 我们同样可以排除比利和苏珊，因为他们在一起，且没有迹象表明他们发生了什么。 4. 因此，戴安娜是唯一剩下且下落不明的朋友。 **以答案结尾：** 答案：1. 戴安娜·斯科特

表49：Minute Mysteries 问答示例（Gemini Pro 基于实例）

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

First Stage:

```
[room_of_murder, murder_weapon, suspect, motive]
```

This representation form can enhance the accuracy of analysis by providing a structured and organized way to capture the key elements of each similar problem. By breaking down the problems into these four categories, it becomes easier to identify patterns and relationships between the different elements. This can help to identify potential suspects, motives, and murder weapons, as well as to eliminate unlikely possibilities. Additionally, this representation form can help to ensure that all of the relevant information is considered when analyzing the problems, which can lead to more accurate and reliable conclusions.
For example, consider the following two problems:
Problem 1:
* A man is found dead in his study, stabbed in the back.
* The study is locked from the inside.
* The only other person in the house is the man's wife, who claims she was asleep at the time of the murder.
Problem 2:
* A woman is found dead in her bedroom, strangled.
* The bedroom window is open, and there are footprints in the flower bed below.
* The only other person in the house is the woman's husband, who claims he was out of town at the time of the murder.
Using the representation form above, we can represent these two problems as follows:

```
Problem 1: [study, knife, wife, unknown]
Problem 2: [bedroom, hands, husband, unknown]
```

By comparing these two representations, we can see that there are some similarities between the two problems. For example, in both cases, the murder took place in a private room, and the only other person in the house is the victim's spouse. This suggests that the spouse may be a potential suspect in both cases.
Additionally, by considering the different elements of each problem, we can start to generate hypotheses about what may have happened. For example, in Problem 1, the fact that the study is locked from the inside suggests that the wife may have been involved in the murder. In Problem 2, the fact that the bedroom window is open and there are footprints in the flower bed below suggests that the husband may have entered the house through the window and killed the wife.
Of course, these are just hypotheses at this point, and more investigation would be needed to determine what actually happened in each case. However, the representation form above provides a useful starting point for analysis, and it can help to ensure that all of the relevant information is considered.

Second Stage:
Initial State Representation:

```
[room_of_incident, object_involved, people_present, time_of_incident]
```

Step 1:
* **room_of_incident:** Music room
* **object_involved:** Canary cage, canary named Rufus
* **people_present:** Mr. Carstairs, Mrs. Carstairs (upstairs), maid Marjorie (an hour ago), children Jimmy and Lydia, neighbor's boy Alex Johnston
* **time_of_incident:** Approximately 20 minutes before Crusher's arrival
Step 2:
* The canary cage was knocked off its perch and fell to the floor, causing the door latch to come undone and the canary to escape.
* The window was open, suggesting that the cage may have been knocked off the perch by someone or something coming in through the window.
Step 3:
* The children were playing frisbee outside the house at the time of the incident.
* Alex admits to running to catch a frisbee and falling, causing the frisbee to go through the window.
* Jimmy also went into the house to retrieve the frisbee, but didn't notice anything about the birdcage.
* Lydia was drinking from her cup near the house and had a clear view of the house, but didn't see the canary escape.
Step 4:
* There are mud stains and grass strands near the French doors, suggesting that someone came in through the window with muddy shoes.
* The cat, Sarabelle, was also in the room at the time of the incident.
Conclusion:
Based on the evidence, it is most likely that Alex is responsible for the disappearance of the canary. He admits to running to catch a frisbee and falling, which could have caused the cage to be knocked off the perch. The mud stains and grass strands near the French doors also suggest that he came in through the window with muddy shoes.
Answer: 0. Alex Johnston

Table 50: Examples of Minute Mysteries QA (Gemini Pro Task-Based)

第一阶段：

```
[房间_的_谋杀，凶器，嫌疑人，动机]。。
```

这种表示形式可以通过提供一种结构化和有序的方式来捕获每个相似问题的关键要素，从而提升分析的准确率。通过将问题分解为这四个类别，更容易识别不同要素之间的模式和关系。这有助于识别潜在的嫌疑人、动机和凶器，以及排除不太可能的可能性。此外，这种表示形式有助于确保在分析问题时考虑所有相关信息，从而得出更准确、更可靠的结论。例如，考虑以下两个问题：

问题1：
* 一名男子被发现在其书房中被从背后刺死。 * 书房从里面锁着。 * 房子里唯一的其他人员是男子的妻子，她声称在谋杀发生时正在熟睡。 **问题2：** * 一名女子被发现在其卧室中被勒死。 * 卧室的窗户开着，下方的花坛里有脚印。 * 房子里唯一的其他人员是女子的丈夫，他声称在谋杀发生时不在城里。 使用上述表示形式，我们可以将这两个问题表示如下：

```
问题1： [书房, 刀,妻子, 未知]问题2： [卧室, 手,
丈夫, 未知]。。
```

通过比较这两种表示形式，我们可以看出这两个问题之间有一些相似之处。例如，在这两种情况下，谋杀都发生在一个私人房间，而房子里唯一的其他人员是受害者的配偶。这表明配偶在这两种情况下都可能是嫌疑人。

此外，通过考虑每个问题的不同要素，我们可以开始生成关于可能发生了什么事情的假设。例如，在问题1中，书房从里面被锁上的事实表明妻子可能参与了谋杀。在问题2中，卧室窗户敞开着且下方花坛里有脚印的事实表明丈夫可能通过窗户进入房间并杀死了妻子。
当然，这些目前只是假设，还需要更多的调查来确定每个案件中实际发生了什么。然而，上述表示形式为分析提供了有用的起点，并且可以帮助确保所有相关信息都被考虑在内。

第二阶段：
初始状态表示：

```
[房间_事件_的，物品_涉及的，人_在场的，事件_发生_的]。。时间
```

第一步：
* **事发房间：** 音乐室_ _* **涉事物品_：** 金丝雀笼，名为鲁弗斯的金丝雀 * **在场人员_：** 卡斯特莱恩斯先生、卡斯特莱恩斯夫人（在楼上）、女仆玛乔丽（一小时前）、孩子吉米和莉迪亚、邻居家的男孩亚历克斯·约翰斯顿 * **事发时间__：** 大约在克拉谢尔到达前20分钟 **步骤 2：** * 金丝雀笼从栖架上掉落，摔到地板上，导致门闩脱落，金丝雀趁机逃脱。 * 窗户是开着的，这表明笼子可能是被从窗户进来的某人或某物碰掉的。 **步骤 3：** * 事发时，孩子们正在房子外面玩飞盘。 * 亚历克斯承认自己跑去接飞盘时摔倒，导致飞盘穿过窗户飞进屋内。 * 吉米也进屋去捡飞盘，但他没有注意到关于鸟笼的任何异常。 * 莉迪亚当时在房子附近端着杯子喝饮料，视野中可以清楚地看到房子，但她没有看到金丝雀逃走。 **步骤 4：** * 法式门附近有泥点和草丝，表明有人穿着沾满泥的鞋子从窗户进了屋。 * 猫萨贝勒在事发时也在房间里。 **结论：** 根据证据，亚历克斯最有可能对金丝雀的失踪负有责任。他承认自己跑去接飞盘时摔倒，这可能导致笼子从栖架上掉落。法式门附近的泥点和草丝也表明他穿着沾满泥的鞋子从窗户进了屋。 **答案：0. 亚历克斯·约翰斯顿**

表50：Minute Mysteries 问答示例（Gemini Pro 基于任务）

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

First Stage (generated by GPT-3.5):
Problem Representation:
Let's represent the problem in the form of a table:

```
| Problem | Given | Unknown | Formula |
|---------|-------|---------|---------|
| 1       | Cost, Profit, Percentage | Selling Price | Selling Price = Cost + Profit |
| 2       | Number of steps | Number of ways | Fibonacci sequence |
| 3       | Vitamin E, Vitamin A | Grams of health food | (0.09 * x) + (0.08 * x) = 1 |
| 4      | Average microwave ovens sold | Average for 22 months | (Total ovens sold in 22 months) / 22 |
| 5       | Size of the cube, Thickness | Number of unpainted faces | (Size - 2)^3 |
```

Explanation:
By representing the problems in a unified table format, it becomes easier to identify the given information, the unknowns, and the relevant formulas or concepts needed to solve each problem. This allows for a more systematic and organized approach to problem-solving, enhancing the accuracy of the analysis. It also helps in identifying patterns and similarities between different problems, making it easier to apply the appropriate solution methods.

Second Stage (generated by GPT-4):
Initial State Representation:

```
| Problem | Given | Unknown | Formula |
|---------|-------|---------|---------|
| Choosing Oranges | 10 oranges in total | Number of ways to choose 2 oranges | Combination formula: C(n, k) = n! / (k!(n - k)!) |
```

Step-by-Step Solution Process:
1. We are given 10 oranges and we need to find the number of ways to choose 2 oranges from these 10.
2. The order in which we choose the oranges does not matter, so we use the combination formula.
3. Using the combination formula C(n, k) = n! / (k!(n - k)!), where n is the total number of items, and k is the number of items to choose, we get:
   `C(10, 2) = 10! / (2!(10 - 2)!)`
4. Simplifying the factorial expressions, we get:
   `C(10, 2) = (10 × 9) / (2 × 1)`
5. Further simplifying, we get:
   `C(10, 2) = 90 / 2`
6. Finally, we get:
   `C(10, 2) = 45`

Concluding with the Answer:
Answer: A

Table 51: Examples of AQuA (GPT-3.5 ⇒ GPT-4 Task-Based)

第一阶段（由 GPT-3.5 生成）：
问题表示：
让我们以表格形式表示该问题：
| 问题 | 已知条件 |未知 | 公式 |---------|------|--------|--------| 1 | 成本、利润、百分比 | 售价 | 售价 = 成本 + 利润 | 2 | 步数 | 方式数量 | 斐波那契数列 | 3 | 维生素 E、维生素 A | 健康食品克数 | (0.09 * x) + (0.08 * x) = 1 | 4 | 微波炉销售平均数 | 22个月的平均值 | (22个月内出售的微波炉总数) / 22 | 5 | 立方体尺寸、厚度 | 未涂漆面数 | (尺寸 - 2)^3 |

解释：
通过将问题表示为统一的表格格式，更容易识别已知条件、未知量以及解决每个问题所需的相关公式或概念。这使得问题解决的方法更加系统和有组织，从而提高了分析的准确率。它还有助于识别不同问题之间的模式和相似之处，从而更容易应用合适的解决方案方法。

第二阶段（由 GPT-4 生成）：我
nitial State R代表ion:
| 问题 | 已知条件 |未知 | 公式 |---------|------|--------|--------| 选择橙子 | 10 橙子总数 | 从 | 中选择2个橙子的方式数量 | 组合公式：C(n, k) = n! / (k!(n - k)!) |逐步求解过程： 1. 已知有10个橙子，需要找出从这10个橙子中选择2个橙子的方式数量。 2. 选择橙子的顺序不重要，因此我们使用组合公式。 3. 使用组合公式C(n, k) = n! / (k!(n- k)!)，其中n是物品的总数，k是要选择的物品数量，我们得到： C(10, 2) = 10! / (2!(10 - 2)!) 4. 简化阶乘表达式，我们得到： C(10, 2) = (10 × 9) / (2 × 1) 5. 进一步简化，我们得到： C(10, 2 = 90 2) / . 最后，我们得到： C(10, 2) = 45以答案结尾： 答案：A

6

表 51：AQuA示例（GPT-3.5 ⇒ GPT-4任务型）

First Stage:
Initial State Representation:
Let's represent the vertices of the triangle as follows:
Vertex 1: (a, 0)
Vertex 2: (b, 0)
Vertex 3: (x, y)
We are looking for the area of the triangle formed by these vertices.

Second Stage:
Step-by-Step Solution Process:
The area of a triangle with vertices (x1, y1), (x2, y2), and (x3, y3) can be calculated using the formula:
Area = |(x1(y2 - y3) + x2(y3 - y1) + x3(y1 - y2))/2|
Given the vertices of the triangle:
Vertex 1: (a, 0)
Vertex 2: (b, 0)
Vertex 3: (x, y)
Using the formula, the area of the triangle is:
Area = |(a(0 - y) + b(y - 0) + x(0 - 0))/2|
Area = |(by - ay)/2|
Therefore, the area of the triangle is represented by:
C) (by - ay)/2
Answer: C

Table 52: Examples of AQuA (GPT-3.5 Instance-Based)

第一阶段：初始状态表示：让我们将三角形的顶点表示如下：
顶点 1：(a, 0) 顶点 2：(b, 0) 顶点 3：(x, y) 我们要求由这些
顶点构成的三角形的面积。

第二阶段：逐步求解过程：已知条件为顶点 (x1, y1)、(x2, y2) 和 (x3, y3) 的三角形的面积，
可使用以下公式计算：面积 = |(x1(y2 - y3) + x2(y3 - y1) + x3(y1 - y2))/2|已知三角形的顶点为：
顶点 1: (a, 0) 顶点 2: (b, 0) 顶点 3: (x, y) 使用公式，该三角形的面积为：面积 = |(a(0 - y) +
b(y - 0) + x(0 - 0))/2|面积 = |(by - ay)/2|因此，该三角形的面积表示为：C) (by - ay)/2 答案：
C

表 52：AQuA示例（GPT-3.5 基于实例）

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

First Stage:
1. Let’s represent the similar problems using a table:

Problem	Variables/Conditions
1	Size of cube = 5cm, thickness = 1cm, size of smaller cubes = 1cm, outer surface painted
2	Q is younger than R, R+T=50
3	Coin size = 2x15mm, weight = 30g, volume of aluminum = volume of silver, silver is twice as heavy as aluminum
4	5 vowels, 20 consonants, middle letter must be vowel, first and third letter must be different and consonants
5	Discount of 5% and 7%, difference in profit = Rs. 15

This representation form can be used to solve each problem by plugging in the specific values and conditions.

Second Stage:
Initial State Representation:
The given equation is (5 + ? × 19 - 15 - 7) / (13 × 13 - 156) = 6
Step-by-Step Solution Process:
1. First, simplify the expression inside the brackets:
   (5 + ? × 19 - 15 - 7) = (5 + 19? - 22)
2. Next, simplify the denominator:
   (13 × 13 - 156) = (169 - 156) = 13
3. Substitute the simplified expressions back into the original equation:
   (5 + 19? - 22) / 13 = 6
4. Multiply both sides by 13 to clear the fraction:
   5 + 19? - 22 = 78
5. Combine like terms:
   19? - 17 = 78
6. Add 17 to both sides:
   19? = 95
7. Divide both sides by 19:
   ? = 5
Concluding with the Answer:
Answer: C) 5

Table 53: Examples of AQuA (GPT-3.5 Task-Based)

第一阶段：1. L

et’我们使用表格来表示类似问题：|问题 | 变量/条件 |---|---| 1 | 立方体尺寸 = 5厘米，厚度 = 1厘米，小立方体尺寸 = 1厘米，外表面涂漆 | 2 | Q比R年轻，R+T=50 | 3 | 硬币尺寸 = 2 x15毫米，重量 = 30克，铝的 = 体积的银，银的重量是铝的两倍 | 4 | 5 元音，20个辅音，中间字母必须是元音，第一个和第三个字母必须不同且为辅音 | 5 | 5%和7%的折扣，利润差异 = 卢比 15 |这种表示形式可用于通过代入具体数值和条件来求解每个问题。第二阶段：初始状态表示：给定的方程是(5 + ? × 19 - 15 - 7) / (13 × 13 - 156) = 6逐步求解过程：1. 首先，简化括号内的表达式：(5 + ? × 19 - 15 - 7) = (5 + 19? - 22) 2. 接下来，简化分母：(13 × 13 156 = 169 156 = 13 -) (-) . 将简化后的表达式代回原方程：(5 + 19 3 22 13 = 6 ? -) / . 两边同乘以清除分数：5 + 19? - 22 = 78 5. 合并同类项：19? - 17 = 78 6. 两边加17：19? = 95 7. 两边除以19：? = 5以答案结尾：答案：C) 5
4 13

表53：AQuA示例（GPT-3.5基于任务）

49

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

First Stage (generated by GPT-4):
JSON (JavaScript Object Notation) is a lightweight data-interchange format that is easy for humans to read and write, and easy for machines to parse and generate. It can be used to represent the state or conditions of the problems provided. Below is a JSON structure that could represent each of the problems:

```json
{
  "problem": {
    "type": "",
    "data": {
      "values": [],
      "units": "",
      "additional_info": {}
    },
    "choices": {
      "A": "",
      "B": "",
      "C": "",
      "D": "",
      "E": ""
    }
  }
}
```

This JSON structure includes a ‘problem‘ object that contains:
- ‘type‘: A string describing the type of problem (e.g., "speed_time_distance", "probability", "geometry", "relative_motion", "number_theory").
- ‘data‘: An object that includes:
  - ‘values‘: An array of numerical values relevant to the problem.
  - ‘units‘: A string indicating the units of measurement used in the problem.
  - ‘additional_info‘: An object that can hold any additional information specific to the problem type (e.g., speed of the ship, width of pickets, etc.).
- ‘choices‘: An object that lists the possible answer choices, with each choice represented by a key-value pair where the key is the choice label (A, B, C, D, E) and the value is the corresponding answer in string format.

This JSON structure can be adapted to represent each of the problems provided by simply filling in the appropriate values for each key.

Second Stage (generated by GPT-3.5):

```json
{
  "problem": {
    "type": "quadratic_equation",
    "data": {
      "values": [-5, "k"],
      "units": "",
      "additional_info": {}
    },
    "choices": {
      "A": "9",
      "B": "7",
      "C": "3",
      "D": "8",
      "E": "12"
    }
  }
}
```

To find the largest integral value of ’k’ for which the quadratic equation x^2 - 5x + k = 0 will have two real and distinct roots, we can use the discriminant formula. The discriminant (D) of the quadratic equation ax^2 + bx + c = 0 is given by D = b^2 - 4ac. For the equation x^2 - 5x + k = 0, the discriminant must be greater than 0 for the equation to have two real and distinct roots.
The discriminant in this case is $(-5)^2$ - 4*1*k = 25 - 4k. For the equation to have two real and distinct roots, 25 - 4k must be greater than 0. Solving for k, we get k < 25/4.
The largest integral value of k that satisfies this condition is 6. Therefore, the largest integral value of k for which the quadratic equation x^2 - 5x + k = 0 will have two real and distinct roots is 6.
Answer: D

Table 54: Examples of AQuA (GPT-4 ⇒ GPT-3.5 Task-Based)

第一阶段（由 GPT-4 生成）：
JSON 这是一种轻量级的数据交换格式，易于人类阅读和书写，也易于机器解析和生成。它可以用来表示所提供问题的状态或条件。下面是一种可用于表示各个问题的 JSON 结构：

```json { "问题": { "类型": "", "
数据": { "数值": []，"单位": "",
" 额外 _ 信息 ": {} }, " 选项 ":
{ "A": "", "B": "", "C": "", "D": "
", "E": "" } } }。。。

此 JSON 结构包含一个“problem”对象，其内容如下：- “type”：描述问题类型的字符串（例如，“速度_时间_距离”、“概率”、“几何”、“相对_运动”、“数字_理论”）。- “data”：包含以下内容的对象：- “values”：与问题相关的数值的数组。- “units”：指示问题中所用测量单位的字符串。- “additional_info”：可保存与问题类型相关的任何其他信息的对象（例如，船速、栅栏宽度等）。- “choices”：列出可能答案选项的对象，每个选项由一个键值对表示，其中键是选项标签（A、B、C、D、E），值是对应的字符串格式答案。只需为每个键填入相应的数值，即可将此 JSON 结构适配于所提供的每个问题。

**第二阶段（由 GPT-3.5 生成）：**

```json { "问题": { "类型": "二次_方
程", "数据": { "数值": [-5, "k"]，"单位
": "", "补充_信息": {} }, "选项": { "A":
"9", "B": "7", "C": "3", "D": "8", "E": "
12" } } }。。。

为了求出使二次方程 $x^2$ - 5x + k = 0 具有两个不相等的实数根的 k 的最大整数值，我们可以使用判别式公式。二次方程 $ax^2$ + bx +c = 0 的判别式 (D) 由 D = $b^2$ - 4ac 给出。对于方程 $x^2$ - 5x + k = 0，要使方程具有两个不相等的实数根，判别式必须大于 0。在这种情况下，判别式为 $(-5)^2$ - 4*1*k = 25 - 4k。要使方程具有两个不相等的实数根，25 - 4k 必须大于 0。解出 k，我们得到 k < 25/4。满足此条件的 k 的最大整数值是 6。因此，使二次方程 $x^2$ - 5x + k = 0 具有两个不相等的实数根的 k 的最大整数值是 6。答案：D

表格 54：AQuA示例（GPT-4 ⇒ GPT-3.5基于任务）


50
翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

**First Stage (generated by GPT-4):**
Unified Representation Form: Algebraic Equations
Explanation:
Algebraic equations are a versatile and powerful tool for representing a wide variety of problems, including the ones listed above. By using variables to represent unknown quantities and forming equations based on the given conditions, we can create a system that captures the essence of each problem. This approach allows us to translate word problems into mathematical language, which can then be solved using algebraic methods.
The algebraic representation enhances the accuracy of analysis by providing a clear, unambiguous framework for the problem. It eliminates the potential for misinterpretation of the problem's conditions and ensures that all relationships between quantities are explicitly stated. Additionally, algebra allows for the systematic application of mathematical rules and operations to find the solution, reducing the likelihood of errors that might arise from ad-hoc or intuitive reasoning.
For example, the first problem can be represented by the following system of equations, where A represents Arjun's money and S represents Sajal's money:

```
1. S + 30 = 2(A - 30)
2. A + 10 = 3(S - 10)
```

Each of the other problems can similarly be translated into algebraic equations that reflect the given conditions. This unified representation form is capable of expressing the relationships and constraints of each problem, making it a suitable choice for a wide range of similar problems.

Second Stage (generated by Gemini Pro):
Initial State Representation:
Let x be the total number of people who took the survey.
Step 1:
60% of the people who took the survey were women. So, the number of women is 0.6x.
Step 2:
Of the men who were surveyed, 75% preferred red cars over green cars. So, the number of men who preferred red cars is 0.75 * (x - 0.6x) = 0.25x.
Step 3:
10 men liked green cars more than red. So, the number of men who preferred green cars is 0.25x - 10.
Step 4:
The total number of people who took the survey is the sum of the number of women and the number of men. So, x = 0.6x + (0.25x - 10).
Step 5:
Solving for x, we get x = 80.
Answer: E

Table 55: Examples of AQuA (GPT-4 ⇒ Gemini Pro Task-Based)

First Stage:
Initial State Representation:
Let's use a math equation to represent the initial state of the problem.
Let x be the length of the longer piece of rope in meters. According to the problem, the shorter piece is 3 meters shorter than the longer piece, so its length would be x - 3 meters.
The total length of the rope is 20 meters. Therefore, the sum of the lengths of the two pieces must equal 20 meters.
The equation representing the relationship between the lengths of the two pieces of rope is:
x + (x - 3) = 20
This equation can be used to solve for the value of x, which represents the length of the longer piece of rope.

Second Stage:
Initial State Representation:
We have the equation x + (x - 3) = 20, where x is the length of the longer piece of rope.
Step-by-Step Solution Process:
1. Combine like terms on the left side of the equation:
```
2x - 3 = 20
```
2. Add 3 to both sides to isolate the term with the variable x:
```
2x - 3 + 3 = 20 + 3
2x = 23
```
3. Divide both sides by 2 to solve for x:
```
2x / 2 = 23 / 2
x = 11.5
```
Concluding with the Answer:
The length of the longer piece of rope is 11.5 meters.
Answer: D

Table 56: Examples of AQuA (GPT-4 Instance-Based)

第一阶段（由GPT-4生成）：
统一表示形式：代数方程 解释：

代数方程是一种功能强大、用途广泛表示问题的工具，其中包括上述所列的问题。通过利用变量表示未知量，并根据已知条件建立方程，我们可以构建一个能够捕捉每个问题精髓的系统。这种方法使我们能够将文字题转化为数学语言，随后利用代数方法求解。
代数表示通过为问题提供清晰、无歧义的框架来提升分析的准确率。它消除了对问题条件误解的可能性，并确保量与量之间的所有关系都被明确陈述。此外，代数允许系统性地应用数学规则和操作来寻找解决方案，从而降低因临时性或非直觉性推理而可能产生的错误概率。
例如，第一个问题可以用以下方程组来表示，其中A表示阿尔君的钱，S表示萨贾尔的钱：

1. S + 30 = 2(A - 30) 2.
A + 10 = 3(S - 10)
其他每个问题都可以类似地转化为反映已知条件的代数方程。这种统一表示形式能够表达每个问题的关系和约束，因此是解决各种类似问题的合适选择。

第二阶段（由 Gemini Pro 生成）：
初始状态表示： 设 x 为参与调查的人数总数。**第一步：** 参与调查的人中有 60% 是女性。因此，女性人数为 0.6x。**步骤2：** 在接受调查的男性中，75% 偏好红色汽车而不是绿色汽车。因此，偏好红色汽车的男性人数为 0.75* (x - 0.6x) = 0.25x。**步骤 3：** 有 10 名男性更喜欢绿色汽车而不是红色汽车。因此，偏好绿色汽车的男性人数为 0.25x - 10。**步骤 4：** 参与调查的人数总数等于女性人数和男性人数之和。因此，x = 0.6x + (0.25x - 10)。**步骤 5：** 求解 x，我们得到 x = 80。**答案：E**

表格 55：AQuA示例（GPT-4 ⇒Gemini Pro 基于任务）

第一阶段：
初始状态表示：让我们使用一个数学公式来表示问题的初始状态。设 x 为较长的那段绳子的长度（以米为单位）。根据问题，较短的那段绳子比较长的那段短 3 米，因此其长度为 x - 3 米。绳子的总长度为 20 米。因此，两段绳子的长度之和必须等于 20 米。表示这两段绳子长度之间关系的方程为：x + (x - 3) = 20 此方程可用于求解 x 的值，该值代表较长的那段绳子的长度。

第二阶段：
初始状态表示：我们有方程 x +(x - 3) = 20，其中 x 是较长的那段绳子的长度。
逐步求解过程：1. 合并方程左侧的同类项：2x - 3 = 202. 两边同时加 3 以分离含变量 x 的项：2x - 3 + 3 = 20 + 32x = 233. 两边同时除以 2 求解 x：2x / 2 = 23/ 2 x = 11.5 以答案结尾：较长的那段绳子的长度是 11.5 米。答案：
D

表格 56：AQuA示例（GPT-4 基于实例）

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

First Stage:
Representation Form: Pseudocode
Explanation:
Pseudocode is a high-level description of an algorithm that uses the structural conventions of programming languages but is intended for human reading rather than machine reading. It allows us to represent the logic of a problem without getting bogged down in the syntax of a particular programming language. By using pseudocode, we can create a generic template that can be adapted to solve a variety of similar problems. This enhances accuracy by ensuring that the logical steps are clearly outlined and can be followed systematically, reducing the risk of errors that might arise from misinterpreting the problem or from syntax issues in actual code.
Pseudocode Template:

```
INITIALIZE variables and ratios as needed

READ input values

CALCULATE based on given ratios or probabilities

IF there are conditions or multiple scenarios THEN
    HANDLE each scenario separately
    CALCULATE probabilities or outcomes for each scenario
    COMBINE results as per the problem's requirements
END IF

CONVERT units if necessary

OUTPUT the final result
```

This pseudocode template can be adapted to represent the logic of each problem presented. It starts with initializing variables and ratios, reads input values, performs calculations based on the given information, handles conditional logic if present, converts units if required, and finally outputs the result. By following this structured approach, we can ensure that each step of the problem-solving process is accounted for, which helps in maintaining accuracy across different problems.

Second Stage:
Initial State Representation:

```
tank_length = 8 // in meters
tank_width = 6 // in meters
tank_height = 2.5 // in meters
capacity_in_liters = 0 // to be calculated
```

Step-by-Step Solution Process:

```
// Step 1: Calculate the volume of the tank in cubic meters
volume_in_cubic_meters = tank_length * tank_width * tank_height

// Step 2: Convert the volume from cubic meters to liters
// 1 cubic meter = 1000 liters
capacity_in_liters = volume_in_cubic_meters * 1000

// Step 3: Substitute the values and calculate the capacity
capacity_in_liters = 8 * 6 * 2.5 * 1000
capacity_in_liters = 120 * 1000
capacity_in_liters = 120000
```

Concluding with the Answer:
Answer: D

Table 57: Examples of AQuA (GPT-4 Task-Based)

第一阶段：
表示形式：伪代码
Explanation:
伪代码是算法的高层描述，它采用编程语言的语法结构约定，但旨在供人类阅读而非机器阅读。它使我们能够表示问题的逻辑，而无需陷入特定编程语言的语法细节。通过使用伪代码，我们可以创建一个通用模板，将其改编以解决一系列类似的问题。这通过确保逻辑步骤被清晰概述并可系统地遵循，从而提高了准确率，降低了因误解问题或实际代码中的语法问题而产生错误的风险。

伪代码模板：

```
按需初始化变量和比率

读取输入值

基于已知条件的比率或概率进行计算

IF 存在条件或多个情景 THEN 分别处理每个情景 计算每个情景
的概率或结果 根据问题的要求合并结果 END IF

如有必要，转换单位

OUTPUT 最终结果、、、
```

此伪代码模板可进行调整，以表示所呈现的每个问题的逻辑。它从初始化变量和比率开始，读取输入值，根据已知信息执行计算，处理条件逻辑（如果存在），在需要时转换单位，并最终输出结果。通过遵循这种结构化方法，我们可以确保问题解决过程中的每个步骤都被考虑在内，这有助于在解决不同问题时保持准确率。

第二阶段：
初始状态表示：

```
水箱_长度 = 8 //以米为单位 水箱_宽度 = 6//
以米为单位 水箱_高度 = 2.5 //以米为单位 容量
_以_升为单位 = 0 //待计算、、、
```

逐步求解过程：

```
// 步骤2：计算水箱的体积，单位为立方米
体积_以_立方_米为单位 =水箱_长度 * 水箱_宽度 * 水箱_高度

// 步骤2：将体积从立方米转换为升 // 1立方米 = 1000 升 容量_
以_升为单位 =体积_以_立方_米为单位 * 1000

// 步骤 3: 代入数值并计算容量 capacity_in_liters = 8 * 6 * 2.5
* 1000 capacity_in_liters = 120* 1000 capacity_in_liters
 = 120000、、、

以答案结尾：
答案: D
```

表57: AQuA示例 (GPT-4任务型)

52
翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

First Stage:
1. Initial State Representation:
Let x be the length of the longer piece of rope.
Length of the shorter piece of rope = x - 3
Total length of the rope = 20 meters
Equation: x + (x - 3) = 20

Second Stage:
1. Initial State Representation:
Let x be the length of the longer piece of wire.
Length of the shorter piece of wire = x - 3
Total length of the wire = 20
Equation: x + (x - 3) = 20
2. Step-by-Step Solution Process:
- Combine like terms: 2x - 3 = 20
- Add 3 to both sides: 2x = 23
- Divide both sides by 2: x = 11.5
3. Concluding with the Answer:
Answer: D) 11.5

Table 58: Examples of AQuA (Gemini Pro Instance-Based)

第一阶段：。
1 初始状态表示：
设x为较长的那段绳子的长度。
较短的那段绳子的长度 = x - 3
绳子的总长度 = 20 米
方程：x +(x - 3) = 20

第二阶段：.
1 初始状态表示：
设 x 为较长金属丝的长度。
较短金属丝的长度 = x - 3
金属丝总长度 = 20
方程：x +(x - 3) = 20
2. 逐步求解过程：
- 合并同类项：2x - 3 = 20
- 两边加3：2x = 23
- 两边除以2：x = 11.5
3. 以答案结尾：
答案：D) 11.5

表格 58：AQuA示例（Gemini Pro 基于实例）

53

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

First Stage:
1. **Representation Form:**

We can represent each problem as a tuple containing the following fields:

- **Problem Type:** A category that classifies the problem, such as "Percentage Gain," "Discount and Price Increase," "Speed and Distance," "Alloy Ratio," or "Purchase Ratio."
- **Given Values:** A dictionary containing the values provided in the problem statement, such as the initial price, selling price, discount percentage, speed, time, alloy ratios, or purchase quantities.
- **Unknown Values:** A list of the values that need to be calculated, such as the percentage gain, new price, length, alloy ratio, or purchase ratio.

For example, the first problem can be represented as:

```
(
    "Problem Type": "Percentage Gain",
    "Given Values": {
        "Initial Price": 450,
        "Selling Price": 7.50,
        "Quantity": 90
    },
    "Unknown Values": ["Percentage Gain"]
)
```

2. **Accuracy Enhancement:**

Using this structured representation form enhances the accuracy of analysis by:

- **Standardization:** It provides a consistent and standardized way to represent problems, making it easier to compare and analyze them.
- **Clarity:** It eliminates ambiguities and ensures that all relevant information is captured, reducing the chances of misinterpretation.
- **Automation:** This representation form can be easily processed by computer programs, enabling automated analysis and solution generation.

3. **Representation Form Generality:**

This representation form can represent every single similar problem because it captures the essential elements common to all problems of a particular type. For example, all percentage gain problems involve calculating the percentage difference between an initial value and a final value. By representing these problems in a structured manner, we can develop generic algorithms or techniques to solve them.

Second Stage:
Initial State Representation:
```
(
    "Problem Type": "Discount and Price Increase",
    "Given Values": {
        "Discount 1": 5,
        "Discount 2": 7,
        "Profit Difference": 15
    },
    "Unknown Values": ["Marked Price"]
)
```

Step-by-Step Solution Process:
1. Let the marked price of the watch be x.
2. Selling price with a 5% discount = x - (5/100) * x = 0.95x
3. Selling price with a 7% discount = x - (7/100) * x = 0.93x
4. According to the problem, the profit difference between the two discounts is Rs. 15. So, we can set up the equation:
   0.95x - 0.93x = 15
5. Solving for x:
   0.02x = 15
   x = 15 / 0.02
   x = 750

Answer: D

Table 59: Examples of AQuA (Gemini Pro Task-Based)

第一阶段：。
1 **表示形式：**
我们可以将每个 问题表示为包含以下字段的元组： 领域：

- **问题类型：** 用于对问题进行分类的类别，例如“百分比收益”、“折扣与涨价”、“速度与距离”、“合金比例”或“购买比例”。

- **已知数值：** 包含问题陈述中提供的数值，例如初始价格、售价、折扣率、速度、时间、合金比例或购买数量。 - **未知数值：** 需要计算的数值列表，例如百分比收益、新价格、长度、合金比例或购买比例。

例如，第一个问题可以表示为：、、、("问题类型": "百分比收益", "已知数值": { "初始价格":450, "售价":7.50, "数量": 90 }, " 未知数值 ": [" 百分比收益 "])、、、

2. **准确性增强：**
使用此结构化 表示形式通过以下方式提高分析的准确率： 分析者：

- **标准化：** 它提供了一种一致且标准化的方式来表示问题，使其更容易进行比较和分析。
- **清晰度：** 它消除了歧义，并确保捕获所有相关信息，从而减少误解的可能性。

- **自动化：** 这种表示形式可以很容易被计算机程序处理，从而实现自动化分析和解决方案生成。

3. **表示形式的通用性：**

这种表示形式可以表示每一个相似的问题，因为它捕捉了所有特定类型问题共有的基本要素。例如，所有百分比增益问题都涉及计算初始值和最终值之间的百分比差异。通过以结构化方式表示这些问题，我们可以开发通用算法或技术来解决它们。

第二阶段：
初始状态表示：
、、、("问题类型": "折扣与涨价", "已知数值": { "折扣 1": 5, "折扣 2": 7, "利润差异": 15 }, "未知数值": ["标价"])
```

**步骤-by-解题过程：**
1. 设手表的标价为 x。 2. 5% 折扣后的售价 =x - (5/100) * x = 0.95x 3. 7% 折扣后的售价 =x - (7/100) * x = 0.93x 4. 根据问题，两种折扣之间的利润差异为 15 卢比。因此，我们可以建立方程： 0.95x - 0.93x = 15 5. 求解 x： 0.02x = 15x = 15 / 0.02 x = 750**答案：D**

表格 59：AQuA示例（Gemini Pro 基于任务）



翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next
```

Elimy:
Context:
- Clive Anderson is a British television and radio presenter, comedy writer, and former barrister. He is best known for being the host of the improvisation show "Whose Line Is It Anyway?".
Foina:
Question:
Where was Clive Anderson born?
Shared Answer:
<A>London</A>

Table 60: Examples of Wiki Hop (GPT-3.5)

艾米丽: 背景: - 克利夫·安德森是英国的电视和广播节目主持人、喜剧编剧，以及前出庭律师。他最为人熟知的身份是即兴表演秀《这是谁的线？》的主持人。 菲娜: **问题**：克利夫·安德森出生在哪里？ **共同答案**：<A>伦敦</A>

表60：Wiki Hop（GPT-3.5）示例

Elimy:
Hey Fiona, I have the question and the options. Do you have any relevant information that can help us answer the question?
Relevant Information:
- Yerim Seowon is a former seowon located in Miryang, Gyeongsangnam-do, South Korea.
- It was built in 1567 to enshrine a native Confucian scholar of Miryang.
- It was designated a tangible cultural treasure of Gyeongsangnam-do in 1974.
Thoughts:
- The information clearly states that Yerim Seowon is in Gyeongsangnam-do, which is a province in South Korea.
- The question asks for the country where Yerim Seowon is located.
Based on the evidence provided, the answer should be "South Korea".
Answer: south korea
Foina:
Great, thanks for the information. Based on the evidence you provided, the answer is indeed "south korea".

Table 61: Examples of Wiki Hop (GPT-3.5, GPT-4)

艾米丽：嗨，菲奥娜，我有问题和问题选项。你有没有任何可以帮助我们回答该问题的相关信息？相关信息：- 汝林书院是韩国庆尚南道密阳的一座前书院。- 它建于1567年，旨在供奉密阳的一位本地儒家学者。- 它于1974年被指定为庆尚南道的有形文化财。思维：- 信息明确指出汝林书院位于韩国庆尚南道，而庆尚南道是韩国的一个省。- 该问题询问汝林书院所在的国家。根据提供的证据，答案应该是“韩国”。答案：southkorea 菲娜：太好了，谢谢你提供的信息。根据你提供的证据，答案确实是“south korea”。

表61：Wiki Hop 示例（GPT-3.5，GPT-4）

Elimy:
Evidence:
- Gymnophiona includes legless caecilians and all amphibians more closely related to them than to frogs or salamanders.
- Caecilians are a group of limbless, serpentine amphibians.
- All modern caecilians and their closest fossil relatives are grouped as a clade, Apoda, within the larger group Gymnophiona.
- Eocaecilia is a caecilian.
My thoughts:
- Since Eocaecilia is a caecilian, and all modern caecilians are grouped within the clade Apoda, which is part of the larger group Gymnophiona, the parent taxon of Eocaecilia is Gymnophiona.
Structured data:

```
{
  "Eocaecilia": {
    "classification": {
      "group": "caecilian",
      "clade": "Apoda",
      "larger_group": "Gymnophiona"
    }
  }
}
```

Conclusion:
Answer: gymnophiona
Fiona:
Great analysis, Emily. I agree with your conclusion. Based on the evidence provided, it’s clear that the parent taxon of Eocaecilia is Gymnophiona. Well done!

Table 62: Examples of Wiki Hop (GPT-4, GPT-3.5)

艾米丽：证据：- 无足目包括无足蚓螈，以及所有与无足蚓螈的亲缘关系比与青蛙或蝾螈更近的两栖动物。- 蚓螈是一类无四肢的蛇形两栖动物。- 所有现生蚓螈及其最近的化石近亲被归类为无足亚纲这一进化支，归属于更大的无足目分类群。- 早蚓螈是一种蚓螈。我的思维：- 由于早蚓螈是一种蚓螈，且所有现生蚓螈都被归类于无足亚纲这一进化支，而无足亚纲属于更大的无足目分类群，因此早蚓螈的亲本分类单元是无足目。结构化数据：{ "Eocaecilia": { "classification": { "group": "蚓螈", "clade": "无足亚纲", "larger_group": "无足目" } } } 结论：答案：无足目 菲奥娜：很好的分析，艾米莉。我同意你的结论。根据所提供的证据，早蚓螈的亲本分类单元无疑是无足目。干得漂亮！

表62：Wiki Hop 示例（GPT-4，GPT-3.5）

55

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

CASE-1:
Elimy:
Evidence:
- Hans Loritz joined the SS in September 1930.
- Hans Loritz joined the NSDAP in August 1930.
- NSDAP is the Nazi Party.
My thoughts:
Based on the evidence, Hans Loritz was a member of the Nazi Party (NSDAP).
Answer: {nazi party}
Foina:
Evidence:
- The context provided mentions the "Schutzstaffel" (SS), which was a major paramilitary organization under Adolf Hitler and the Nazi Party (NSDAP).
- Heinrich Himmler was a leading member of the Nazi Party and head of the SS.
- The context also references the Nazi Party (NSDAP) as being active between 1920 and 1945 and practicing Nazism.
My thoughts:
Given the information about the SS and its connection to the Nazi Party, as well as the direct mention of the Nazi Party (NSDAP), it is logical to conclude that Hans Loritz, associated with the SS, would be a member of the Nazi Party.
Answer: nazi party

CASE-2:
Elimy:
@Fiona: Requesting information on the film "The Eighth Day" to determine its genre. Please provide relevant data.
Foina:
@Emily: The film "The Eighth Day" is described as a Belgian comedy-drama. It tells the story of a friendship between a businessman and a man with Down syndrome. The film includes dream sequences with magical realism elements. It is not specified as a tragicomedy, but the blend of comedy and drama suggests it could fit that category as well.
Evidence:
- Described as a comedy-drama film.
- Contains elements of magical realism.
- Not explicitly labeled as tragicomedy, but the mix of genres implies a potential fit.
My thoughts: Based on the evidence, the genre of "The Eighth Day" aligns with both "comedy" and "drama," and potentially "tragicomedy" due to the blend of these elements.
Recommendation: We should consider "comedy," "drama," and "tragicomedy" as possible answers. Please cross-verify and confirm.
Elimy:
@Fiona: Confirming that "The Eighth Day" is a comedy-drama, which aligns with the genre "tragicomedy" due to its blend of comedy and drama elements.
Fiona: @Emily:Evidence:
- Comedy-drama is a subgenre of tragicomedy.
- "The Eighth Day" is described as a comedy-drama.
My thoughts: Given the evidence, "tragicomedy" is the most accurate genre that encompasses the elements of "The Eighth Day."
Conclusion: Answer: tragicomedy

Table 63: Examples of Wiki Hop (GPT-4)

案例1：
艾米丽：
证据：
- 汉斯·洛里茨于1930年9月加入了党卫军。
- 汉斯·洛里茨于1930年8月加入了国家社会主义德国工人党。
- 国家社会主义德国工人党即纳粹党。
我的思维：
根据证据，汉斯·洛里茨是纳粹党（国家社会主义德国工人党）的成员。
答案：{nazi party}
菲娜：
证据：
- 提供的背景中提到了“保护队”（党卫军），它是阿道夫·希特勒治下的一支主要准军事组织，且纳粹党（国家社会主义德国工人党）。
- 海因里希·希姆莱是纳粹党的高级成员，并担任党卫军首领。
- 背景中还提到，纳粹党（国家社会主义德国工人党）在1920年至1945年间处于活跃状态，并奉行纳粹主义。
我的思维：
已知条件关于党卫军及其与纳粹党的联系，以及直接提及了纳粹党（国家社会主义德国工人党），由此合乎逻辑地得出结论，与党卫军相关的汉斯·洛里茨会是纳粹党党员。
答案：纳粹党

案例-2：
艾米丽：
@菲奥娜：请求获取电影《第八天》的信息以确定其类型。请提供相关数据。
菲娜：
@艾米莉：电影《第八天》被描述为一部比利时剧情喜剧。它讲述了一位商人与一位唐氏综合征患者之间友谊的故事。该片包含带有魔幻现实主义元素的梦境片段。它并未被明确归类为悲喜剧，但喜剧与戏剧的结合表明它也可能符合该类别。证据：
- 被描述为一部剧情喜剧电影。
- 包含魔幻现实主义元素。
- 未明确标注为悲喜剧，但多种类型的融合蕴含了潜在的契合性。
我的想法：根据证据，《第八天》的类型与“喜剧”和“戏剧”相符，并且由于这些元素的融合，可能也属于“悲喜剧”。

建议：我们可以将“喜剧”、“戏剧”和“悲喜剧”视为可能的答案。请交叉验证并确认。埃利

my:
@菲奥娜: 确认《第八天》是一部剧情喜剧，由于其融合了喜剧和戏剧元素，这与“悲喜剧”这一类型相符，原因是它的融合喜剧和戏剧元素。
菲奥娜: @艾米莉:证据:
- 剧情喜剧是悲喜剧的一个子类型。
- 《第八天》被描述为一部剧情喜剧。
我的思维：根据证据，“悲喜剧”是涵盖《第八天》元素的最准确类型。
结论：答案：悲喜剧

表63：Wiki Hop（GPT-4）的示例

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

Elimy:
Context
```json
{
  "Ruggero Deodato": {
    "nationality": "Italian",
    "occupation": ["film director", "screenwriter", "actor"]
  },
  "Alejandro Springall": {
    "nationality": "Mexican",
    "occupation": ["film director"]
  }
}
```
Question
```json
{
  "question": "Are Ruggero Deodato from Italy, and Mexican Alejandro Springall, both film directors?"
}
```
Foina:
Answer: ```json
{
 "response": "Yes, Ruggero Deodato from Italy and Alejandro Springall from Mexico are both film directors."
}

Elimy:
Context:
- Pollokshields Burgh Hall
- President of the Queen's Privy Council for Canada
- Accession Council
- Great Hall of the People
- St James's Palace
Question:
What building is opposite the ceremonial meeting place of the Accession Council in the United Kingdom?
Foina:
Answer:
```json
{
  "Building": "Great Hall of the People"
}
```
[...]

Table 64: Examples of Hotpot QA (GPT-3.5)

Elimy:
Hey Fiona, I think we can solve this question by counting the number of bands each musician has been a member of. We can use a simple list of bands for each musician and then compare the lengths of the lists to find the answer.
Pete Doherty: ["The Libertines", "Babyshambles", "Peter Doherty and the Puta Madres"]
Jonny Craig: ["Dance Gavin Dance", "Emarosa", "Ghost Runner on Third", "Slaves", "westerHALTS", "Isles & Glaciers"]
[...]

Table 65: Examples of Hotpot QA (GPT-3.5, GPT-4)

Elimy:
@Fiona: Requesting information on the actress who played "Celene" in "Walk All over Me" and her roles in "Starcraft 2" trilogy.
Foina:
@Emily: Tricia Helfer played "Celene" in "Walk All over Me" and voiced Sarah Kerrigan in the "Starcraft 2" trilogy.
Elimy: Answer: Sarah Kerrigan

Table 66: Examples of Hotpot QA (GPT-4, GPT-3.5)

艾米丽： **背景**```json { "鲁杰罗·迪奥达托": { "国籍": "意大利", "职业": ["电影导演", "编剧", "演员"]}, "亚历杭德罗·斯普林戈尔": { "国籍": "墨西哥", "occupation": ["电影导演"]} }、、、**问题**```json { "question": "Are Ruggero Deodato from Italy, and Mexican Alejandro Springall, both film directors?" } 菲娜： **答案：** ```json { "response": "Yes, Ruggero Deodato from Italy and Alejandro Springall from Mexico are both film directors." } 艾米丽： **背景：** - 波洛克希尔斯市政厅 - 加拿大枢密院议长 - 继位委员会 - 人民大会堂 - 圣詹姆斯宫 **问题：** 在英国，与继位委员会的仪式性会址相对的建筑是什么？ 菲娜： **答案：**```json { "Building": " Great Hall of the People" }、、[...]

表64：Hotpot QA 示例（GPT-3.5）

艾米丽：嘿，菲奥娜，我想我们可以通过统计每位音乐家曾是成员乐队的数量来解答这个问题。我们可以为每位音乐家列一个简单的乐队列表，然后比较列表的长度以找到答案。皮特·多赫蒂： ["解放者乐队"、"巴比珊布尔斯乐队"、"彼得·多赫蒂与母牛乐队"]乔尼·克雷格： [...] ["Dance Gavin Dance"、"Emarosa"、"Ghost Runner on Third"、"Slaves"、"westerHALTS"、"Isles & Glaciers"]

表65：Hotpot QA 示例（GPT-3.5、GPT-4）

艾米丽：@菲奥娜：请求获取在《Walk All Over Me》中饰演"塞琳"的女演员的信息，以及她在《星际争霸2》三部曲中的角色。菲娜：@艾米莉：特里西亚·海尔弗在《Walk All Over Me》中饰演了"塞琳"，并在《星际争霸2》三部曲中为莎拉·凯瑞甘配音。艾米丽：答案：莎拉·凯瑞甘

表66：Hotpot QA 示例（GPT-4、GPT-3.5）

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

Elimy:
@Fiona: Requesting information on the designer of a 1917 Progressive Era community.
Fiona:
@Emily: Context indicates Howard Van Doren Shaw designed Deerpath Hill Estates in 1926. Not applicable for 1917 community.
Elimy:
@Fiona: Context provided states Howard Van Doren Shaw designed Marktown in 1917. This matches the query for a 1917 Progressive Era community.
Answer: Marktown
Fiona:
@Emily: Confirming receipt of information on Marktown designed by Howard Van Doren Shaw in 1917.
Answer: Marktown

Elimy:
Based on the provided context:
Jonny Craig has been a member of the following bands:
- Dance Gavin Dance
- Emarosa
- Ghost Runner on Third
- Slaves
- westerHALTS
Total bands: 5
Fiona, please provide the number of bands Pete Doherty has been a member of according to your context.
Foina:
Based on the provided context:
Pete Doherty has been a member of the following bands:
- The Libertines
- Babyshambles
- Peter Doherty and the Puta Madres
Total bands: 3
Answer: Jonny Craig

Table 67: Examples of Hotpot QA (GPT-4)

艾米丽：@菲奥娜：请求获取1917年进步时代社区的设计师信息。菲娜：@艾米莉：背景表明霍华德·范·多伦·肖在1926年设计了Deerpath Hill Estates。不适用于1917年的社区。艾米丽：@菲奥娜：提供的背景信息指出，霍华德·范·多伦·肖在1917年设计了马克镇。这与查询中关于1917年进步时代社区的描述相符。答案：马克镇 菲娜：@艾米莉：确认已收到关于1917年由霍华德·范·多伦·肖设计的马克镇的信息。答案：马克镇

艾米丽：
基于提供的背景：
乔尼·克雷格曾是以下乐队的成员：
- Dance Gavin Dance
- Emarosa
- Ghost Runner on Third
- Slaves
- westerHALTS
乐队总数：5
菲奥娜，请根据你提供的背景说明皮特·多赫蒂曾作为成员参加过的乐队数量。
菲娜：
根据提供的背景：
皮特·多赫蒂曾是以下乐队的成员：
- 解放者乐队
- 巴比珊布尔斯乐队
- 彼得·多赫蒂与母牛乐队
乐队总数：3
答案：乔尼·克雷格

表67：Hotpot QA（GPT-4）示例

Elimy:
Context:
- The context is a passage from the story of Peter Pan in Kensington Gardens by J. M. Barrie. The passage describes the adventures of Maimie, a little girl, in the magical world of fairies and the little house she encounters.
Question:
- What is the crow's name?
Your message to Fiona:
- "The crow's name is not mentioned in the given context. We need to look for more information to find the answer."
Foina:
Your message to Emily:
- "Confirmed. The provided context does not contain the crow's name. No further information available in my context. Suggest checking your context for the crow's name."

Table 68: Examples of Narrative QA (GPT-3.5, GPT-4)

艾米丽：**背景：** - 背景是J. M. 巴里的《肯辛顿花园里的彼得·潘》中的段落。该段落描述了小女孩梅米在仙女的魔法世界中以及她遇到的小房子的冒险。**问题：** - 乌鸦的名字是什么？**你发给菲奥娜的消息：** - “乌鸦的名字在已知条件中没有提到。我们需要查找更多信息来找到答案。”菲娜：**你发给艾米莉的消息：** - “确认。提供的已知条件中没有包含乌鸦的名字。我的已知条件中没有更多信息。建议查看你的已知条件以获取乌鸦的名字。”

Elimy:
I have the context that includes information about Mr. Jackson and his appearance in Mrs. Tittlemouse's house. I will now extract the relevant data to answer the question about which of Mr. Jackson's articles of clothing is dripping water.
Extracting relevant data:
- Mr. Jackson is described as having water dripping off his coat tails.
- He is sitting with his feet on the fender, implying that his coat is the source of the dripping water.
Conclusion:
- The article of Mr. Jackson's clothing that drips water is his coat.
Communicating conclusion to Fiona:
- Article of clothing: Coat
- Action: Dripping water
Answer: coat
Foina:
Got it. Based on the information you provided, the article of Mr. Jackson's clothing that drips water is his coat.
Answer: coat

Table 69: Examples of Narrative QA (GPT-4, GPT-3.5)

艾米丽：我拥有包含杰克逊先生及其在蒂特尔鼠标夫人房子中出现的背景信息。我现在将提取相关数据来回答关于杰克逊先生的哪件衣服在滴水的问题。提取相关数据： - 杰克逊先生被描述为外套下摆在滴水。 - 他坐着，双脚放在挡泥板上，这意味着他的外套是滴水的来源。结论： - 杰克逊先生滴水的衣服是他的外套。向菲娜传达结论： - 衣服：外套 - 动作：滴水 答案：外套 菲娜：明白了。根据你提供的信息，杰克逊先生滴水的衣服是他的外套。答案：外套

表69：叙事问答（GPT-4、GPT-3.5）示例

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

Elimy:
@Fiona: Requesting information on contents used to fill the sack after bunnies' escape.
Foina:
@Emily: The provided context does not contain the information required to answer the question about the contents of the sack. Please provide the relevant details from your context.
Elimy:
@Fiona: The contents used to fill the sack after the bunnies escape are "three rotten vegetable marrows, an old blacking-brush and two decayed turnips." Please confirm.
<A>three rotten vegetable marrows, an old blacking-brush and two decayed turnips</A>

Table 70: Examples of Narrative QA (GPT-4)

1. Initial State Representation:
The initial state is that the coin is heads up.
2. Step-by-Step Solution Process:
- oshan flips the coin: The coin will change from heads to tails.
- quentin flips the coin: The coin will change from tails to heads.
- hayaat flips the coin: The coin will change from heads to tails.
- polewarzyk flips the coin: The coin will change from tails to heads.
3. Concluding with the Answer:
The coin is still heads up.
Answer: Yes

Table 71: Example response of Coin Flip (GPT-3.5)

Initial State Representation: The coin is initially heads up.
Step-by-Step Solution Process:
1. Itzamar flips the coin. State changes; coin: tails up.
2. Clione flips the coin. State changes; coin: heads up.
3. Alely does not flip the coin. State remains unchanged; coin: heads up.
4. Sieghardt flips the coin. State changes; coin: tails up.
Answer: no

Table 72: Example response of Coin Flip (GPT-4)

[...]
House 1:
- Cannot be the pianist
- Cannot be the person eating oranges
- Cannot have a television
House 2:
- Cannot be the person eating oranges
- Cannot have a television
House 3:
- Cannot be the person eating kiwis
- Cannot be the person wearing boots
[...]

Table 73: Example response of Logic Grid (GPT-3.5)

[...]
Now we can fill in the details:
- House 1: Minivan, Daffodils, History Book, (not cat lover)
- House 2: Bird Keeper, Carnations, Romance Book
- House 3: SUV, (not cat lover), (not grapes)
- House 4: Convertible, Grapes
[...]

Table 74: Example response of Logic Grid (GPT-4)

艾米丽: @菲奥娜: 请求提供兔子逃跑后用于填充麻袋的内容物的信息。菲娜: @艾米莉: 提供的背景中不包含回答关于麻袋内容物问题所需的信息。请提供您背景中的相关细节。艾米丽: @菲奥娜: 兔子逃跑后用于填充麻袋的内容物是“三个腐烂的蔬菜西葫芦、一个旧鞋油刷和两个腐烂的芜菁”。请确认。<A>三个腐烂的蔬菜西葫芦、一个旧鞋油刷和两个腐烂的芜菁</A>

表70：叙事问答示例（GPT-4）

1. 初始状态表示：初始状态是硬币正面朝上。2. 逐步求解过程：
- 奥山翻转硬币：硬币将从正面变为反面。 - 昆廷翻转硬币：硬币将从反面变为正面。 - 海亚特翻转硬币：硬币将从正面变为反面。 - 波瓦日克翻转硬币：硬币将从反面变为正面。3. 以答案结尾：硬币仍然是正面朝上。答案：是

表71：抛硬币示例响应（GPT-3.5）

初始状态表示：硬币初始状态为正面朝上。逐步求解过程：1. 伊察马尔翻转硬币。状态改变；硬币：反面朝上。2. 克莱奥翻转硬币。状态改变；硬币：正面朝上。3. 阿莱利不翻转硬币。状态保持不变；硬币：正面朝上。4. 齐格哈特翻转硬币。状态改变；硬币：反面朝上。答案：否

表格 72：抛硬币示例响应（GPT-4）

[...]
1号房屋:
- 不可能是钢琴家
- 不可能是吃橙子的人员
- 不可能有电视
2号房屋:
- 不能是吃橙子的
- 不能有电视
3号房屋:
- 不能是吃猕猴桃的人
- 不能是穿靴子的人
[...]

表 73：逻辑网格示例响应（GPT-3.5）

[...]
现在我们可以填入细节:
- 1号房屋：小型厢式车，水仙花，历史书，（非爱猫者）
- 2号房屋：鸟主人，康乃馨，浪漫小说
- 3号房屋：运动型多用途车，（非爱猫者），（非葡萄）
- 4号房屋：敞篷车，葡萄
[...]

表 74：逻辑网格的示例响应（GPT-4）

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

[Olivia]: Speaking:
Structured language for efficient communication.
Initial State Representation:
- Psychologist's tires slashed between 10:30 AM and 11:00 AM.
- Four "red-zone" patients are suspects: Mrs. Roberts, Randy Peters, Mr. Carlson, Mr. Leamington.
Step-by-Step Solution Process:
- Mrs. Roberts shows frustration but lacks a clear motive or physical capability for tire slashing.
- Randy Peters is visually impaired, but his ability to pick up pens suggests he could perform the act. His dream indicates subconscious aggression towards the psychologist.
- Mr. Carlson has a germ phobia, which would likely prevent him from slashing tires, but he expressed a desire to take out his frustration on someone.
- Mr. Leamington shows impulsive behavior and anger issues, but no direct link to the tire slashing incident.
Re-evaluation of Evidence:
- Mrs. Roberts and Mr. Leamington lack a clear motive or opportunity.
- Randy Peters' dream and subconscious aggression towards the psychologist, combined with his ability to navigate without clear vision, make him a plausible suspect.
- Mr. Carlson's expressed frustration and ability to overcome his phobia to sit in the chair suggest he could have slashed the tires.
Conclusion:
- Randy Peters and Mr. Carlson are the most likely suspects based on the evidence provided.
- Randy Peters' dream and actions during the session suggest a subconscious link to aggression towards the psychologist, making him the most likely suspect.
[...]

Table 75: Example response of Minute Mysteries QA (GPT-4)

[奥利维亚]: 发言：结构化语言以实现高效通信。初始状态表示： - 心理学家在上午10:30至11:00之间被割破轮胎。 - 四位“红色区域”患者是嫌疑人：罗伯茨夫人、兰迪·彼得斯、卡尔森先生、利明顿先生。 逐步求解过程： - 罗伯茨夫人表现出挫败感，但缺乏明确的动机或割破轮胎的体力条件。 - 兰迪·彼得斯视力受损，但他能够捡起笔，这表明他有可能实施该行为。他的梦显示了对心理学家潜意识的攻击性。 - 卡尔森先生有细菌恐惧症，这通常会阻止他割破轮胎，但他表达了想对某人发泄挫败感的愿望。 - 利明顿先生表现出冲动行为和愤怒问题，但没有与割胎事件直接相关的线索。 证据重新评估： - 罗伯茨夫人和利明顿先生缺乏明确的动机或机会。 - 兰迪·彼得斯的梦和对心理学家的潜意识攻击性，加上他能够在视力不清晰的情况下移动的能力，使他成为合理的嫌疑人。 - 卡尔森先生表达的挫败感以及他能够克服恐惧症坐在椅子上的能力，表明他可能割破了轮胎。 结论： - 根据所提供的证据，兰迪·彼得斯和卡尔森先生是最可能的嫌疑人。 - 兰迪·彼得斯的梦和在会谈期间的行为表明他与对心理学家的攻击性之间存在潜意识联系，使他成为最可能的嫌疑人。 [...]

表75：Minute Mysteries 问答的示例响应（GPT-4）

Initial State Representation:
- Socks: P, Q, R, S, T, U
- Pairs: 3 pairs in total
- P pairs with neither Q nor R
Step-by-Step Solution Process:
1. Let's start by analyzing statement 1: "Neither S nor T pairs with either Q or R."
- This statement provides information about the pairs involving S and T, but it doesn't directly address the pairs involving P, Q, and R. Therefore, statement 1 alone is insufficient to determine the pairs.
2. Now, let's consider statement 2: "Neither P nor Q pairs with either S or U."
- This statement provides information about the pairs involving P and Q, but it also doesn't directly address the pairs involving S, T, R, and U. Therefore, statement 2 alone is also insufficient to determine the pairs.
3. Next, let's analyze both statements together:
- When we combine both statements, we have information about the pairs involving P, Q, R, S, T, and U. However, the information is still not sufficient to uniquely determine the pairs.
Conclusion:
Neither statement 1 nor statement 2 nor statements 1 and 2 taken together are sufficient.

Table 76: Example response of Information Essentiality (GPT-3.5)

初始状态表示： - 袜子：P、Q、R、S、T、U - 配对：共3对 - P不与Q配对，也不与R配对 **逐步求解过程：** 1. 让我们先分析陈述1：“S和T都不与Q或R配对。” - 该陈述提供了关于S和T所参与配对的信息，但并未直接涉及P、Q和R的配对情况。因此，仅凭陈述1不足以确定配对关系。 2. 现在，让我们考虑陈述2：“P和Q都不与S或U配对。” - 该陈述提供了关于 P 和 Q 所参与配对的信息，但同样并未直接涉及 S 、 T 、 R 和 U 的配对情况。因此，仅凭陈述2也不足以确定配对关系。 3. 接下来，让我们联合分析这两个陈述： - 当我们结合两个陈述时，我们获得了关于P、Q、R、S、T和U所参与配对的信息。然而，这些信息仍不足以唯一确定配对关系。 **结论：** 无论是陈述1、陈述2，还是将陈述1和陈述2结合起来，都不足以确定配对关系。

表格 76：信息必要性（GPT-3.5）的示例响应

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next

Representation after Statement 1:

```
  ?       ?
H   E   ?   ?
  ?       ?
```

Statement 1 alone is insufficient.
Statement 2: F is sitting opposite to H.
This statement tells us the exact positions of F and H. Since there are only four players, we can deduce that E and G must be sitting to the left and right of F and H, respectively. However, we do not know which one is to the left and which one is to the right.
Representation after Statement 2:

```
  E/G     G/E
F   ?   ?   H
  ?       ?
```

Statement 2 alone is insufficient.
Now, let's combine both statements:
Statement 1 and 2 together:
From Statement 1, we know E is to the right of H. From Statement 2, we know F is opposite H. Therefore, G must be to the left of H, and E must be opposite G.
Final Representation:

```
  E       G
F   H   E   H
  G       F
```

Combining statements 1 and 2, we can determine that G is sitting opposite to E.

Statement 1: "Q is to the north of R, who is to the west of P."
- If Q is north of R, we can represent this as Q ↑ R.
- If R is west of P, we can represent this as R ← P.
- Combining these two pieces of information, we can deduce that P is northeast of R and southeast of Q.
Representation:

```
P
↑↖
R ← Q
```

From this representation, we can conclude that P is to the southeast of Q.
[...]

Table 77: Example responses of Information Essentiality (GPT-4)

陈述1之后的表示：

```
  ?       ?
H   E   ?   ?
  ?       ?
```

仅陈述1是充分的。
陈述2：F坐在H的对面。
此陈述告诉我们F和H的精确位置。由于只有四名玩家，我们可以推断E和G必须分别坐在F和H的左侧和右侧。然而，我们不知道哪一位在左侧，哪一位在右侧。

陈述2之后的表示：

```
  E/G     G/E
F   ?   ?   H
  ?       ?
```

仅陈述2是充分的。
Now, 让我们将这两个陈述结合起来：
陈述1和陈述2合在一起：
根据陈述1，我们知道E在正面的右侧。根据陈述2，我们知道F与正面相对。因此，G一定在正面的左侧，并且E一定与G相对。
最终表示：

```
  E       G
F   H   E   H
  G       F
```

结合陈述1和陈述2，我们可以确定G坐在E的对面。

陈述1："Q 在 R 的北面，而 R 在 P 的西面。" - 如果 Q 在 R 的北面，我们可以将其表示为 Q ↑ R。 - 如果 R 在 P 的西面，我们可以将其表示为 R ← P。 - 结合这两条信息，我们可以推断出 P 在 R 的东北面和 Q 的东南面。 表示：、、、P↑↖R ← Q、、、根据此表示，我们可以得出结论：P 在 Q 的东南面。 [...]

表 77：信息必要性（GPT-4）的示例响应

翻译服务由 ChatECNU 大语言模型平台提供 https://chat.ecnu.edu.cn/html/
技术方案基于 PDFMathTranslate-next 项目 https://github.com/PDFMathTranslate/PDFMathTranslate-next