2642♂

什么是一个好的 Natural Language 提示？

Do Xuan Long[1,3], **Duy Dinh**[1*], **Ngoc-Hai Nguyen**[1*],
Kenji Kawaguchi[1], **Nancy F. Chen**[3], **Shafiq Joty**[2], **Min-Yen Kan**[1]
[1]National University of Singapore, [2]Salesforce AI Research,
[3]Institute for Infocomm Research (I^{2}R), A*STAR
xuanlong.do@u.nus.edu, { dinhcongduy131200, haibeo2552001 } @gmail.com,
{ kenji,knmnyn } @nus.edu.sg, sjoty@salesforce.com, nfychen@i2r.a-star.edu.sg

Abstract

随着大型语言模型（LLM）向更类人化且人机交流日益普遍的方向发展，提示词已经成为一个决定性组成部分。然而，关于什么确切量化了自然语言提示词，尚未达成概念上的共识。我们尝试通过对来自领先 NLP 和 AI 会议（2022–2025）以及博客的 150 多篇与提示词相关的论文进行元分析来解决这个问题。我们提出了一个以属性和人为中心的框架用于评估提示词质量，涵盖 21 个属性，这些属性被归类为六个维度。然后我们检验现有研究如何评估这些属性对 LLM 的影响，揭示其对不同模型和任务支持不平衡的问题以及显著的研究差距。此外，我们分析高质量自然语言提示词中属性之间的相关性，从中得出提示词推荐。随后，我们在推理任务中对多属性提示词增强进行实证探索，发现单一属性增强通常具有最大的影响。最后，我们发现对增强属性的提示词进行指令微调可以得到更好的推理模型。我们的研究结果为以属性为中心的提示词评估和优化奠定了基础，弥合了人与 AI 沟通之间的差距，并开启了新的提示词研究方向[1]。

1 引言

预训练 LLMs (Brown et al., 2020; Chowdhery et al., 2022; OpenAI, 2022; Touvron et al., 2023a; Team et al., 2023; Guo et al., 2025) 以其生成类似人类文本的能力而闻名，并且在各种自然语言处理任务中表现出色。尽管它们的有效性受到自然语言提示 (Sahoo et al., 2024) 质量的深刻影响，但有效提示的艺术和科学仍未被充分探索。随着人机交互变得普遍，深入理解这些自然语言提示至关重要，因为它们是人类与人工智能系统之间的主要交流界面。

尽管理解自然语言提示非常重要，但在如何量化它们方面仍然缺乏共识。当前的方法主要依赖于以结果为中心的测量，如特定模型的性能指标 (Deng et al., 2022; Lin et al., 2024; Shi et al., 2024) 以及可能导致提示被优化为机器解释而非人类理解的迭代试错测试 (Pryzant et al., 2023; Long et al., 2024a)。这可能导致在解释和验证它们时遇到困难，并可能在大型语言模型中引入对抗性行为 (Zou et al., 2023; Zhu et al., 2023)，引发关于一致性、透明度、整体可靠性，甚至人类与人工智能通信方面的担忧。

最近，一些提示研究 (Bsharat et al., 2023; Lin, 2024) 和指南 (OpenAI, 2024b; Anthropic, 2024) 引入了增强某些提示属性的建议，例如“指定输出的所需长度”。这些以属性为中心的建议侧重于提示质量而非模型性能，提供了可解释的策略，并可以补充以结果为中心的方法。然而，它们有关键的局限性。首先，没有统一或理论上以属性为中心的框架在抽象上涵盖此类实际建议，这阻碍了对这些策略的系统理解、分析和比较。其次，不清楚这些建议是否在跨模型和任务中提供普遍的好处，还是更倾向于特定的模型或任务。第三，这些建议之间的相互作用及其对模型性能的综合影响仍然未被充分研究。

为了应对这些限制，我们提出了一项元分析来系统地研究自然语言提示。我们调查了 2022 年至 2025 年顶级 NLP 和 AI 会议的提示论文以及顶级科技公司撰写的博客（完整列表见 §B），并在六个评估维度中识别了 21 个提示级属性，从而提供了一种新颖的以属性和人为中心的视角（§3）。基于此，我们研究了先前的研究如何评估通过增强每个属性来使哪些模型和任务受益，揭示了支持每个属性的 # 论文在模型和任务中存在显著的不平衡分布以及研究空白（§4）。接下来，我们分析了这些属性在一个高质量自然语言提示子集中的相关性，得出提示设计的实际建议（§5）。随后，我们进行了一个关于推理任务的案例研究，以了解增强多个提示属性对模型性能的影响（§6）。值得注意的是，我们观察到不同的提示属性在任务中对模型的影响不同，增强多个属性并不总是带来更大的提升；单一属性通常是最有效的，并且在属性增强的指令上对模型进行微调会进一步提高这种有效性。我们的贡献总结如

*Equal contribution. Works done during the internship at WING, NUS.

[1]我们的代码和数据将在 这里 公开提供。

下：

1. 我们引入了一种新颖的以属性和人为为中心的框架，用于评估自然语言提示的质量，识别了跨越六个评估维度的 21 个关键属性，以实现从以结果为中心到以属性为中心的评估转变。
2. 我们对 2022-2025 年 NLP/AI 会议和博客中的先前研究进行了荟萃分析，调查这些属性如何影响模型性能，揭示了显著的研究不平衡和空白。
3. 我们分析了一组高质量提示中的这些属性之间的相关性，并得出实用的建议以指导有效的提示设计。
4. 我们研究了推理任务中的提示和微调模型，发现优化单一提示属性通常比结合多个属性表现更优，且这种效果在不同的任务和模型中有所不同。

2 相关工作

提示在发挥大型语言模型（LLMs）全部潜力中起关键作用，引发了对提示分析研究的广泛兴趣。现有研究主要集中在两个主要方向。第一个方向分析提示的结构组成部分，强调它们在格式和措辞方面的变化如何导致性能上的显著差异，以及它们的出现频率。这些研究旨在理解提示组件及其对模型性能的影响。第二个方向通过实际实验分析提示，提供设计建议，如链式思维提示、对 LLMs 保持礼貌，甚至是一些通用指导原则。然而，这些提示分析研究往往是任务特定的，或关注于提示的特定属性。在这项工作中，我们首次引入一个统一的属性中心框架，该框架抽象地组合这些实际建议，便于系统理解、分析和比较提示策略。

提示工程与优化。 提示工程 (Wei et al., 2022; Zhang et al., 2023; Zhou et al., 2023c) 和优化 (Deng et al., 2022; Pryzant et al., 2023; Long et al., 2024a) 旨在寻找能够最大化语言模型在特定任务中表现的提示。尽管现有的大多数研究集中在提高基准性能上，但最近有一些新兴的努力强调更广泛的提示属性，例如清晰度 (Lin, 2024; Anthropic, 2024) 、礼貌性 (Bsharat et al., 2023; Yin et al., 2024) 、结构化格式 (OpenAI, 2024b) ，甚至是输出生成中的公平性 (Ji et al., 2023; Yuan et al., 2023) 。然而，目前尚不清楚这些属性是否能在不同模型和任务中提供普遍的好处，或者它们的效果是否是特定于模型或任务的。此外，它们的相互作用及其对模型性能的综合影响仍然基本未探索。我们在 Sections 4 to 6 中解决了这些差距。

3 提示质量评估

我们开始研究，通过对超过 150 篇论文和博客进行全面调查。我们的方法很简单：首先，我们检查从 2022 年至 2025 年在 ACL Anthology [2] 上发表的 ACL、EMNLP、NAACL 的论文，以及在 OpenReview [3] 上发表的 ICLR 和 NeurIPS 的论文。通过在 Google 上进行关键词搜索，进一步识别相关论文。在力求全面的同时，我们也承认可能会无意中遗漏一些相关的论文。然后，我们手动识别这些论文中影响模型性能的提示目标和建议，并将其概念化为提示属性。这些属性及其证据（缩写为 e.b.）定义如下。

一、**交流与语言。** 先前的研究强调了特定的沟通特性对于获得期望的语言模型结果的重要性。例如，Yin et al. (2024) 发现不礼貌的提示会在各种任务和语言中降低模型结果，而 Shi et al. (2023) 发现不相关的背景会分散大型语言模型的注意力，更明确的提示则能够提升模型性能 (Bsharat et al., 2023; Lin, 2024) 。受到这些研究的启发以及大型语言模型更具 humanoid 特性的影响，我们认为在评估提示时应该考虑类似人类的沟通特性。我们引入了四个用于评估的特性，部分受到 Grice 会话准则的激励 (Grice, 1975)：

- 令牌数量：提示在提供最佳和相关信息的同时最小化令牌使用的程度，在信息完整性和效率之间取得平衡（例如，Shi et al. (2023); Jiang et al. (2023b) ）。
- 方式：提示的清晰和直接程度（跨回合），同时最大限度地减少不必要的歧义、复杂性和混淆（例如 Anthropic (2024) ）。
- 互动和参与：提示在多大程度上通过提出澄清或确认的问题（例如 Deng et al. (2023) ）来明确鼓励模型收集必要的细节和要求。
- 礼貌：提示保持尊重、专业和上下文特定礼貌的程度，包括礼貌语言的使用（例如，“请”、“谢谢”）（例如，Yin et al. (2024) ）。

二、**认知。** Wei et al. (2022); Zhou et al. (2023a) 首创了引入提示方法，将复杂的推理任务分解为更简单的步骤，从而提升了大型语言模型（LLM）的性能。后续研究广泛探讨了优化子任务的策略，以进一步使其符合模型的能力 (Khot et al., 2023; Suzgun and Kalai, 2024) 。此外，Sun et al. (2022) 表明整合自我生成的知识

[2] https://aclanthology.org/
[3] https://openreview.net/

Property	Real-world chat AlpacaEval/ATLAS/ ShareGPT/...	Eval. suit MMLU/C-Eval/ BIG-Bench/...	Reasoning/QA GSM8K/Comm.QA/ HotpotQA/ELI5/...	Generation CNN/Arxiv-March23/ HumanEval/Translation/...	NLU GLUE/CommitmentBank/ DBPedia/...	Others Safety/Persona./ Judging/Retrieval/...
Better quantity	4	4	9	4	1	0
Better manner	0	0	0	0	0	0
Better engagement	2	0	1	2	0	1
Better politeness	1	2	1	4	2	2
Better intrinsic	3	2	7	2	3	8
Lower extraneous	0	1	3	0	0	3
Better germane	1	1	2	1	0	0
Better objective(s)	1	1	1	1	1	0
Better external tool(s)	1	2	2	1	0	1
Better metacognition	0	2	2	0	1	1
Better demo(s)	1	2	8	4	3	1
Better reward(s)	1	2	2	1	0	1
Better structure	1	1	4	2	1	0
Better context logic	0	0	1	0	0	1
Better hallu. awa.	0	0	1	1	0	0
Better fact. and cre.	0	0	0	0	0	0
Lower bias	1	0	0	1	0	2
Better safety	0	0	0	0	0	1
Better privacy	0	0	0	0	0	1
Better reliability	0	1	1	0	0	1
Better societal norms	0	0	0	0	0	0

Table 1: Summary of the number of papers supporting specific properties across various tasks and models. Model logos are used as follows: : ChatGPT / Codex; : LLaMa / OPT / RoBERTa / BART; : Qwen; : Mistral; : Alpaca; : Yi; : PaLM / FLAN / Gemma; : BLOOM / LongChat / T0; : ChatGLM; : Claude; : Command R; : DeepSeek; : EleutherAI; : InternLM; : LLaVa; : mDeBERTa / Orca / WizardLM; OFA : OFA; : OpenChat; : Pegasus; : PolyLM; : Swallow; : Vicuna; : XGLM。支持各种属性的论文分布在模型和任务之间是高度不平衡的。我们将在 §4 中详细讨论结果。

可以提高 LLM 的问答性能。从哲学上讲，这些工作意味着，要最大化 LLM 的学习和问题解决能力，必须仔细管理它们的认知负荷。

Sweller and Chandler (1991) 引入了认知负荷理论，将认知负荷分类为 固有（任务复杂性）、外来的 （不清晰或设计不佳的指令）和 锗烷（理解、记忆和组织信息的努力）。受此启发，提示评估应关注 LLMs 上的三种负荷：

- 管理内在负荷：这评估了提示在显式指导模型将复杂任务分解为与语言模型技能（例如 Zhou et al. (2023a) ）一致的可执行步骤方面的效果。
- 减少无关负担：提示通过简化语言和移除冗余或不相关的信息以减轻不必要的复杂性程度（例如 OpenAI (2024b) ）。
- 鼓励相关负荷：提示在多大程度上明确地让模型与其先验知识或深层工作记忆（例如，“自问”(Press et al., 2023) ）进行互动，以便将其与现有和新知识结合用于解决问题（例如 Sun et al. (2022); Mialon et al. (2023); Fan et al. (2024) ）。

III. 指导。 提示的教学价值对于实现期望输出 (Sahoo et al., 2024) 至关重要。借鉴加涅的九项教学事件 (Gagné, 1985) 和元认知理论 (Schraw and Moshman, 1995) ，我们提出教学标准来评估它们，与其他维度不重叠。

- 目标：提示在多大程度上明确传达任务目标，包括预期的人物角色、输出结果、格式、限制、受众以及其他适用的标准（例如 Chang (2023); Long et al. (2025b) ）。
- 外部工具：提示明确引导模型识别何时需要超出任务目标的特定外部工具或知识资源的程度，以及执行相应外部调用（例如 Yao et al. (2023) ）。
- 元认知：其评估明确指导模型推理、自我监控和自我验证输出，以满足期望并增强可靠性（例如 Wang and Zhao (2024) ）。
- 示例：提示中明确包含示例、演示和反例以说明所需输出的程度（例如 Dong et al. (2024) ）。
- 奖励：提示在多大程度上明确建立反馈和强化机制，以鼓励模型达到期望的输出（例如，Bsharat et al. (2023) ）。

四、逻辑与结构。 事实证明，连贯的结构性提示在各种任务中都很有效 (Wang et al., 2024a; Huang et al., 2024a) 。此外，提示指南 (Guide, 2024; OpenAI, 2024b) 还建议对输入和输出进行结构化，以获得性能更佳的提示。对于逻辑，最近的研究 (Wang et al., 2024g; Pham et al., 2024) 强调了上下文一致性的重要性，其中提示内的知识冲突会严重降低语言模型的性能。

基于这些见解以及已建立的人类逻辑有效沟通标准 (Grice, 1975; Mercier and Sperber, 2011)，我们引入了两个逻辑标准：

- 结构逻辑：这评估了提示的结构在逻辑上的清晰度和连贯性，以及组件之间的进展（例如，Wang et al. (2024a); Zhou et al. (2024b)）。
- 情境逻辑：这用于评估提示内及交互轮次间指令、术语、概念、事实和其他组成部分的逻辑一致性和连贯性（例如，Pham et al. (2024)）。

五. 幻觉。提示可能导致幻觉，其中模型生成看似合理但实际上不准确的内容 (Huang et al., 2024b)。尽管预测一个提示是否以及何时会引发幻觉仍然具有挑战性 (Farquhar et al., 2024)，但可以设计提示来鼓励模型意识到这个关键问题。我们建议，提示评估应解决两个与幻觉相关的标准：

- 幻觉意识：提示在多大程度上明确引导模型生成基于事实和证据的响应，同时尽量减少推测性或无根据的断言（例如，Gao et al. (2023)）。
- 在事实性与创造性之间取得平衡：指明提示如何明确引导模型在创意生成与事实准确性之间取得平衡，包括在哪些任务中及何时强调创造性和事实性孰轻孰重。迄今为止，我们尚未观察到专门为这一标准设计的提示方法。然而，Sinha et al. (2023) 提出了一个用于平衡语言模型这两个方面的训练方法。

在这一维度中，我们不评估提示中的幻觉，因为它与沟通的“数量”部分重叠。

这种维度强调负责任的提示，旨在缓解与包容性、隐私、安全性、偏见、可靠性、公平性、透明性和社会规范相关的担忧，特别是在涉及敏感话题或多样化观众的任务中。

- 偏见：指的是提示在多大程度上不带有偏见，并明确鼓励模型生成不含文化、性别、种族或社会经济偏见的内容，并避免刻板印象（例如，Si et al. (2023b)）。
- 安全性：指提示中不含不安全内容并明确鼓励模型生成安全输出的程度，避免有害内容，如关于危险活动或武器制造的指导（例如，Zou et al. (2023); Zheng et al. (2024a)）。
- 隐私：提示不包含敏感隐私信息的程度，并明确鼓励模型生成不含个人敏感或可识别信息的内容（例如，Edemacu and Wu (2024)）。
- 可靠性：提示在多大程度上能够明确地鼓励显式的推理过程和归因，包括对模型局限性和不确定性的承认（例如 Si et al. (2023b); Long et al. (2024b)）。
- 社会规范：提示在何种程度上排除有害的规范，并明确鼓励模型生成符合广泛接受的文化、伦理和道德标准（例如，Yuan et al. (2024b)）的包容性和适当内容。

4 属性如何影响模型性能？

为了评估 §3 中的属性如何影响模型性能，我们分析了截至目前的调查论文，以确定这些方面是否被研究过。我们将所探讨的任务分类为六组：(1) 现实世界对话，包括从真实用户中收集的基准测试，如 AlpacaEval (Li et al., 2023c) 和 ShareGPT (ShareGPT, 2023)；(2) 评估套件，其中有多个评估任务，如 MMLU (Hendrycks et al., 2021) 和 C-Eval (Huang et al., 2023c)；(3) 推理/问答，涵盖推理和问答任务，如 GSM8K (Cobbe et al., 2021) 和 HotpotQA (Yang et al., 2018)；(4) 生成，专注于文本生成基准测试，如摘要 (Nallapati et al., 2016) 和翻译；(5) 自然语言理解（NLU），包括自然语言理解任务，如 GLUE (Wang et al., 2018) 和 CommitmentBank (De Marneffe et al., 2019)；(6) 其他，包括安全性、个性化、判断和检索任务。我们在下面的 Table 1 中讨论我们的发现作为可操作的提示建议。

首先，在现实世界的聊天中，通信属性是最受到支持的，其次是指令和认知属性。这是因为在实际使用大型语言模型时，用户通常会精心制作丰富且信息量大的提示来处理复杂和多样的任务。这些提示可以扩大到数万个标记，有时可能包含冗余细节或缺乏重点，特别是在多轮交互中。此外，指令属性的重要性反映了聊天的互动性，而认知属性对于实现期望的结果至关重要。其次，对于评估套件，认知、指令和通信属性被研究得最多，其中逻辑在推理/问答任务中尤为强调。这与这些基准的性质一致，其中良好的认知指令对于加强大型语言模型的推理能力至关重要。此外，逻辑和结构逻辑也强调了对于此类任务系统解决方法的重要性。第三，对于生成任务，通信属性得到的支持最多，其次是指令。这种观察反映了在生成任务中高效标记管理的重要性。有趣的是，一些研究强调了礼貌的有效性，可能反映了大型语言

模型在处理良性查询而非非正式查询时的固有偏见。第四，目前针对自然语言理解（NLU）任务的提示研究有限，最常被探索的是指令属性，其次是认知属性。这可以用 NLU 任务要求模型准确解读提示，以便深入推理语言意义或超越表面理解的含义来解释。最后，较低的外部性和更好的保护提示已被证明能够有效增强安全性，更好的内在性可以用于个性化，更好的内在性和较低的偏见可以用于判断，较低的外部性可以用于检索。这些发现虽然突出了任务需求与显示的属性之间的细致对齐，但在探索如何增强其他属性以进一步提高模型在这些任务上的表现方面仍然存在显著的研究空白。

具体来说，OpenAI 的专有模型（CodeX (Chen et al., 2021)，InstructGPT (Ouyang et al., 2022)，ChatGPT (OpenAI, 2022)，GPT-4/4o (OpenAI, 2023, 2024a)）是被研究得最为广泛的，其次是开源的 LLaMa 模型 (Touvron et al., 2023a,b; Dubey et al., 2024)，以及谷歌的模型（FLAN (Chung et al., 2024)，PaLM (Chowdhery et al., 2022)，Gemma (Team et al., 2024)）。我们假设不同的属性对模型有不同的益处，并且这些益处在不同的任务中也可能有所不同，并在 §6 中进行了验证。

我们的分析揭示了任务特定与普遍属性：虽然更好的内在负载管理、示范以及外部工具被认为是普遍有效的，幻觉意识和责任感似乎更具任务特定性。更好的内在负载突出了当前 LLM 在无需明确指导的情况下，隐性且有效地将复杂任务分解为较易管理的子任务的薄弱环节。此外，示范属性强调了从例子中学习的价值，而使用外部工具表明即便降低了认知负担并提供了良好的示范，LLM 仍然能够受益于工具以处理某些任务。

开放问题（Oq）。（Oq1）不同模型由于其固有知识的差异，属性的有效性也会有所不同，因此，一个对某个模型有利的属性是否对另一个模型有用还是一个未解的问题。此外，Table 1 中缺失的条目强调了几个关键但未被探索的属性。例如，（Oq2）虽然推理是人类解决任务 (Pearl, 1998) 的基础，但是否培养更深层次的推理（改善基本认知负荷）、反思行为（增强元认知）或责任感能促进在真实世界的聊天、评估套件和自然语言理解任务中提升大语言模型的结果尚待研究。此外，（Oq3）尽管创造力对生成等多项任务具有直观的重要性，但其对大语言模型的有效性仍然是一个未解的问题。此外，对属性动态的理解仍存在显著差距，特别是（Oq4）某些相关或甚至与任务无关的属性 (Taveekitworachai et al., 2024) 在何种条件下变得有效以及原因。最后，（Oq5）关于特定任务属性和通用属性的观察引发了关于提示工程和优化是否应当优先考虑其一以及哪一个更重要的重要问题。研究（Oq1）-（Oq5）对推动大语言模型的效率、可靠性和一致性具有巨大潜力。未来的研究可以对不同的大语言模型和任务进行比较研究，开发可量化的指标以评估多个维度的提示，并探索结合特定任务和通用提示属性的混合策略。

5 这些属性如何在高质量提示中出现并相互关联？

我们研究高质量的自然语言提示，以调查这些属性之间的相关性，从而得出提示的建议。我们手动收集了我们的测试集，该测试集由 765 个单轮提示组成，来自提示工程论文、ChatGPT 提示集合 [4]、Awesome ChatGPT Prompts [5]、Alpaca (Taori et al., 2023)、Natural Instructions (Mishra et al., 2022)、Complex Instructions (He et al., 2024)，以及 50 个来自 LMSYS-Chat-1M (Zheng et al., 2024b)、包含 204 个提示的真实世界多轮（> 2 轮）对话，总计 969 个提示于附录 Table 4 中。我们使用 GPT-4o-2024-11-20 (OpenAI, 2024a)，设 Self-consistency (Wang et al., 2022) 为评判标准，对这 21 个提议的属性进行了评估。我们还测试了开源模型，包括 DeepSeek R1 Distill Qwen 32B (Guo et al., 2025) 和 Mistral Small 24B It 2501 (Jiang et al., 2023a)，作为评委。然而，由于 DeepSeek 和 Mistral 只取得了 65.42 % 和 71.19 %，并且我们在遵循显著评价格式时遇到了问题 (Long et al., 2025a)，因此最终没有使用它们。除了 GPT-4o，我们还补充了来自 Gemini-2.0-flash (Team et al., 2023) 的发现以支持我们的相关性结果，见附录 **??**。

方法。使用大语言模型进行的自动评估可能不可靠，特别是考虑到评估提示中的可变性 (Doostmohammadi et al., 2024)。这在从这些评估中得出可靠的相关性结论时造成了重大挑战。为了解决这个问题，我们首先在 21 个属性中随机手动标记 50 个提示，然后设计评估提示以与人工判断紧密对齐。每个注释均由我们的三位具有学士学位并至少有六个月经验的提示研究人员一致同意。

对于每个评估维度，我们从一个与 Zheng et al. (2023) 提出的 1-10 分无参考评判提示相似的提示开始。然而，我们发现这种方法与人类评分者的 Cohen's Kappa 一致性结果非常低；在 21 个主题中有 15 个主题得分低于 0.15，参见附录 Fig. 2 和 原始评估。然后，我们为每个

[4] ChatGPT Prompts Collections
[5] https://github.com/f/awesome-chatgpt-prompts

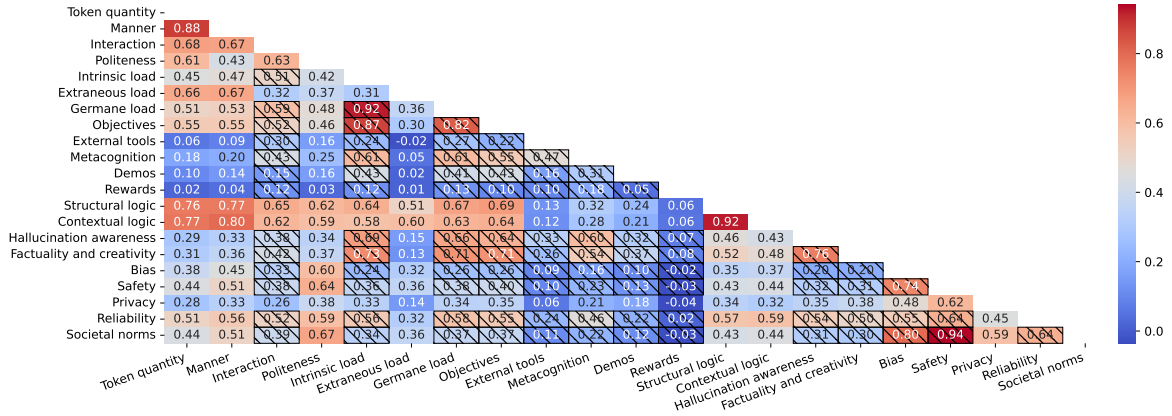

Figure 1: Correlations of properties evaluated by GPT-4o. We do not consider correlations between pairs of properties concurrently having average scores below 5/10 (hatched by "\\") since they naturally but may falsely suggest correlations.

标准补充了类似于 (Yuan et al., 2024a) 的增量评分系统，这显著提高了一致性。然而，相关负担、目标、奖励和责任属性的得分仍然很低。这是因为评估者倾向于根据隐式指令而不是预期中的明确提示给它们比人类更高的评分 。为解决此问题，我们明确指示评估者评判明确的信号，结果显著提高了一致性（附录 Fig. 2 中的 我们的 ）。我们用 "我们的" 评估了所有提示。

发现。 对于这组特定的提示 ，属性相关性在 Fig. 1 中提供。如果两个属性的平均得分均低于 5/10（由 "\\" 标识），我们不考虑属性之间的相关性，因为低平均得分自然会但可能错误地建议相关性。我们观察到在 21 个属性中有 17/210 个强相关性（≥ 0.7）。其中一些与其现实世界中的重叠一致。例如，令牌数量、方式、结构逻辑、上下文逻辑和额外负载反映了令牌效率、清晰度、直接性、排除不相关细节和逻辑一致性之间的自然相关性。在维度中，我们注意到结构逻辑与上下文逻辑强相关；幻觉意识与真实性和创造力；安全性与社会规范。令人惊讶的是，我们注意到目标与内在负载之间以及目标与有益负载之间的强相关性；幻觉意识与可靠性之间的强相关性。这些可以归因于有效的人类提示的性质：当我们优化内在和/或有益负载时，我们倾向于更清晰地表述目标。同样，增强幻觉意识本质上有助于增强可靠性意识。

我们从对这一组提示的分析中学习到提示推荐。首先，优化提示的直接性、清晰性和简洁性可能会提高词汇效率和逻辑一致性，并减少额外的认知负荷。其次，当提示具有逻辑结构时，自然会产生清晰的目标，引导模型自我监控其生成过程或逐步执行任务。第三，在提示中明确加入幻觉意识可能会导致更好的可靠性能意识。最后，由于这些提示是由人类精心选择的，某些不明显的关联，如结构逻辑、上下文逻辑、词汇数量和方式之间的关联，表明这些属性应该共同优化。

开放问题 (Oq)。 虽然我们的分析揭示了提示属性之间的某些相关性，但仍有若干开放性问题需要未来研究。首先，(Oq6) 我们假设不同提示集，尤其是那些特定任务的提示集，之间的相关性可能会有所不同，从而导致不同的提示建议。我们将这一点留待未来研究。其次，(Oq7) 当两个属性表现出强相关性时，尚需确定是在一个属性上增强提示是否会因果地增强另一个属性，还是这些属性仅仅在我们的数据集中同时出现。最后，(Oq8)，了解这些相关性如何影响模型性能对推进提示优化方法至关重要。对（Oq6）-（Oq8）的研究提供了一条通过分析属性相关性和消除优化冗余来优化 LLM 提示的途径。未来的工作可以使用因果推断工具，例如结构方程模型，以区分单纯的共现和影响，并进行多样的模型和任务特定实验，以更精确地量化这些效果。

6 在实验过程中，我们是否应该增强提示的属性？

我们对结合这些属性对模型推理性能的影响进行了初步调查。我们的实验在两种设置下进行：（1）提示（§6.1 ），（2）微调（§6.2 ），并在 MMLU (Hendrycks et al., 2021)，常识问答 (Talmor et al., 2019) 和 ARC-挑战 (Clark et al., 2018)，以及 GSM8K 数据集上进行。

6.1 属性增强提示

我们的提示实验是在 Llama-3.1-8B-it (Dubey et al., 2024)、Qwen2.5-7B-it (Qwen Team, 2024) 和 OpenAI o3-mini (OpenAI, 2025) 上进行的，主要关注三个维度：沟通、认知负担和指导。我们排除了演示、目标和外部工具，因为之前的研究已经广泛探讨了这些特性。我们从零次

		MMLU	Comm.QA	ARC-C	GSM8K
Llama-3.1-8B-It	Zero-shot CoT	65.00	76.00	81.50	82.0
	+ Politeness	68.00↑	83.50 ↑	84.50 ↑	87.5 ↑
	+ Germane load	66.00↑	75.50↓	82.00↑	82.0↓
	+ Metacognition	61.00↓	81.50↑	81.00↓	81.5↓
	+ Rewards	64.00↓	80.50↑	82.00↑	84.0↑
	+ Pol. + Ger.	67.00↑	79.50↑	80.50↓	80.5↓
	+ Met. + Rew.	66.00↑	80.00↑	83.50↑	83.5↑
	+ Pol. + Ger. + Met.	69.50 ↑	75.00↓	82.50↑	81.5↓
Qwen-2.5-8B-It	Zero-shot CoT	45.50	55.00	59.50	76.5
	+ Politeness	41.00↓	45.50↓	54.00↓	79.0↑
	+ Germane load	44.50↓	56.50 ↑	53.50↓	90.0 ↑
	+ Metacognition	52.50 ↑	56.50 ↑	62.00 ↑	83.5↑
	+ Rewards	40.50↓	48.00↓	52.00↓	66.0↓
	+ Pol. + Ger.	46.00↑	54.00↓	59.00↓	86.5↑
	+ Met. + Rew.	41.00↓	55.50↑	54.50↓	88.5↑
	+ Pol. + Ger. + Met.	46.50↑	53.50↓	62.00 ↑	89.5↑
o3-mini	Zero-shot CoT	92.00	88.50	94.50	97.0
	+ Politeness	88.50↓	87.00↓	93.50↓	96.0↓
	+ Germane load	88.00↓	82.00↓	95.00 ↑	96.5↓
	+ Metacognition	90.00↓	85.00↓	94.00↓	95.5↓
	+ Rewards	89.50↓	85.50↓	94.50	96.0↓
	+ Pol. + Ger.	81.00↓	71.00↓	88.50↓	97.0

Table 2: 模型（% ）在不同配置下各种任务上的性能。箭头表示相对于零样本 CoT 的变化。

CoT 提示 (Kojima et al., 2022) “逐步回答以下问题。” 开始。然后我们引入以下修改：(1) 添加 “请” 以促进 礼貌 ；(2) “在解决问题之前，反思你已有的知识，以更深入地理解问题。” 以鼓励 贴切负荷 ；(3) “彻底自我验证你的回答，以确保每个推理步骤都是正确的。” 以促进 元认知 ；(4) “每正确推理一步将奖励你 100 美元。” 以改善 奖励 。

研究结果。 我们在 Table 2 中的结果表明 不同的提示属性以不同的方式影响模型，其影响在不同任务中有所不同 。总体而言，大多数属性组合对 Llama-3.1 有利，但对其他模型产生负面影响。此外，我们观察到，结合多个正面属性并不一定会产生更好的改进; 相反，单一属性通常更为有效。具体而言，在 Comm.QA 和 ARC-C 数据集上，礼貌属性为 Llama 带来了最佳效果，而在所有任务中，元认知为 Qwen 取得了最高性能。关于属性的结合，尽管礼貌和重要负荷单独都能提升 Llama 在 MMLU 和 ARC-C 上的表现，二者结合反而导致性能低于仅使用礼貌时的表现。在 CommQA 数据集中，类似模式在将元认知与奖励结合用于 Llama 时也被观察到。令人惊讶的是，对于 o3-mini 模型，我们观察到大多数属性导致负面影响 。我们假设这可能是由于模型过度训练在链式思维数据上，导致属性将提示推向分布之外。最后，我们还注意到，在没有观察到任何改进的情况下，这并不意味着这些属性缺乏影响。相反，更复杂或优化的提示方法可能更好地促进这些属性，从而带来改进。我们将这些探索留给未来的研究。

6.2 属性增强微调

为了更好地理解模型特定因素，特别是指令微调，如何影响提示属性的有效性，我们在 Qwen-2.5-7B-It 模型上进行了一项有针对性的微调实验。我们选择这个模型是因为在使用更礼貌的提示时，它并没有表现出更好的推理能力。我们使用充满礼貌的数据或保持原始形式的数据，对 Qwen-2.5-7B-It 的两个变体进行微调。具体而言，我们从 Alpaca-GPT-4o 数据集 [6] 中采样 2,500 个例子，并创建两个微调集：一个在每个指令中添加 “请”，另一个则保持不变。

如 Table 3 所示，首先，在有礼貌提示上微调 Qwen-2.5-7B-It，在输入中添加 “请” 字时，性能有显著提升。这表明，将带有明确礼貌标识的数据用于指令微调可以增强模型对礼貌提示风格的敏感性，实现仅靠简单提示水平的礼貌无法达到的性能改进（§6.1 ）。其次，令人惊讶的是，与原始数据相比，使用礼貌增强的数据进行指令微调在几乎所有属性增强的实验中都取得了更好的结果。这表明，在指令微调过程中融入礼貌或更广泛地说，某些属性，可以导致更有效和更有鲁棒性的推理模型。

7 结论

本文通过一种新颖的基于属性的视角，探讨了自然语言提示及其对模型性能的影响。我们对超过 150 项提示研究进行了调查，并引入了一种 21 个关键属性的分类法，用于评估提示质量及其对模型性能的影响。我们的分析揭示了在不同模型和任务中对不同属性的不均衡重视，暴露了基于属性的提示优化中的显著研究空白。我们进一步识别了良好自然语言提示中的属性之间的相关性，从而提供了可操作的提示建议。在一个推理任务的案例研究中，我们发现增强单个提示属性往往比多属性组合表现更好，并且对这些属性进行微调可以改进推理，这挑战了组合属性总是能产生更好结果的假设。随着该领域的不断发展，我们希望这项工作能激励研究人员对提示属性与模型行为之间的关系进行更深入的研究，并推进提示评估方法及其在不同应用中的影响。

8

局限性

尽管我们尽最大努力进行严格和全面的研究，但我们承认我们的方法中存在几个局限性。

[6] https://huggingface.co/datasets/vicgalle/alpaca-gpt4

Method	MMLU	CQA	ARC	GSM8K	Avg.
Zero-shot CoT	60.0 / 67.00	67.5 / 69.00	73.5 / 68.50	85.00 / 85.00	71.50 / 72.38
+Politeness	69.5 ↑ / 62.50↓	72.5↑ / 70.00↑	85.0↑ / 79.50↑	85.00 / 88.50↑	78.00↑ / 75.13↑
+Germane load	49.0↓ / 45.00↓	47.5↓ / 43.00↓	49.0↓ / 51.00↓	84.00↓ / 88.00↑	57.38↓ / 56.80↓
+Metacognition	61.0↑ / 54.00↓	72.0↑ / 68.00↓	75.0↑ / 71.00↑	86.50↑ / 89.00↑	73.63↑ / 70.50↓
+Rewards	61.0↑ / 65.00↓	72.5↑ / 69.50↑	76.5↑ / 74.00↑	81.50↓ / 82.50↓	72.88↑ / 72.75↑
+Pol. + Ger.	49.5↓ / 51.50↓	62.5↓ / 63.00↓	70.0↓ / 67.50↓	85.00 / 78.00↓	66.75↓ / 65.00↓
+Met. + Rew.	54.5↓ / 57.00↓	69.5↓ / 68.00↓	68.0↓ / 67.50↓	85.00 / 85.50↑	69.25↓ / 69.50↓
+Pol. + Ger. + Met.	69.0↑ / 66.50↓	77.5 ↑ / 79.50 ↑	86.5 ↑ / 83.50 ↑	82.50↓ / 81.50↓	78.88 ↑ / 77.75 ↑

Table 3: 在不同设置下，两种微调的 Qwen-2.5-7B-it 模型（%）在礼貌数据/非礼貌数据上的表现。

首先，我们的研究受限于我们所调查文献的范围。由于人力资源的限制，我们无法覆盖该领域的所有相关论文。尽管我们努力通过调查来自不同会议和主题的多元出版物来减轻这一影响，但仍有可能遗漏了一些相关研究。这可能会影响我们研究结果的全面性，从而影响我们得出的结论。

其次，我们的相关性属性分析仅限于一组预定义的属性。尽管这些属性经过精心选择，以代表多样且有意义的维度，但分析其他属性可能会产生不同的结果。为了解决这个问题，我们确保收集的提示是多样化的，并通过人工审核进行验证。然而，属性选择的固有变化性给我们的研究结果的普遍性带来了潜在限制，在将这些结果推广到其他背景时应谨慎行事。

我们也同意某些维度，尤其是“责任”（包括“偏见”、“安全”、“隐私”、“可靠性”和“社会规范”），可能过于宽泛，涵盖了多种复杂问题。尽管更细致的细分可以提高分析精度，我们当前的方法主要是基于这样一个事实：先前研究中很少有探索这些维度的提示。如在 Table 1 中所示，这个维度仍然基本上未被充分探讨，大多数单元格为空。然而，随着更多研究的涌现，我们认识到进一步细化的重要性。随着该领域研究的进展和更细致的调查结果的出现，我们将相应地更新我们的研究以反映更细致的分类。

最后，我们的多属性提示增强实验是使用最简单形式的补充提示进行的，没有针对特定模型进行优化。虽然这种方法建立了基础分析，但可能导致在处理某些属性时效果不佳，并且忽视了这些属性对于单个模型的更精致提示的潜在优势。这一限制影响了我们研究结果的稳健性，并强调了未来研究提示优化技术的必要性。

总之，虽然我们采取了重要措施来减轻这些限制，但它们反映了在进行如此规模和复杂性的研究时固有的挑战。我们希望我们的工作能够成为该领域进一步探索和改进的基础。

9

伦理考虑 我们的分析可能被滥用于优化恶意目的的提示，例如生成错误信息、仇恨言论或侵犯隐私。虽然我们的研究并非旨在此类应用，但要完全防止潜在的滥用本质上是具有挑战的。尽管我们的研究可能会提高对抗性应用和恶意行为者的有效性，但我们并不认为其在恶意目的上会比在积极应用上更有优势。最后，我们以每小时 $ 20 的工资补偿我们的注释员，这高于当地的最低工资。

10

致谢 本研究得到了由新加坡国家研究基金会支持的人工智能新加坡计划（奖项编号：AISG2-GC-2022-005，AISG 奖项编号：AISG2-TC2023-010-SGIL）以及新加坡教育部学术研究基金一级（奖项编号：T1 251RES2207）的支持。DXL 得到了 A*STAR 计算与信息科学（ACIS）奖学金的支持。我们感谢国立大学 WING 和深度学习实验室的成员以及 ACL RR 匿名审稿人所提供的建设性反馈。

References

Anirudh Ajith, Chris Pan, Mengzhou Xia, Ameet Deshpande, and Karthik Narasimhan. 2023. Instructeval: Systematic evaluation of instruction selection methods. In *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*.

Afra Feyza Akyürek, Sejin Paik, Muhammed Kocyigit, Seda Akbiyik, Serife Leman Runyun, and Derry Wijaya. 2022. On measuring social biases in prompt-based multi-task learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 551–564, Seattle, United States. Association for Computational Linguistics.

Sarah Alnegheimish, Alicia Guo, and Yi Sun. 2022. Using natural sentence prompts for understanding biases in language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2824–2830, Seattle,

United States. Association for Computational Linguistics.

Reinald Kim Amplayo, Kellie Webster, Michael Collins, Dipanjan Das, and Shashi Narayan. 2023. Query refinement prompts for closed-book long-form QA. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7997–8012, Toronto, Canada. Association for Computational Linguistics.

Anthropic. 2024. Be clear, direct, and detailed. Accessed: 2025-01-15.

Siddhant Arora, Hayato Futami, Jee-weon Jung, Yifan Peng, Roshan Sharma, Yosuke Kashiwagi, Emiru Tsunoo, Karen Livescu, and Shinji Watanabe. 2024. UniverSLU: Universal spoken language understanding for diverse tasks with natural language instructions. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2754–2774, Mexico City, Mexico. Association for Computational Linguistics.

Simran Arora, Avanika Narayan, Mayee F Chen, Laurel Orr, Neel Guha, Kush Bhatia, Ines Chami, and Christopher Re. 2023. Ask me anything: A simple strategy for prompting language models. In *The Eleventh International Conference on Learning Representations*.

Liam Barkley and Brink van der Merwe. 2024. Investigating the role of prompting and external tools in hallucination rates of large language models. *arXiv preprint arXiv:2410.19385*.

Neeladri Bhuiya, Viktor Schlegel, and Stefan Winkler. 2024. Seemingly plausible distractors in multi-hop reasoning: Are large language models attentive readers? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2514–2528, Miami, Florida, USA. Association for Computational Linguistics.

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2023. Prompting language models for linguistic structure. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6649–6663, Toronto, Canada. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sondos Mahmoud Bsharat, Aidar Myrzakhan, and Zhiqiang Shen. 2023. Principled instructions are all you need for questioning llama-1/2, gpt-3.5/4. *arXiv preprint arXiv:2312.16171*.

Kyubyung Chae, Jaepill Choi, Yohan Jo, and Taesup Kim. 2024. Mitigating hallucination in abstractive summarization with domain-conditional mutual information. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1809–1820, Mexico City, Mexico. Association for Computational Linguistics.

Edward Y Chang. 2023. Prompting large language models with the socratic method. In *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0351–0360. IEEE.

Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. 2023a. Plot: Prompt learning with optimal transport for vision-language models. In *International Conference on Learning Representations (ICLR)*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Wei-Lin Chen, Cheng-Kuang Wu, Yun-Nung Chen, and Hsin-Hsi Chen. 2023b. Self-ICL: Zero-shot in-context learning with self-generated demonstrations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15651–15662, Singapore. Association for Computational Linguistics.

Yongchao Chen, Jacob Arkin, Yilun Hao, Yang Zhang, Nicholas Roy, and Chuchu Fan. 2024. PRompt optimization in multi-step tasks (PROMST): Integrating human feedback and heuristic-based sampling. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3859–3920, Miami, Florida, USA. Association for Computational Linguistics.

Yulin Chen, Ning Ding, Xiaobin Wang, Shengding Hu, Haitao Zheng, Zhiyuan Liu, and Pengjun Xie. 2023c. Exploring lottery prompts for pre-trained language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15428–15444, Toronto, Canada. Association for Computational Linguistics.

Jiale Cheng, Xiao Liu, Kehan Zheng, Pei Ke, Hongning Wang, Yuxiao Dong, Jie Tang, and Minlie Huang. 2024a. Black-box prompt optimization: Aligning large language models without model training. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3201–3219, Bangkok, Thailand. Association for Computational Linguistics.

Kewei Cheng, Nesreen K. Ahmed, Theodore L. Willke, and Yizhou Sun. 2024b. Structure guided prompt: Instructing large language model in multi-step reasoning by exploring graph structure of the text. In

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 9407–9430, Miami, Florida, USA. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113.

Yu-Neng Chuang, Tianwei Xing, Chia-Yuan Chang, Zirui Liu, Xun Chen, and Xia Hu. 2024. Learning to compress prompt in natural language formats. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7756–7767, Mexico City, Mexico. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Debrup Das, Debopriyo Banerjee, Somak Aditya, and Ashish Kulkarni. 2024. MATHSENSEI: A tool-augmented large language model for mathematical reasoning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 942–966, Mexico City, Mexico. Association for Computational Linguistics.

Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse. In *proceedings of Sinn und Bedeutung*, volume 23, pages 107–124.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric Xing, and Zhiting Hu. 2022. RLPrompt: Optimizing discrete text prompts with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3369–3391, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10602–10621, Singapore. Association for Computational Linguistics.

Dario Di Palma, Giovanni Maria Biancofiore, Vito Walter Anelli, Fedelucio Narducci, Tommaso Di Noia, and Eugenio Di Sciascio. 2023. Evaluating chatgpt as a recommender system: A rigorous approach. *arXiv preprint arXiv:2309.03613*.

Shizhe Diao, Pengcheng Wang, Yong Lin, Rui Pan, Xiang Liu, and Tong Zhang. 2024. Active prompting with chain-of-thought for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1330–1350. Association for Computational Linguistics.

Xuan Long Do, Kenji Kawaguchi, Min-Yen Kan, and Nancy Chen. 2025. Aligning large language models with human opinions through persona selection and value–belief–norm reasoning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2526–2547, Abu Dhabi, UAE. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Ehsan Doostmohammadi, Oskar Holmström, and Marco Kuhlmann. 2024. How reliable are automatic evaluation methods for instruction-tuned LLMs? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6321–6336, Miami, Florida, USA. Association for Computational Linguistics.

Vishnu Sashank Dorbala, Sanjoy Chowdhury, and Dinesh Manocha. 2024. Can LLM's generate human-like wayfinding instructions? towards platform-agnostic embodied instruction synthesis. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 258–271, Mexico City, Mexico. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Satyam Dwivedi, Sanjukta Ghosh, and Shivam Dwivedi. 2023. Breaking the bias: Gender fairness in llms using prompt engineering and in-context learning. *Rupkatha Journal on Interdisciplinary Studies in Humanities*, 15(4).

Jessica Maria Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. 2024. Cognitive bias in decision-making with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12640–12653, Miami, Florida, USA. Association for Computational Linguistics.

Kennedy Edemacu and Xintao Wu. 2024. Privacy preserving prompt engineering: A survey. *arXiv preprint arXiv:2404.06001*.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Amila Ferron, Amber Shore, Ekata Mitra, and Ameeta Agrawal. 2023. MEEP: Is this engaging? prompting large language models for dialogue evaluation in multilingual settings. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2078–2100, Singapore. Association for Computational Linguistics.

R.M. Gagné. 1985. *The Conditions of Learning and Theory of Instruction*. Holt, Rinehart and Winston.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Zhibin Gou, Qingyan Guo, and Yujiu Yang. 2023. MvP: Multi-view prompting improves aspect sentiment tuple prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4380–4397, Toronto, Canada. Association for Computational Linguistics.

Herbert Paul Grice. 1975. Logic and conversation. *Syntax and semantics*, 3:43–58.

Prompting Guide. 2024. Optimizing prompts. Accessed: 2024-12-22.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. 2024. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. In *The Twelfth International Conference on Learning Representations*.

Qianyu He, Jie Zeng, Qianxi He, Jiaqing Liang, and Yanghua Xiao. 2024. From complex to simple: Enhancing multi-constraint complex instruction following ability of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10864–10882, Miami, Florida, USA. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Wenyang Hu, Yao Shu, Zongmin Yu, Zhaoxuan Wu, Xiaoqiang Lin, Zhongxiang Dai, See-Kiong Ng, and Bryan Kian Hsiang Low. 2024. Localized zeroth-order prompt optimization. In *Advances in Neural Information Processing Systems*.

Jin Huang, Xingjian Zhang, Qiaozhu Mei, and Jiaqi Ma. 2024a. Can LLMs effectively leverage graph structural information through prompts, and why? *Transactions on Machine Learning Research*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2024b. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.* Just Accepted.

Xu Huang, Jianxun Lian, Yuxuan Lei, Jing Yao, Defu Lian, and Xing Xie. 2023a. Recommender ai agent: Integrating large language models for interactive recommendations. *arXiv preprint arXiv:2308.16505*.

Yongfeng Huang, Yanyang Li, Yicong Xu, Lin Zhang, Ruyi Gan, Jiaxing Zhang, and Liwei Wang. 2023b. Mvp-tuning: Multi-view knowledge retrieval with prompt tuning for commonsense reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*, pages 13417–13432. Association for Computational Linguistics.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, jiayi lei, Yao Fu, Maosong Sun, and Junxian He. 2023c. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

EunJeong Hwang, Yichao Zhou, James Bradley Wendt, Beliz Gunel, Nguyen Vo, Jing Xie, and Sandeep Tata. 2024. Enhancing incremental summarization with structured representations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3830–3842, Miami, Florida, USA. Association for Computational Linguistics.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, Singapore. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023a. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023b. LLMLingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13358–13376, Singapore. Association for Computational Linguistics.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2024. LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1658–1677, Bangkok, Thailand. Association for Computational Linguistics.

Zhiwei Jiang, Tianyi Gao, Yafeng Yin, Meng Liu, Hua Yu, Zifeng Cheng, and Qing Gu. 2023c. Improving domain generalization for prompt-aware essay scoring via disentangled representation learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12456–12470, Toronto, Canada. Association for Computational Linguistics.

Hoyoun Jung and Kyung-Joong Kim. 2024. Discrete prompt compression with reinforcement learning. *IEEE Access*.

Zhigang Kan, Linbo Qiao, Hao Yu, Liwen Peng, Yifu Gao, and Dongsheng Li. 2023. Protecting user privacy in remote conversational systems: A privacy-preserving framework based on text sanitization. *arXiv preprint arXiv:2306.08223*.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, and Xiaoyan Bai. 2023. PromptRank: Unsupervised keyphrase extraction using prompt. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9788–9801, Toronto, Canada. Association for Computational Linguistics.

Xiaojun Kuang, C. L. Philip Chen, Shuzhen Li, and Tong Zhang. 2024. Multi-scale prompt memory-augmented model for black-box scenarios. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1743–1757. Association for Computational Linguistics.

Joshua Lee, Wyatt Fong, Alexander Le, Sur Shah, Kevin Han, and Kevin Zhu. 2025. Pragmatic metacognitive prompting improves LLM performance on sarcasm detection. In *Proceedings of the 1st Workshop on Computational Humor (CHum)*, pages 63–70, Online. Association for Computational Linguistics.

Itay Levy, Ben Bogin, and Jonathan Berant. 2023. Diverse demonstrations improve in-context compositional generalization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1401–1422, Toronto, Canada. Association for Computational Linguistics.

Chengzhengxu Li, Xiaoming Liu, Zhaohan Zhang, Yichen Wang, Chen Liu, Yu Lan, and Chao Shen. 2024a. Concentrate attention: Towards domain-generalizable prompt optimization for language models. In *Advances in Neural Information Processing Systems*.

Junlong Li, Jinyuan Wang, Zhuosheng Zhang, and Hai Zhao. 2024b. Self-prompting large language models for zero-shot open-domain QA. In *Proceedings*

of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 296–310, Mexico City, Mexico. Association for Computational Linguistics.

Moxin Li, Wenjie Wang, Fuli Feng, Yixin Cao, Jizhi Zhang, and Tat-Seng Chua. 2023a. Robust prompt optimization for large language models against distribution shifts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1539–1554, Singapore. Association for Computational Linguistics.

Qian Li, Zhuo Chen, Cheng Ji, Shiqi Jiang, and Jianxin Li. 2024c. Llm-based multi-level knowledge generation for few-shot knowledge graph completion. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24.

Sha Li, Ruining Zhao, Manling Li, Heng Ji, Chris Callison-Burch, and Jiawei Han. 2023b. Open-domain hierarchical event schema induction by incremental prompting and verification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5677–5697, Toronto, Canada. Association for Computational Linguistics.

Shiyang Li, Jun Yan, Hai Wang, Zheng Tang, Xiang Ren, Vijay Srinivasan, and Hongxia Jin. 2024d. Instruction-following evaluation through verbalizer manipulation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3678–3692, Mexico City, Mexico. Association for Computational Linguistics.

Siqi Li, Danni Liu, and Jan Niehues. 2024e. Optimizing rare word accuracy in direct speech translation with a retrieval-and-demonstration approach. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12703–12719, Miami, Florida, USA. Association for Computational Linguistics.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023c. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Yingji Li, Mengnan Du, Xin Wang, and Ying Wang. 2023d. Prompt tuning pushes farther, contrastive learning pulls closer: A two-stage approach to mitigate social biases. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14254–14267, Toronto, Canada. Association for Computational Linguistics.

Yiwei Li, Jiayi Shi, Shaoxiong Feng, Peiwen Yuan, Xinglin Wang, Boyuan Pan, Heda Wang, Yao Hu, and Kan Li. 2024f. Instruction embedding: Latent representations of instructions towards task identification. In *Advances in Neural Information Processing Systems*.

Yucheng Li, Bo Dong, Frank Guerin, and Chenghua Lin. 2023e. Compressing context to enhance inference efficiency of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6342–6353, Singapore. Association for Computational Linguistics.

Yuxin Liang, Zhuoyang Song, Hao Wang, and Jiaxing Zhang. 2024. Learning to trust your feelings: Leveraging self-awareness in LLMs for hallucination mitigation. In *Proceedings of the 3rd Workshop on Knowledge Augmented Methods for NLP*, pages 44–58, Bangkok, Thailand. Association for Computational Linguistics.

Zujie Liang, Feng Wei, Yin Jie, Yuxi Qian, Zhenghong Hao, and Bing Han. 2023. Prompts can play lottery tickets well: Achieving lifelong information extraction via lottery prompt tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 277–292, Toronto, Canada. Association for Computational Linguistics.

Xiaoqiang Lin, Zhaoxuan Wu, Zhongxiang Dai, Wenyang Hu, Yao Shu, See-Kiong Ng, Patrick Jaillet, and Bryan Kian Hsiang Low. 2024. Use your INSTINCT: INSTruction optimization using neural bandits coupled with transformers.

Zhicheng Lin. 2024. How to write effective prompts for large language models. *Nature human behaviour*, 8(4):611–615.

Barys Liskavets, Maxim Ushakov, Shuvendu Roy, Mark Klibanov, Ali Etemad, and Shane Luke. 2024. Prompt compression with context-aware sentence encoding for fast and improved llm inference. *arXiv preprint arXiv:2409.01227*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024a. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Tongxuan Liu, Wenjiang Xu, Weizhe Huang, Xingyu Wang, Jiaxing Wang, Hailong Yang, and Jing Li. 2024b. Logic-of-thought: Injecting logic into contexts for full reasoning in large language models. *arXiv preprint arXiv:2409.17539*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yixin Liu, Alexander Fabbri, Jiawen Chen, Yilun Zhao, Simeng Han, Shafiq Joty, Pengfei Liu, Dragomir Radev, Chien-Sheng Wu, and Arman Cohan. 2024c.

Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4481–4501, Mexico City, Mexico. Association for Computational Linguistics.

Jia Li, Ge Li, Yongmin Li, and Zhi Jin. 2023. Structured chain-of-thought prompting for code generation. *ACM Transactions on Software Engineering and Methodology*.

Do Long, Yiran Zhao, Hannah Brown, Yuxi Xie, James Zhao, Nancy Chen, Kenji Kawaguchi, Michael Shieh, and Junxian He. 2024a. Prompt optimization via adversarial in-context learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7308–7327, Bangkok, Thailand. Association for Computational Linguistics.

Do Xuan Long, Ngoc-Hai Nguyen, Tiviatis Sim, Hieu Dao, Shafiq Joty, Kenji Kawaguchi, Nancy F. Chen, and Min-Yen Kan. 2025a. LLMs are biased towards output formats! systematically evaluating and mitigating output format bias of LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 299–330, Albuquerque, New Mexico. Association for Computational Linguistics.

Do Xuan Long, Duong Ngoc Yen, Anh Tuan Luu, Kenji Kawaguchi, Min-Yen Kan, and Nancy F. Chen. 2024b. Multi-expert prompting improves reliability, safety and usefulness of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20370–20401, Miami, Florida, USA. Association for Computational Linguistics.

Do Xuan Long, Duong Ngoc Yen, Do Xuan Trong, Luu Anh Tuan, Kenji Kawaguchi, Shafiq Joty, Min-Yen Kan, and Nancy F Chen. 2025b. Beyond in-context learning: Aligning long-form generation of large language models via task-inherent attribute guidelines. *arXiv preprint arXiv:2506.01265*.

Renze Lou, Kai Zhang, Jian Xie, Yuxuan Sun, Janice Ahn, Hanzi Xu, Yu Su, and Wenpeng Yin. 2024. Muffin: Curating multi-faceted instructions for improving instruction-following. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Sheng Lu, Hendrik Schuff, and Iryna Gurevych. 2024. How are prompts different in terms of sensitivity? *arXiv preprint arXiv:2311.07230*.

Yihan Ma, Xinyue Shen, Yixin Wu, Boyang Zhang, Michael Backes, and Yang Zhang. 2024. The death and life of great prompts: Analyzing the evolution of LLM prompts from the structural perspective. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21990–22001, Miami, Florida, USA. Association for Computational Linguistics.

Aman Madaan, Katherine Hermann, and Amir Yazdanbakhsh. 2023. What makes chain-of-thought prompting effective? a counterfactual study. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1448–1535, Singapore. Association for Computational Linguistics.

Junyu Mao, Stuart E. Middleton, and Mahesan Niranjan. 2024. Do prompt positions really matter? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4102–4130, Mexico City, Mexico. Association for Computational Linguistics.

Hugo Mercier and Dan Sperber. 2011. Why do humans reason? arguments for an argumentative theory. *Behavioral and Brain Sciences*, 34(2):57–74.

Grégoire Mialon, Roberto Dessi, Maria Lomeli, Christoforos Nalmpantis, Ramakanth Pasunuru, Roberta Raileanu, Baptiste Roziere, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. 2023. Augmented language models: a survey. *Transactions on Machine Learning Research*.

James Michaelov, Catherine Arnett, Tyler Chang, and Ben Bergen. 2023. Structural priming demonstrates abstract grammatical representations in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3703–3720, Singapore. Association for Computational Linguistics.

Kshitij Mishra, Manisha Burja, and Asif Ekbal. 2024. ABLE: Personalized disability support with politeness and empathy integration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22445–22470, Miami, Florida, USA. Association for Computational Linguistics.

Kshitij Mishra, Priyanshu Priya, and Asif Ekbal. 2023. PAL to lend a helping hand: Towards building an emotion adaptive polite and empathetic counseling conversational agent. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12254–12271, Toronto, Canada. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

Nikita Moghe, Patrick Xia, Jacob Andreas, Jason Eisner, Benjamin Van Durme, and Harsh Jhamtani. 2024. Interpreting user requests in the context of natural language standing instructions. In *Findings of the Association for Computational Linguistics:*

NAACL 2024, pages 4043–4060, Mexico City, Mexico. Association for Computational Linguistics.

Fangwen Mu, Lin Shi, Song Wang, Zhuohao Yu, Binquan Zhang, ChenXue Wang, Shichao Liu, and Qing Wang. 2024. Clarifygpt: A framework for enhancing llm-based code generation via requirements clarification. *Proc. ACM Softw. Eng.*, 1(FSE).

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Cauglar Gulccehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Hoang Nguyen, Ye Liu, Chenwei Zhang, Tao Zhang, and Philip Yu. 2023. CoF-CoT: Enhancing large language models with coarse-to-fine chain-of-thought prompting for multi-domain NLU tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12109–12119, Singapore. Association for Computational Linguistics.

Xuan-Phi Nguyen, Mahani Aljunied, Shafiq Joty, and Lidong Bing. 2024. Democratizing LLMs for low-resource languages by leveraging their English dominant abilities with linguistically-diverse prompts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3501–3516, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI. 2022. Introducing chatgpt.

OpenAI. 2023. Gpt-4 is openai' s most advanced system, producing safer and more useful responses.

OpenAI. 2024a. Introducing gpt-4o and more tools to chatgpt free users. Accessed: 2025-02-02.

OpenAI. 2024b. Prompt engineering guide. `https://platform.openai.com/docs/guides/prompt-engineering`. Accessed: 2024-12-17.

OpenAI. 2025. Openai o3-mini. Accessed: 2025-02-13.

Krista Opsahl-Ong, Michael J Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. 2024. Optimizing instructions and demonstrations for multi-stage language model programs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9340–9366, Miami, Florida, USA. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *Advances in neural information processing systems*, volume 35, pages 27730–27744.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Dongmei Zhang. 2024. LLMLingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 963–981, Bangkok, Thailand. Association for Computational Linguistics.

Judea Pearl. 1998. Graphical models for probabilistic and causal reasoning. *Quantified representation of uncertainty and imprecision*, pages 367–389.

Keqin Peng, Liang Ding, Yancheng Yuan, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2024a. Revisiting demonstration selection strategies in in-context learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9090–9101, Bangkok, Thailand. Association for Computational Linguistics.

Letian Peng, Yuwei Zhang, Zilong Wang, Jayanth Srinivasa, Gaowen Liu, Zihan Wang, and Jingbo Shang. 2024b. Answer is all you need: Instruction-following text embedding via answering the question. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 459–477, Bangkok, Thailand. Association for Computational Linguistics.

Quang Hieu Pham, Hoang Ngo, Anh Tuan Luu, and Dat Quoc Nguyen. 2024. Who's who: Large language models meet knowledge conflicts in practice. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10142–10151, Miami, Florida, USA. Association for Computational Linguistics.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, Singapore. Association for Computational Linguistics.

Reid Pryzant, Dan Iter, Jerry Li, Yin Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with "gradient descent" and beam search. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7957–7968, Singapore. Association for Computational Linguistics.

Valentina Pyatkin, Jena D. Hwang, Vivek Srikumar, Ximing Lu, Liwei Jiang, Yejin Choi, and Chandra Bhagavatula. 2023. ClarifyDelphi: Reinforced clarification questions with defeasibility rewards for social and moral situations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11253–11271, Toronto, Canada. Association for Computational Linguistics.

Chengwei Qin, Aston Zhang, Chen Chen, Anirudh Dagar, and Wenming Ye. 2024. In-context learning with iterative demonstration selection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7441–7455, Miami, Florida, USA. Association for Computational Linguistics.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2695–2709, Singapore. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

Gregory Schraw and David Moshman. 1995. Metacognitive theories. *Educational psychology review*, 7:351–371.

Shivam Shandilya, Menglin Xia, Supriyo Ghosh, Huiqiang Jiang, Jue Zhang, Qianhui Wu, and Victor Rühle. 2024. Taco-rl: Task aware prompt compression optimization with reinforcement learning. *arXiv preprint arXiv:2409.13035*.

ShareGPT. 2023. ShareGPT: Share your wildest ChatGPT conversations with one click. `https://sharegpt.com/`.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023a. HuggingGPT: Solving AI tasks with chatGPT and its friends in hugging face. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yongliang Shen, Zeqi Tan, Shuhui Wu, Wenqi Zhang, Rongsheng Zhang, Yadong Xi, Weiming Lu, and Yueting Zhuang. 2023b. PromptNER: Prompt locating and typing for named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12492–12507, Toronto, Canada. Association for Computational Linguistics.

Chengshuai Shi, Kun Yang, Zihan Chen, Jundong Li, Jing Yang, and Cong Shen. 2024. Efficient prompt optimization through the lens of best arm identification. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. 2022. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 14274–14289.

Chenglei Si, Dan Friedman, Nitish Joshi, Shi Feng, Danqi Chen, and He He. 2023a. Measuring inductive biases of in-context learning with underspecified demonstrations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11289–11310, Toronto, Canada. Association for Computational Linguistics.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Lee Boyd-Graber, and Lijuan Wang. 2023b. Prompting GPT-3 to be reliable. In *The Eleventh International Conference on Learning Representations*.

Somanshu Singla, Zhen Wang, Tianyang Liu, Abdullah Ashfaq, Zhiting Hu, and Eric P. Xing. 2024. Dynamic rewarding with prompt optimization enables tuning-free self-alignment of language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21889–21909, Miami, Florida, USA. Association for Computational Linguistics.

Ritwik Sinha, Zhao Song, and Tianyi Zhou. 2023. A mathematical abstraction for balancing the trade-off between creativity and reality in large language models. *arXiv preprint arXiv:2306.02295*.

Dilara Soylu, Christopher Potts, and Omar Khattab. 2024. Fine-tuning and prompt optimization: Two great steps that work better together. *Stanford Institute for Human-Centered Artificial Intelligence (HAI)*.

Sam Spilsbury, Pekka Marttinen, and Alexander Ilin. 2024. Generating demonstrations for in-context compositional generalization in grounded language learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15960–15991, Miami, Florida, USA. Association for Computational Linguistics.

Yusheng Su, Xiaozhi Wang, Yujia Qin, Chi-Min Chan, Yankai Lin, Huadong Wang, Kaiyue Wen, Zhiyuan Liu, Peng Li, Juanzi Li, Lei Hou, Maosong Sun, and Jie Zhou. 2021. On transferability of prompt tuning for natural language processing. *arXiv preprint arXiv:2111.06719*.

Theodore Sumers, Robert Hawkins, Mark K Ho, Tom Griffiths, and Dylan Hadfield-Menell. 2022. How to talk so ai will learn: Instructions, descriptions, and autonomy. In *Advances in neural information processing systems*, volume 35, pages 34762–34775.

Weiwei Sun, Hengyi Cai, Hongshen Chen, Pengjie Ren, Zhumin Chen, Maarten de Rijke, and

Zhaochun Ren. 2023. Answering ambiguous questions via iterative prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7669–7683, Toronto, Canada. Association for Computational Linguistics.

Yueqing Sun, Yu Zhang, Le Qi, and Qi Shi. 2022. TSGP: Two-stage generative prompting for unsupervised commonsense question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 968–980, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mirac Suzgun and Adam Tauman Kalai. 2024. Meta-prompting: Enhancing language models with task-agnostic scaffolding. *arXiv preprint arXiv:2401.12954*.

John Sweller and Paul Chandler. 1991. Evidence for cognitive load theory. *Cognition and Instruction*, 8(4):351–362.

Chang-Yu Tai, Ziru Chen, Tianshu Zhang, Xiang Deng, and Huan Sun. 2023. Exploring chain of thought style prompting for text-to-SQL. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5376–5393, Singapore. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Pittawat Taveekitworachai, Febri Abdullah, and Ruck Thawonmas. 2024. Null-shot prompting: Rethinking prompting large language models with hallucination. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13321–13361, Miami, Florida, USA. Association for Computational Linguistics.

Gemini Team, 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*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Yuan Tian, Nan Xu, and Wenji Mao. 2024. A theory guided scaffolding instruction framework for LLM-enabled metaphor reasoning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7738–7755, Mexico City, Mexico. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Xingchen Wan, Ruoxi Sun, Hootan Nakhost, and Sercan O Arik. 2024. Teach better or show smarter? on instructions and exemplars in automatic prompt optimization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023a. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2717–2739, Toronto, Canada. Association for Computational Linguistics.

Hongru Wang, Rui Wang, Fei Mi, Yang Deng, Zezhong Wang, Bin Liang, Ruifeng Xu, and Kam-Fai Wong. 2023b. Cue-CoT: Chain-of-thought prompting for responding to in-depth dialogue questions with

LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12047–12064, Singapore. Association for Computational Linguistics.

Jinyuan Wang, Junlong Li, and Hai Zhao. 2023c. Self-prompted chain-of-thought on large language models for open-domain multi-hop reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2717–2731, Singapore. Association for Computational Linguistics.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023d. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2609–2634. Association for Computational Linguistics.

Ming Wang, Yuanzhong Liu, Xiaoyu Liang, Songlian Li, Yijie Huang, Xiaoming Zhang, Sijia Shen, Chaofeng Guan, Daling Wang, Shi Feng, et al. 2024a. Langgpt: Rethinking structured reusable prompt design framework for llms from the programming language. *arXiv preprint arXiv:2402.16929*.

Peng Wang, Xiaobin Wang, Chao Lou, Shengyu Mao, Pengjun Xie, and Yong Jiang. 2024b. Effective demonstration annotation for in-context learning via language model-based determinantal point process. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1266–1280, Miami, Florida, USA. Association for Computational Linguistics.

Rui Wang, Fei Mi, Yi Chen, Boyang Xue, Hongru Wang, Qi Zhu, Kam-Fai Wong, and Ruifeng Xu. 2024c. Role prompting guided domain adaptation with general capability preserve for large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2243–2255, Mexico City, Mexico. Association for Computational Linguistics.

Rui Wang, Hongru Wang, Fei Mi, Boyang Xue, Yi Chen, Kam-Fai Wong, and Ruifeng Xu. 2024d. Enhancing large language models against inductive instructions with dual-critique prompting. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5345–5363, Mexico City, Mexico. Association for Computational Linguistics.

Sijia Wang, Mo Yu, and Lifu Huang. 2023e. The art of prompting: Event detection based on type specific prompts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1286–1299, Toronto, Canada. Association for Computational Linguistics.

Xinyuan Wang, Chenxi Li, Zhen Wang, Fan Bai, Haotian Luo, Jiayou Zhang, Nebojsa Jojic, Eric Xing, and Zhiting Hu. 2023f. Promptagent: Strategic planning with language models enables expert-level prompt optimization. In *The Twelfth International Conference on Learning Representations*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Yan Wang, Zhixuan Chu, Xin Ouyang, Simeng Wang, Hongyan Hao, Yue Shen, Jinjie Gu, Siqiao Xue, James Y Zhang, Qing Cui, et al. 2023g. Enhancing recommender systems with large language model reasoning graphs. *arXiv preprint arXiv:2308.10835*.

Yancheng Wang, Ziyan Jiang, Zheng Chen, Fan Yang, Yingxue Zhou, Eunah Cho, Xing Fan, Yanbin Lu, Xiaojiang Huang, and Yingzhen Yang. 2024e. RecMind: Large language model powered agent for recommendation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4351–4364, Mexico City, Mexico. Association for Computational Linguistics.

Yau-Shian Wang, Ta-Chung Chi, Ruohong Zhang, and Yiming Yang. 2023h. PESCO: Prompt-enhanced self contrastive learning for zero-shot text classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14897–14911, Toronto, Canada. Association for Computational Linguistics.

Yifan Wang, Yafei Liu, Chufan Shi, Haoling Li, Chen Chen, Haonan Lu, and Yujiu Yang. 2024f. InsCL: A data-efficient continual learning paradigm for fine-tuning large language models with instructions. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 663–677, Mexico City, Mexico. Association for Computational Linguistics.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2024g. Resolving knowledge conflicts in large language models. In *First Conference on Language Modeling*.

Yuqing Wang and Yun Zhao. 2024. Metacognitive prompting improves understanding in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1914–1926, Mexico City, Mexico. Association for Computational Linguistics.

Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng

Chau. 2023i. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 893–911, Toronto, Canada. Association for Computational Linguistics.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in neural information processing systems*, volume 35, pages 24824–24837.

Bosi Wen, Pei Ke, Xiaotao Gu, Lindong Wu, Hao Huang, Jinfeng Zhou, Wenchuang Li, Binxin Hu, Wendy Gao, Jiaxing Xu, Yiming Liu, Jie Tang, Hongning Wang, and Minlie Huang. 2024. Benchmarking complex instruction-following with multiple constraints composition. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Guojun Wu. 2023. ICU: Conquering language barriers in vision-and-language modeling by dividing the tasks into image captioning and language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14740–14746, Singapore. Association for Computational Linguistics.

Qinzhuo Wu, Wei Liu, Jian Luan, and Bin Wang. 2024a. ToolPlanner: A tool augmented LLM for multi granularity instructions with path planning and feedback. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18315–18339, Miami, Florida, USA. Association for Computational Linguistics.

Xuansheng Wu, Wenlin Yao, Jianshu Chen, Xiaoman Pan, Xiaoyang Wang, Ninghao Liu, and Dong Yu. 2024b. From language modeling to instruction following: Understanding the behavior shift in LLMs after instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2341–2369, Mexico City, Mexico. Association for Computational Linguistics.

Zhaoxuan Wu, Xiaoqiang Lin, Zhongxiang Dai, Wenyang Hu, Yao Shu, See-Kiong Ng, Patrick Jaillet, and Bryan Kian Hsiang Low. 2024c. Prompt optimization with EASE? efficient ordering-aware automated selection of exemplars. In *ICML 2024 Workshop on In-Context Learning*.

Zongyu Wu, Hongcheng Gao, Yueze Wang, Xiang Zhang, and Suhang Wang. 2024d. Universal prompt optimizer for safe text-to-image generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6340–6354, Mexico City, Mexico. Association for Computational Linguistics.

Zeguan Xiao, Yan Yang, Guanhua Chen, and Yun Chen. 2024. Distract large language models for automatic jailbreak attack. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16230–16244, Miami, Florida, USA. Association for Computational Linguistics.

Binfeng Xu, Xukun Liu, Hua Shen, Zeyu Han, Yuhan Li, Murong Yue, Zhiyuan Peng, Yuchen Liu, Ziyu Yao, and Dongkuan Xu. 2023a. Gentopia.AI: A collaborative platform for tool-augmented LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 237–245, Singapore. Association for Computational Linguistics.

Jialiang Xu, Shenglan Li, Zhaozhuo Xu, and Denghui Zhang. 2024. Do LLMs know to respect copyright notice? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20604–20619, Miami, Florida, USA. Association for Computational Linguistics.

Lei Xu, Yangyi Chen, Ganqu Cui, Hongcheng Gao, and Zhiyuan Liu. 2022. Exploring the universal vulnerability of prompt-based learning paradigm. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1799–1810, Seattle, United States. Association for Computational Linguistics.

Yuanjian Xu, Qi An, Jiahuan Zhang, Peng Li, and Zaiqing Nie. 2023b. Hard sample aware prompt-tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12356–12369, Toronto, Canada. Association for Computational Linguistics.

Jinghan Yang, Shuming Ma, and Furu Wei. 2023a. Auto-icl: In-context learning without human supervision. *arXiv preprint arXiv:2311.09263*.

Kexin Yang, Dayiheng Liu, Wenqiang Lei, Baosong Yang, Mingfeng Xue, Boxing Chen, and Jun Xie. 2023b. Tailor: A soft-prompt-based approach to attribute-based controlled text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 410–427, Toronto, Canada. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William 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*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Ziqi Yin, Hao Wang, Kaito Horio, Daisuike Kawahara, and Satoshi Sekine. 2024. Should we respect LLMs? a cross-lingual study on the influence of prompt politeness on LLM performance. In *Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024)*, pages 9–35, Miami, Florida, USA. Association for Computational Linguistics.

Siyu Yuan, Deqing Yang, Jinxi Liu, Shuyu Tian, Jiaqing Liang, Yanghua Xiao, and Rui Xie. 2023. Causality-aware concept extraction based on knowledge-guided prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9255–9272, Toronto, Canada. Association for Computational Linguistics.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. 2024a. Self-rewarding language models. In *Forty-first International Conference on Machine Learning*.

Ye Yuan, Kexin Tang, Jianhao Shen, Ming Zhang, and Chenguang Wang. 2024b. Measuring social norms of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 650–699, Mexico City, Mexico. Association for Computational Linguistics.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2024. Evaluating large language models at evaluating instruction following. In *International Conference on Learning Representations (ICLR)*.

Jingtao Zhan, Qingyao Ai, Yiqun Liu, Yingwei Pan, Ting Yao, Jiaxin Mao, Shaoping Ma, and Tao Mei. 2024. Prompt refinement with image pivot for text-to-image generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 941–954, Bangkok, Thailand. Association for Computational Linguistics.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024a. R-tuning: Instructing large language models to say 'I don't know'. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7113–7139, Mexico City, Mexico. Association for Computational Linguistics.

Qianchi Zhang, Hainan Zhang, Liang Pang, Hongwei Zheng, and Zhiming Zheng. 2024b. Adacomp: Extractive context compression with adaptive predictor for retrieval-augmented large language models. *arXiv preprint arXiv:2409.01579*.

Tianjun Zhang, Xuezhi Wang, Denny Zhou, Dale Schuurmans, and Joseph E Gonzalez. 2022. Tempera: Test-time prompt editing via reinforcement learning. In *The Eleventh International Conference on Learning Representations*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*.

Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Xuan Long Do, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, and Shafiq Joty. 2023. Retrieving multimodal information for augmented generation: A survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4736–4756, Singapore. Association for Computational Linguistics.

Chujie Zheng, Fan Yin, Hao Zhou, Fandong Meng, Jie Zhou, Kai-Wei Chang, Minlie Huang, and Nanyun Peng. 2024a. On prompt-driven safeguarding for large language models. In *International Conference on Machine Learning*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2024b. LMSYS-chat-1m: A large-scale real-world LLM conversation dataset. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023a. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023b. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

Pei Zhou, Jay Pujara, Xiang Ren, Xinyun Chen, Heng-Tze Cheng, Quoc V Le, Ed H Chi, Denny Zhou, Swaroop Mishra, and Huaixiu Steven Zheng. 2024a. Self-discover: Large language models self-compose reasoning structures. *arXiv preprint arXiv:2402.03620*.

Pei Zhou, Jay Pujara, Xiang Ren, Xinyun Chen, Heng-Tze Cheng, Quoc V Le, Ed H. Chi, Denny Zhou, Swaroop Mishra, and Steven Zheng. 2024b. SELF-DISCOVER: Large language models self-compose reasoning structures. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Xiaoling Zhou, Wei Ye, Yidong Wang, Chaoya Jiang, Zhemg Lee, Rui Xie, and Shikun Zhang. 2024c. Enhancing in-context learning via implicit demonstration augmentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2810–2828, Bangkok, Thailand. Association for Computational Linguistics.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023c. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*.

Yujia Zhou, Zheng Liu, Jiajie Jin, Jian-Yun Nie, and Zhicheng Dou. 2024d. Metacognitive retrieval-augmented large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 1453–1463.

Dongsheng Zhu, Daniel Tang, Weidong Han, Jinghui Lu, Yukun Zhao, Guoliang Xing, Junfeng Wang, and Dawei Yin. 2024. VisLingInstruct: Elevating zero-shot learning in multi-modal language models with autonomous instruction optimization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2122–2135, Mexico City, Mexico. Association for Computational Linguistics.

Rongxin Zhu, Jey Han Lau, and Jianzhong Qi. 2025. Factual dialogue summarization via learning from large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4474–4492, Abu Dhabi, UAE. Association for Computational Linguistics.

Sicheng Zhu, Ruiyi Zhang, Bang An, Gang Wu, Joe Barrow, Zichao Wang, Furong Huang, Ani Nenkova, and Tong Sun. 2023. Autodan: Automatic and interpretable adversarial attacks on large language models. *arXiv preprint arXiv:2310.15140*.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

关于我们的代码实现和本研究中使用的自定义提示的更全面概述，请参阅附加的补充材料。

A 补充结果

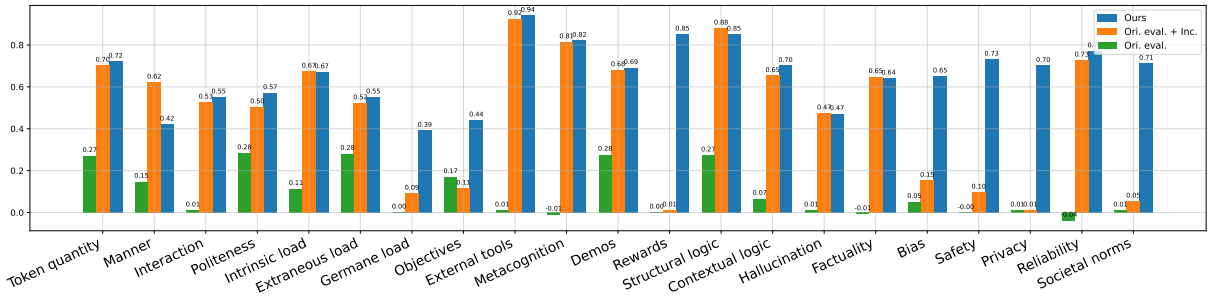

Figure 2: Agreements between human evaluators and LLM-based evaluation methods measured by Cohen's Kappa.

PE papers	ChatGPT PC	Awe. ChatGPT Prompts	Alpaca	NI	CI	Multi-turn	Total
25	66	44	108	462	60	204	969
Human	Human	Human	Machine	Human	Machine	Human	

Table 4: Prompt evaluation statistics

B 调研论文

Table 5: 具有自动索引增长的表格

Index	Category	Title	Conference and year	Best prompt means?
	PE	Structured Chain-of-Thought Prompting for Code Generation (Li et al., 2023)	ACM Transactions 2022	Highest Performance
	PE	TSGP: Two-Stage Generative Prompting for Unsupervised Commonsense ... (Sun et al., 2022)	EMNLP 2022	Prior Knowledge Engagement
	PE	Chain-of-Thought Prompting Elicits Reasoning in Large Language Models (Wei et al., 2022)	NeurIPS 2022	Highest Performance
	PE	Ask Me Anything: A Simple Strategy for Prompting Language Models (Arora et al., 2023)	ICLR 2023	Highest performance
	PE	Augmented Language Models: a Survey (Zhao et al., 2023)	Preprint 2023	Enhanced Task Decomposition
	PE	Large Language Models are Human-Level Prompt Engineers (Zhou et al., 2023c)	ICLR 2023	Highest Performance
	PE	Least-to-Most Prompting Enables Complex Reasoning ... (Zhou et al., 2023a)	ICLR 2023	Enhanced Task Decomposition
	PE	Decomposed Prompting: A Modular Approach for Solving Complex Tasks (Khot et al., 2023)	ICLR 2023	Enhanced Task Decomposition
	PE	Chain-of-Thought Prompting Elicits Reasoning in Large Language Models (Wei et al., 2022)	ICLR 2023	Highest Performance
	PE	Prompting GPT-3 to be Reliable (Si et al., 2023b)	ICLR 2023	Reliability Enhancement
	PE	Large Language Models Can Be Easily Distracted by Irrelevant Context (Shi et al., 2023)	ICML 2023	Contextual Relavance
	PE	Answering Ambiguous Questions via Iterative Prompting (Sun et al., 2023)	ACL 2023	Performance-Diversity Balance
	PE	Causality-aware Concept Extraction based on Knowledge-guided Prompting (Yuan et al., 2023)	ACL 2023	Bias Mitigation

PE	DIFFUSIONDB: A Large-scale ProWe agree that some dimensions, particularly "Responsibility" (including “Bias”, “Safety”, “Privacy”, “Reliability”, and “Societal norms”) may be too broad and encompass multiple complex issues. While a more fine-grained subdivision could enhance analytical precision, our current approach is mainly motivated by the fact that there is a lack of prior studies that explore prompting with these dimensions. As reflected in Table 1, this dimension remains largely underexplored, with most cells empty. However, we recognize the importance of further refinement as more studies emerge. As research in this area advances and more fine-grained investigations become available, we will update our study accordingly to reflect a more nuanced categorization.mpt Gallery Dataset for Text-to-Image ... (Wang et al., 2023i)	ACL 2023	Highest Performance
PE	Exploring Lottery Prompts for Pre-trained Language Models (Chen et al., 2023c)	ACL 2023	Highest performance
PE	Improving Domain Generalization for Prompt-Aware Essay Scoring via... (Jiang et al., 2023c)	ACL 2023	Domain Generalization Capability
PE	MVP: Multi-view Prompting Improves Aspect Sentiment Tuple Prediction (Gou et al., 2023)	ACL 2023	Diverse Outcomes
PE	Prompting Language Models for Linguistic Structure (Blevins et al., 2023)	ACL 2023	Highest performance
PE	PromptRank: Unsupervised Keyphrase Extraction Using Prompt (Kong et al., 2023)	ACL 2023	Highest Performance
PE	Prompting PaLM for Translation: Assessing Strategies and Performance (Vilar et al., 2023)	ACL 2023	Highest Performance
PE	PromptNER: Prompt Locating and Typing for Named Entity Recognition (Shen et al., 2023b)	ACL 2023	Highest Performance
PE	Open-Domain Hierarchical Event Schema Induction ... (Li et al., 2023b)	ACL 2023	Enhanced Task Decomposition
PE	Retrieving Multimodal Information for Augmented Generation: A Survey (Zhao et al., 2023)	ACL 2023	Multimodal Enhancement
PE	Towards Understanding Chain-of-Thought Prompting ... (Wang et al., 2023a)	ACL 2023	Coherence and Relevance
PE	The Art of Prompting: Event Detection based on Type Specific Prompts (Wang et al., 2023e)	ACL 2023	Highest performance
PE	Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought ... (Wang et al., 2023d)	ACL 2023	Highest Performance
PE	PESCO: Prompt-enhanced Self-Contrastive Learning for Zero-shot ... (Wang et al., 2023h)	ACL 2023	Highest Performance
PE	MEEP: Is this Engaging? Prompting Large Language Models for Dialogue Evaluation in Multilingual Settings (Ferron et al., 2023)	ACL 2023	Engagingness Evaluation
PE	PAL to Lend a Helping Hand: Towards Building an Emotion Adaptive Polite and Empathetic Counseling Conversational Agent (Mishra et al., 2023)	ACL 2023	Emotion-Aware Interaction
PE	Query Refinement Prompts for Closed-Book Long-Form QA (Amplayo et al., 2023)	ACL 2023	Enhanced Task Decomposition
PE	Tailor: A Soft-Prompt-Based Approach to Attribute-Based Controlled ... (Yang et al., 2023b)	ACL 2023	Highest Performance
PE	Prompting and Evaluating Large Language Models for Proactive Dialogues ... (Deng et al., 2023)	EMNLP 2023	Highest Performance
PE	Cross-lingual Prompting: Improving Zero-shot Chain-of-Thought Reasoning across Languages (Qin et al., 2023)	EMNLP 2023	Highest Performance

PE	CoF-CoT: Enhancing Large Language Models with Coarse-to-Fine Chain-of-Thought Prompting for Multi-domain NLU Tasks (Nguyen et al., 2023)	EMNLP 2023	Highest Perfomance
PE	Exploring Chain of Thought Style Prompting for Text-to-SQL (Tai et al., 2023)	EMNLP 2023	Effective Reasoning Support
PE	G-EVAL: NLG Evaluation using GPT-4 with Better Human Alignment (Liu et al., 2023)	EMNLP 2023	Highest Performance
PE	Gentopia.AI: A Collaborative Platform for Tool-Augmented LLMs (Xu et al., 2023a)	EMNLP 2023	Highest Perfomance
PE	Self-prompted Chain-of-Thought on Large Language Models for Open-domain Multi-hop Reasoning (Wang et al., 2023c)	EMNLP 2023	Highest Perfomance
PE	LLMLingua: Compressing Prompts for Accelerated Inference of Large Language Models (Jiang et al., 2023b)	EMNLP 2023	Performance-Preserving Semantic Compression
PE	Towards Mitigating LLM Hallucination via Self Reflection (Ji et al., 2023)	EMNLP 2023	Hallucination Mitigation
PE	ClarifyGPT: A Framework for Enhancing LLM-Based Code Generation via Requirements Clarification (Mu et al., 2024)	ACM 2023	Highest Performance
PE	Breaking the Bias: Gender Fairness in LLMs Using Prompt Engineering and In-Context Learning (Dwivedi et al., 2023)	Journal 2023	Bias Mitigation
PE	Enhancing Recommender Systems with Large Language Model Reasoning Graphs (Wang et al., 2023g)	Preprint 2023	Highest Performance
PE	Who' s Who: Large Language Models Meet Knowledge Conflicts in Practice (Pham et al., 2024)	EMNLP 2024	Conflict Resolution
PE	The Death and Life of Great Prompts: Analyzing the Evolution of LLM ... (Ma et al., 2024)	EMNLP 2024	Coherent Structure
PE	Enhancing Incremental Summarization with Structured Representations (Hwang et al., 2024)	EMNLP 2024	Effective Structured Representations
PE	A Survey on In-context Learning (Dong et al., 2024)	EMNLP 2024	Effective Demonstrations
PE	Distract Large Language Models for Automatic Jailbreak Attack (Xiao et al., 2024)	EMNLP 2024	High Attack Success Rate
PE	Multi-expert Prompting Improves Reliability, Safety and Usefulness of Large ... (Long et al., 2024b)	EMNLP 2024	Reliability and Usefulness Enhancement
PE	How are Prompts Different in Terms of Sensitivity? (Lu et al., 2024)	NAACL 2024	Highest Performance
PE	Role Prompting Guided Domain Adaptation with General Capability Preserve... (Wang et al., 2024c)	NAACL 2024	Effective Role Assignment
PE	Mitigating Hallucination in Abstractive Summarization with Domain-Conditional Mutual Information (Chae et al., 2024)	NAACL 2024	Hallucination Mitigation
PE	Metacognitive Prompting Improves Understanding in Large Language Models (Wang and Zhao, 2024)	NAACL 2024	Highest Performance
PE	Effective Demonstration Annotation for In-Context Learning via Language Model-Based Determinantal Point Process (Wang et al., 2024b)	EMNLP 2024	Highest Performance
PE	Self-Prompting Large Language Models for Zero-Shot Open-Domain QA (Li et al., 2024b)	NAACL 2024	Effective Contextualization
PE	Learning to Compress Prompt in Natural Language Formats, (Chuang et al., 2024)	NAACL 2024	Token efficiency
PE	Should We Respect LLMs? A Cross-Lingual Study on the Influence of ... (Yin et al., 2024)	SICon 2024	Prompt Politeness
PE	Resolving Knowledge Conflicts in Large Language Models (Wang et al., 2024g)	COLM 2024	Conflict Resolution

PE	A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented ... (Fan et al., 2024)	KDD 2024	Effective Knowledge Integration
PE	Can LLMs Effectively Leverage Graph Structural Information ... (Huang et al., 2024a)	TMLR 2024	Coherent Structure
PE	A Survey on Hallucination in Large Language Models: Principles, ... (Huang et al., 2024b)	ACM 2024	Hallucination Mitigation
PE	Democratizing LLMs for Low-Resource Languages by Leveraging their English Dominant Abilities with Linguistically-Diverse Prompts (Nguyen et al., 2024)	ACL 2024	Effective Exemplars
PE	Active Prompting with Chain-of-Thought for Large Language Models (Diao et al., 2024)	ACL 2024	Enhanced Task Decomposition
PE	Prompt Refinement with Image Pivot for Text-to-Image Generation (Zhan et al., 2024)	ACL 2024	Highest Performance
PE	Learning to Trust Your Feelings: Leveraging Self-awareness in LLMs for ... (Liang et al., 2024)	KnowledgeNLP 2024	Hallucination Mitigation
PE	Should We Respect LLMs? A Cross-Lingual Study ... (Yin et al., 2024)	SICon 2024	Optimal Politeness Level
PE	LLM-based Multi-Level Knowledge Generation for Few-shot Knowledge Graph Completion (Li et al., 2024c)	IJCAI 2024	Knowledge Integrity
PE	AdaComp: Extractive Context Compression with Adaptive Predictor ... (Zhang et al., 2024b)	Preprint 2024	Relevance and Efficiency
PE	LangGPT: Rethinking Structured Reusable Prompt Design Framework for LLMs from the Programming Language (Wang et al., 2024a)	Preprint 2024	Reusable Prompts
PE	TACO-RL: Task Aware Prompt Compression Optimization with Reinforcement Learning (Shandilya et al., 2024)	Preprint 2024	Highest Performance
PE	LangGPT: Rethinking Structured Reusable Prompt Design Framework ... (Wang et al., 2024a)	Preprint 2024	Coherent Structure
PE	Meta-Prompting: Enhancing Language Models with Task-Agnostic ... (Suzgun and Kalai, 2024)	Preprint 2024	Task-Agnostic Scaffolding
PE	Investigating the Role of Prompting and External Tools ... (Barkley and van der Merwe, 2024)	Preprint 2024	Hallucination Mitigation
PE	Principled Instructions Are All You Need for Questioning LLaMA-1/2 ... (Bsharat et al., 2023)	Preprint 2024	Designed Principles Guidance
PE	Privacy Preserving Prompt Engineering: A Survey (Edemacu and Wu, 2024)	Preprint 2024	Privacy Risks Mitigation
PE	Aligning Large Language Models with Human Opinions through Persona Selection and Value–Belief–Norm Reasoning (Do et al., 2025)	COLING 2025	Effective Persona Utilization
PO	Do Prompt-Based Models Really Understand the Meaning ... (Webson and Pavlick, 2022)	NAACL 2022	Highest Performance
PO	Exploring the Universal Vulnerability of Prompt-based Learning Paradigm (Xu et al., 2022)	NAACL 2022	Highest Performance
PO	Using Natural Sentences for Understanding Biases in ... (Alnegheimish et al., 2022)	NAACL 2022	Bias Mitigation
PO	On Measuring Social Biases in Prompt-Based Multi-Task Learning (Akyürek et al., 2022)	NAACL 2022	Bias Mitigation
PO	On Transferability of Prompt Tuning for Natural Language Processing (Su et al., 2021)	NAACL 2022	Domain Generalization Capability
PO	Test-Time Prompt Tuning for Zero-Shot Generalization in Vision-Language ... (Shu et al., 2022)	NeurIPS 2022	Consistent Performance
PO	PLOT: Prompt Learning with Optimal Transport for Vision-Language ... (Chen et al., 2023a)	NeurIPS 2022	Domain Generalization Capability

PO	ASK ME ANYTHING: A SIMPLE STRATEGY FOR PROMPTING ... (Arora et al., 2023)	ICLR 2023	Highest Performance
PO	TEMPERA: Test-Time Prompt Editing via Reinforcement Learning (Zhang et al., 2022)	ICLR 2023	Highest Performance
PO	Automatic Prompt Optimization with "Gradient Descent" and Beam Search (Pryzant et al., 2023)	EMNLP 2023	Highest Performance
PO	Compressing Context to Enhance Inference Efficiency of Large Language Models (Li et al., 2023e)	EMNLP 2023	Efficiency and Performance
PO	Robust Prompt Optimization for Large Language Models Against ... (Li et al., 2023a)	EMNLP 2023	Domain Generalization Capability
PO	Hard Sample Aware Prompt-Tuning (Xu et al., 2023b)	ACL 2023	Effective Sample Utilization
PO	MVP-Tuning: Multi-View Knowledge Retrieval with Prompt Tuning for ... (Huang et al., 2023b)	ACL 2023	Highest Performance
PO	Prompt Tuning Pushes Farther, Contrastive Learning Pulls Closer ... (Li et al., 2023d)	ACL 2023	Effective Representation
PO	Prompts Can Play Lottery Tickets Well ... (Liang et al., 2023)	ACL 2023	Domain Generalization Capability
PO	Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters (Wang et al., 2023a)	ACL 2023	Coherence and Relevance
PO	Large Language Models Can Be Easily Distracted by Irrelevant Context (Shi et al., 2023)	ICML 2023	Relevance Maintenance
PO	Discrete Prompt Compression with Reinforcement Learning (Jung and Kim, 2024)	Preprint 2023	Highest Performance
PO	VisLingInstruct: Elevating Zero-Shot Learning in Multi-Modal Language ... (Zhu et al., 2024)	Preprint 2024	Highest Performance
PO	Concentrate Attention: Towards Domain-Generalizable Prompt Optimization ... (Li et al., 2024a)	NeurIPS 2024	Domain Generalization Capability
PO	Efficient Prompt Optimization Through the Lens of Best Arm Identification (Shi et al., 2024)	NeurIPS 2024	Highest Performance
PO	Localized Zeroth-Order Prompt Optimization (Hu et al., 2024)	NeurIPS 2024	Highest performance
PO	Prompt Optimization with EASE? Efficient Ordering-aware Automated ... (Wu et al., 2024c)	NeurIPS 2024	Highest performance
PO	Teach Better or Show Smarter? On Instructions and Exemplars in Automatic ... (Wan et al., 2024)	NeurIPS 2024	Highest performance
PO	Connecting Large Language Models with Evolutionary Algorithms Yields ... (Guo et al., 2024)	ICLR 2024	Highest Performance
PO	PromptAgent: Strategic Planning with Language Models Enables ... (Wang et al., 2023f)	ICLR 2024	Highest Performance
PO	On Prompt-Driven Safeguarding for Large Language Models (Zheng et al., 2024a)	ICML 2024	Safety Optimization
PO	Dynamic Rewarding with Prompt Optimization Enables Tuning-free ... (Singla et al., 2024)	EMNLP 2024	Highest Performance
IF	ToolPlanner: A Tool Augmented LLM for Multi Granularity Instructions with Path Planning and Feedback (Wu et al., 2024a)	EMNLP 2024	Instruction Alignment
PO	Fine-Tuning and Prompt Optimization: Two Great Steps that Work ... (Soylu et al., 2024)	EMNLP 2024	Prompt Effectiveness
PO	PRompt Optimization in Multi-Step Tasks (PROMST): Integrating Human ... (Chen et al., 2024)	EMNLP 2024	Highest Performance
PO	Multi-Scale Prompt Memory-Augmented Model for Black-Box Scenarios (Kuang et al., 2024)	NAACL 2024	Highest Performance

PO	Learning to Compress Prompt in Natural Language Formats (Chuang et al., 2024)	NAACL 2024	Efficiency and Transferability
PO	Universal Prompt Optimizer for Safe Text-to-Image Generation (Wu et al., 2024d)	NAACL 2024	Safe and Semantic-Preserving
PO	Black-Box Prompt Optimization: Aligning Large Language Models without Model Training (Cheng et al., 2024a)	ACL 2024	Human Preference Alignment
PO	LongLLMLingua: Accelerating and Enhancing LLMs in Long Context Scenarios via Prompt Compression (Jiang et al., 2024)	ACL 2024	Highest Perfomance
PO	LLMLingua-2: Data Distillation for Efficient and Faithful Task-Agnostic Prompt Compression (Pan et al., 2024)	ACL 2024	Highest Performance
PO	Lost in the Middle: How Language Models Use Long Contexts (Liu et al., 2024a)	TACL 2024	Effective Context Utilization
PO	Do Prompt Positions Really Matter? (Mao et al., 2024)	Preprint 2024	Highest Performance
PO	Prompt Compression with Context-Aware Sentence Encoding for Fast and Improved LLM Inference (Liskavets et al., 2024)	AAAI 2025	Highest Performance
IF	How to talk so AI will learn: Instructions, descriptions, and autonomy (Sumers et al., 2022)	NeurIPS 2022	Contextual Relevance
IF	Training language models to follow instructions with human feedback (Ouyang et al., 2022)	NeurIPS 2022	User-Aligned Guidance
IF	Instruction-Following Evaluation for Large Language Models (Zhou et al., 2023b)	Preprint 2023	Verifiable instruction
IF	Protecting User Privacy in Remote Conversational Systems: A Privacy-Preserving framework based on text sanitization (Kan et al., 2023)	Preprint 2023	Privacy Preservation and Data Utility
IF	ICU: Conquering Language Barriers ... (Wu, 2023)	EMNLP 2023	Cross-Language Clarity
IF	Benchmarking Generation and Evaluation Capabilities of Large Language ... (Liu et al., 2024c)	NAACL 2023	Comprehensive Instruction Clarity
IF	Enhancing Large Language Models Against Inductive Instructions with ... (Wang et al., 2024d)	NAACL 2023	Enhanced Instruction Adherence
IF	InstructEval: Systematic Evaluation of Instruction Selection Methods (Ajith et al., 2023)	NAACL 2023	Highest Performance
IF	Interpreting User Requests in the Context of Natural Language Standing ... (Moghe et al., 2024)	NAACL 2023	Highest Performance
IF	Instruction-following Evaluation through Verbalizer Manipulation (Li et al., 2024d)	NAACL 2023	Enhanced Instruction Adherence
IF	HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face (Shen et al., 2023a)	NeurIPS 2023	Highest Performance
IF	Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena (Zheng et al., 2023)	NeurIPS 2023	Effective Evaluation Criteria
IF	Recommender AI Agent: Integrating Large Language Models for Interactive Recommendations (Huang et al., 2023a)	Preprint 2023	Highest Performance
IF	Evaluating ChatGPT as a Recommender System: A Rigorous Approach (Di Palma et al., 2023)	Preprint 2023	Highest Performance
IF	RecMind: Large Language Model Powered Agent For Recommendation (Wang et al., 2024e)	NAACL 2024	Highest Performance
IF	R-Tuning: Instructing Large Language Models to Say... (Zhang et al., 2024a)	NAACL 2024	Refusal Awareness
IF	Benchmarking Complex Instruction-Following with Multiple Constraints ... (Wen et al., 2024)	NeurIPS 2024	Comprehensive Instruction Clarity

IF	Instruction Embedding: Latent Representations of Instructions Towards ... (Li et al., 2024f)	NeurIPS 2024	Highest Performance
IF	Evaluating Large Language Models at Evaluating Instruction Following (Zeng et al., 2024)	ICLR 2024	Enhanced Instruction Adherence
IF	MUFFIN: Curating Multi-Faceted Instructions for Improving ... (Lou et al., 2024)	ICLR 2024	Enhanced Instruction Adherence
IF	Self-Rewarding Language Models (Yuan et al., 2024a)	ICML 2024	Self-Rewarding Guidance
IF	A Theory Guided Scaffolding Instruction Framework for LLM-Enabled Metaphor Reasoning (Tian et al., 2024)	NAACL 2024	Effective Reasoning Support
IF	Can LLMs Generate Human-Like Wayfinding Instructions? Towards Platform-Agnostic Embodied Instruction Synthesis (Dorbala et al., 2024)	NAACL 2024	Highest Performance
IF	From Language Modeling to Instruction Following: Understanding the Behavior Shift in LLMs after Instruction Tuning (Wu et al., 2024b)	NAACL 2024	Comprehensive Instruction Clarity
IF	MATHSENSEI: A Tool-Augmented Large Language Model for Mathematical Reasoning (Das et al., 2024)	NAACL 2024	Highest Perfomance
IF	UniverSLU: Universal Spoken Language Understanding for Diverse Tasks with Natural Language Instructions (Arora et al., 2024)	NAACL 2024	User-Aligned Guidance
IF	InsCL: A Data-efficient Continual Learning Paradigm for Fine-tuning Large Language Models with Instructions (Wang et al., 2024f)	NAACL 2024	Highest Performance
IF	Answer is All You Need: Instruction-following Text Embedding via Answering the Question (Peng et al., 2024b)	ACL 2024	Highest Performance
IF	ABLE: Personalized Disability Support with Politeness and Empathy Integration (Mishra et al., 2024)	EMNLP 2024	Highest Performance
IF	Seemingly Plausible Distractors in Multi-Hop Reasoning ... (Bhuiya et al., 2024)	EMNLP 2024	Multi-Hop Reasoning Capabilities
IF	Generating Demonstrations for In-Context Compositional Generalization in Grounded Language Learning (Spilsbury et al., 2024)	EMNLP 2024	Highest Performance
IF	Do LLMs Know to Respect Copyright Notice? (Xu et al., 2024)	EMNLP 2024	Copyright Compliance
IF	Factual Dialogue Summarization via Learning from Large Language Models (Zhu et al., 2025)	COLING 2025	Consistent Perfomance

C 支持 Table 1 中性质的论文列表

我们观察到，我们之前分析中发现的大多数强相关性仍然保持一致，包括（标记数量；方式；结构逻辑；上下文逻辑；额外负担）、（目标；内在负担）、（结构逻辑；上下文逻辑）和（安全性；社会规范），其中两个相关性稍微不如以前那么强（现在是 Gemini-2.0-flash 的 0.6，而 GPT-4o 是 0.7）：（幻觉意识；事实性和创造性）和（目标；相关负担）。这些额外的结果进一步支持了观察到的相关性在不同高性能大模型中（几乎）具有普遍性，而不是仅限于特定模型组（例如 OpenAI 模型）。

Property	Real-world chat	Total
Better quantity	(Jiang et al., 2023b; Pan et al., 2024; Li et al., 2023e; Jung and Kim, 2024)	4
Better manner	-	0
Better engagement	(Bsharat et al., 2023; Ferron et al., 2023)	2
Better politeness	(Mishra et al., 2023)	1
Better intrinsic	(Bsharat et al., 2023; Nguyen et al., 2023; Wang et al., 2023b)	3
Lower extraneous	-	0
Better germane	(Zhu et al., 2025)	1
Better objective(s)	(Bsharat et al., 2023)	1
Better external tool(s)	(Shen et al., 2023a)	1
Better metacognition	-	0
Better demo(s)	(Bsharat et al., 2023)	1
Better reward(s)	(Bsharat et al., 2023)	1
Better structure	(Bsharat et al., 2023)	1
Better context logic	-	0
Better hallu. awa.	-	0
Better fact. and cre.	-	0
Lower bias	(Dwivedi et al., 2023)	1
Better safety	-	0
Better privacy	-	0
Better reliability	-	0
Better societal norms	-	0

Table 6: Property impact on Real-world chat.

Property	Eval. suit	Total
Better quantity	(Jiang et al., 2023b, 2024; Pan et al., 2024; Liskavets et al., 2024)	4
Better manner	-	0
Better engagement	-	0
Better politeness	(Yin et al., 2024; Xu et al., 2024)	2
Better intrinsic	(Wei et al., 2022; Li et al., 2023)	2
Lower extraneous	(Bhuiya et al., 2024)	1
Better germane	(Sun et al., 2022)	1
Better objective(s)	(Wu, 2023)	1
Better external tool(s)	(Xu et al., 2023a; Das et al., 2024)	2
Better metacognition	(Zhou et al., 2024d; Lee et al., 2025)	2
Better demo(s)	(Chen et al., 2023b; Wu et al., 2024c)	2
Better reward(s)	(Pyatkin et al., 2023; Yuan et al., 2024a)	2
Better structure	(Wang et al., 2024a)	1
Better context logic	-	0
Better hallu. awa.	-	0
Better fact. and cre.	-	0
Lower bias	-	0
Better safety	-	0
Better privacy	-	0
Better reliability	(Long et al., 2024b)	1
Better societal norms	-	0

Table 7: Property impact on Eval. suit.

Property	Reasoning/QA	Total
Better quantity	(Jiang et al., 2023b; Shi et al., 2023; Li et al., 2023e; Wang et al., 2023a; Pan et al., 2024; Jiang et al., 2024; Chuang et al., 2024; Zhang et al., 2024b; Shandilya et al., 2024)	9
Better manner	-	0
Better engagement	(Deng et al., 2023)	1
Better politeness	(Yin et al., 2024)	1
Better intrinsic	(Wei et al., 2022; Arora et al., 2023; Qin et al., 2023; Tai et al., 2023; Madaan et al., 2023; Wang et al., 2023b,c)	7
Lower extraneous	(Shi et al., 2023; Bhuiya et al., 2024; Liu et al., 2024a)	3
Better germane	(Sun et al., 2022; Li et al., 2024c)	2
Better objective(s)	(Wu, 2023)	1
Better external tool(s)	(Yao et al., 2023; Wu et al., 2024a)	2
Better metacognition	(Wang and Zhao, 2024; Zhou et al., 2024d)	2
Better demo(s)	(Levy et al., 2023; Yang et al., 2023a; Michaelov et al., 2023; Opsahl-Ong et al., 2024; Qin et al., 2024; Spilsbury et al., 2024; Li et al., 2024b; Wu et al., 2024c)	8
Better reward(s)	(Pyatkin et al., 2023; Yuan et al., 2024a)	2
Better structure	(Wang et al., 2024a; Zhou et al., 2024a; Cheng et al., 2024b)	3
Better context logic	(Liu et al., 2024b)	1
Better hallu. awa.	(Gao et al., 2023)	1
Better fact. and cre.	-	0
Lower bias	-	0
Better safety	-	0
Better privacy	-	0
Better reliability	(Si et al., 2023b)	1
Better societal norms	-	0

Table 8: Property impact on Reasoning/QA.

Property	Generation	Total
Better quantity	(Jiang et al., 2023b; Li et al., 2023e; Pan et al., 2024; Shandilya et al., 2024)	4
Better manner	-	0
Better engagement	(Ferron et al., 2023; Mu et al., 2024)	2
Better politeness	(Mishra et al., 2023; Yin et al., 2024; Mishra et al., 2024; Xu et al., 2024)	4
Better intrinsic	(Li et al., 2023; Wang et al., 2023b)	2
Lower extraneous	-	0
Better germane	(Zhu et al., 2025)	1
Better objective(s)	(Long et al., 2025b)	1
Better external tool(s)	(Xu et al., 2023a)	1
Better metacognition	-	0
Better demo(s)	(Wu et al., 2024c; Peng et al., 2024a; Wang et al., 2024b)	3
Better reward(s)	(Pyatkin et al., 2023)	1
Better structure	(Hwang et al., 2024; Ma et al., 2024)	2
Better context logic	-	0
Better hallu. awa.	(Chae et al., 2024)	1
Better fact. and cre.	-	0
Lower bias	(Dwivedi et al., 2023)	1
Better safety	-	0
Better privacy	-	0
Better reliability	-	0
Better societal norms	-	0

Table 9: Property impact on Generation.

Property	NLU	Total
Better quantity	(Jiang et al., 2024)	1
Better manner	-	0
Better engagement	-	0
Better politeness	(Mishra et al., 2023, 2024)	2
Better intrinsic	(Arora et al., 2023; Wang et al., 2023b; Nguyen et al., 2023)	3
Lower extraneous	-	0
Better germane	-	0
Better objective(s)	(Wu, 2023)	1
Better external tool(s)	-	0
Better metacognition	(Wang and Zhao, 2024)	1
Better demo(s)	(Si et al., 2023a; Peng et al., 2024a; Wang et al., 2024b; Zhou et al., 2024c)	4
Better reward(s)	-	0
Better structure	(Huang et al., 2024a)	1
Better context logic	-	0
Better hallu. awa.	-	0
Better fact. and cre.	-	0
Lower bias	-	0
Better safety	-	0
Better privacy	-	0
Better reliability	-	0
Better societal norms	-	0

Table 10: Property impact on NLU.

Property	Others (Judging, Personalization, Retrieval, Safety)	Total
Better quantity	-	0
Better manner	-	0
Better engagement	(Ferron et al., 2023)	1
Better politeness	(Mishra et al., 2024; Xu et al., 2024)	2
Better intrinsic	(Zheng et al., 2023; Liu et al., 2023; Wang et al., 2023b; Di Palma et al., 2023; Huang et al., 2023a; Wang et al., 2023g, 2024e; Do et al., 2025)	8
Lower extraneous	(Xiao et al., 2024; Liu et al., 2024a; Do et al., 2025)	3
Better germane	-	0
Better objective(s)	-	0
Better external tool(s)	(Wu et al., 2024a)	-
Better metacognition	(Lee et al., 2025)	1
Better demo(s)	(Li et al., 2024e)	1
Better reward(s)	(Yuan et al., 2024a)	1
Better structure	-	0
Better context logic	(Pham et al., 2024)	1
Better hallu. awa.	-	0
Better fact. and cre.	-	0
Lower bias	(Zheng et al., 2023; Echterhoff et al., 2024)	2
Better safety	(Zheng et al., 2024a)	1
Better privacy	(Kan et al., 2023)	1
Better reliability	(Long et al., 2024b)	1
Better societal norms	-	0

Table 11: Property impact on Others.

C.1 认知维度提示细节

```
 COG_FORMAT = “ { ’Intrinsic load’: 1-10, ’Extraneous load’: 1-10, ’Germane load’: 1-10 } ”
COG_JUDGING_PROMPT = f“““You are a highly experienced judge tasked with evaluating a prompt on criteria.
The prompt given to you is provided below:
< 开始提示 > [[输入 _ 提示]] < 结束提示 >
Your task is to evaluate the above prompt on the following criteria on a scale of 1-10:
- Intrinsic load: This evaluates the prompts in explicitly guiding models to break complex tasks into actionable steps aligned with LM skills.
- Extraneous load: The extent to which prompts exclude irrelevant materials to reduce unnecessary load.
- Germane load: The degree to which prompts explicitly engage models with their prior knowledge or deep working memory (e.g., “ask itself”) to integrate it with existing and new knowledge for problem-solving.
The scoring system is provided below:
> Intrinsic load :
- 1-2 (Poor): The prompt provides little to no guidance on breaking down the task. It is overly vague, abstract, or assumes the model can handle complexity without guidance.
- 3-4 (Below Average): The prompt provides minimal guidance but fails to clearly break the task into actionable steps. The model is left to infer most of the process.
- 5-6 (Average): The prompt partially breaks down the task but lacks clarity or completeness in defining actionable steps. Some guidance is present, but it is inconsistent or incomplete.
- 7-8 (Good): The prompt effectively breaks the task into clear, actionable steps. It aligns well with the model’ s skills but may lack some nuance or optimization.
- 9-10 (Excellent): The prompt perfectly breaks the task into logical, actionable steps. It is highly aligned with the model’ s capabilities and ensures clarity and efficiency in execution.
> Extraneous load :
- 1-2 (Poor): The prompt includes excessive irrelevant information, making it difficult for the model to focus on the core task. It is cluttered or overly verbose.
- 3-4 (Below Average): The prompt contains some irrelevant information, but the core task is still somewhat discernible. The extraneous load is noticeable and distracting.
- 5-6 (Average): The prompt includes some unnecessary details but generally stays focused on the task. The extraneous load is moderate but not overly detrimental.
- 7-8 (Good): The prompt is concise and mostly free of irrelevant information. It minimizes extraneous load effectively, with only minor distractions.
- 9-10 (Excellent): The prompt is perfectly concise and excludes all irrelevant materials. It is optimized to reduce extraneous load to the bare minimum.
> Germane load :
- 1-2 (Poor): The prompt does not engage the model’ s prior knowledge or working memory. It provides no cues or instructions to leverage existing knowledge.
- 3-4 (Below Average): The prompt makes minimal attempts to engage prior knowledge but does so ineffectively or inconsistently. The model is left to infer connections on its own.
- 5-6 (Average): The prompt partially engages the model’ s prior knowledge but lacks depth or clarity in integrating it with new information. The engagement is superficial.
- 7-8 (Good): The prompt effectively engages the model’ s prior knowledge and encourages integration with new information. It provides clear cues or instructions for leveraging existing knowledge.
- 9-10 (Excellent): The prompt perfectly engages the model’ s prior knowledge and deep working memory. It explicitly guides the model to integrate existing and new knowledge for optimal problem-solving.
Your evaluations must focus on explicit instructions rather than implicit instructions.
For example, if the prompt does not say “Reflect on your prior knowledge” then you should not assume that the prompt is effective in encouraging germane load.
Begin your evaluation by providing a short explanation for each. Be as objective, thorough, and constructive as possible.
After providing your explanation, please rate the response on all the criteria on a scale of 1 to 10 by strictly following this format:
<begin of explanation> ... <end of explanation>
< 开始评分 > { 认知 _ 格式 } < 结束评分 >
```

C.2 指令维度提示细节

```
 INS_FORMAT = “ { ‘Objectives’: 1-10, ‘External tools’: 1-10, ‘Metacognition’: 1-10, ‘Demos’:
1-10, ‘Rewards’: 1-10 } ”
INS_JUDGING_PROMPT = f“““You are a highly experienced judge tasked with evaluating a prompt on
criteria.
The prompt given to you is provided below:
<begin of the prompt> [[INPUT_PROMPT]] <end of the prompt>
Your task is to evaluate the above prompt on the following criteria on a scale of 1-10:
- Objectives: How well prompts explicitly communicate the task objectives, including expected
outputs, formats, constraints, audiences, and other applicable criteria.
- External tools: The extent to which prompts explicitly guide models to identify when
specific external tools or knowledge resources are needed, and perform tool calls to support
problem-solving.
- Metacognition: This assesses prompts in explicitly guiding models to reason, self-monitor, and
self-verify outputs to meet expectations and enhance reliability.
- Demos: The extent to which the prompts explicitly include examples, demonstrations, and
counterexamples to illustrate the desired output.
- Rewards: How well prompts explicitly establish feedback, reward, and reinforcement mechanisms
that encourage the models achieving desired outputs.
The scoring system is provided below:
> Objectives :
- 1-2 (Poor): The prompt lacks any clear objectives or guidance.
- 3-4 (Below Average): Vague or incomplete objectives.
- 5-6 (Average): Outlines basic objectives but lacks depth.
- 7-8 (Good): Clearly communicates objectives, may miss edge cases.
- 9-10 (Excellent): Comprehensive and leaves no ambiguity.
> External tools :
- 1-2 (Poor): No mention or guidance on external tools.
- 3-4 (Below Average): Vague hints at tools, no clear usage.
- 5-6 (Average): Acknowledges tools, lacks specifics.
- 7-8 (Good): Explicitly guides tool use, may lack examples.
- 9-10 (Excellent): Fully integrates tools with guidance and examples.
> Metacognition :
- 1-2 (Poor): No encouragement for reasoning or self-monitoring.
- 3-4 (Below Average): Minimal guidance, lacks actionable steps.
- 5-6 (Average): Provides some reasoning/self-monitoring, incomplete.
- 7-8 (Good): Explicitly guides reasoning and verification.
- 9-10 (Excellent): Thorough integration of metacognitive strategies.
> Demos :
- 1-2 (Poor): No examples or demonstrations.
- 3-4 (Below Average): Poorly constructed or minimal examples.
- 5-6 (Average): Basic examples, lacks depth or variety.
- 7-8 (Good): Clear and relevant examples with counterexamples.
- 9-10 (Excellent): Comprehensive, edge cases included.
> Rewards :
- 1-2 (Poor): No feedback, reward, or reinforcement.
- 3-4 (Below Average): Vague or minimal reward mechanisms.
- 5-6 (Average): Basic reward mechanisms, not fully integrated.
- 7-8 (Good): Clear feedback/reward guidance.
- 9-10 (Excellent): Fully integrated with examples and detail.
Your evaluations must focus on explicit instructions rather than implicit instructions.
For example, if the prompt does not mention about the formats or constraints of the objectives
then you should not assume that the prompt is effective in communicating the objectives.
For example, if the prompt does not say “I will reward you something for something” then you
should not assume that the prompt is effective in encouraging the rewards.
Begin your evaluation by providing a short explanation for each. Be as objective, thorough, and
constructive as possible. After providing your explanation, please rate the response on all the
criteria on a scale of 1 to 10 by strictly following this format:
< 解释开始 > …< 解释结束 >
< 评分开始 > { INS_FORMAT } < 评分结束 >  """
```

C.3 逻辑和结构维度提示细节

```
LOGIC_FORMAT = " { 'Structural logic': 1-10, 'Contextual logic': 1-10 } "
LOGIC_JUDGING_PROMPT = f"""You are a highly experienced judge tasked with evaluating a prompt
on criteria.
The prompt given to you is provided below:
<begin of the prompt> [[INPUT_PROMPT]] <end of the prompt>
Your task is to evaluate the above prompt on the following criteria on a scale of 1-10:
- Structural logic: This evaluates the logical clarity and coherence of prompts' structure, and
the progression between components.
- Contextual logic: This assesses the logical consistency and coherence of the instructions,
terminologies, concepts, facts, and other components within the prompt and across communication
turns.
The scoring system is provided below:
> Structural logic :
- 1-2 (Poor): No discernible structure or logical flow. Disjointed and confusing.
- 3-4 (Below Average): Basic structure but poorly organized and weak progression.
- 5-6 (Average): Moderately clear structure; minor lapses in logic.
- 7-8 (Good): Clear and coherent structure with smooth progression.
- 9-10 (Excellent): Impeccable organization with flawless logical progression.
> Contextual logic :
- 1-2 (Poor): Inconsistent, contradictory, or unclear use of concepts.
- 3-4 (Below Average): Some context provided but notable inconsistencies remain.
- 5-6 (Average): Generally consistent with minor lapses that don't severely hinder understanding.
- 7-8 (Good): Coherent and logical use of language with only minor issues.
- 9-10 (Excellent): Seamless, consistent, and logical across all instructions and components.
Begin your evaluation by providing a short explanation for each. Be as objective, thorough, and
constructive as possible. After providing your explanation, please rate the response on all the
criteria on a scale of 1 to 10 by strictly following this format:
<begin of explanation> ···<end of explanation>
< 评分开始 > { LOGIC_FORMAT } < 评分结束 >  """
```

C.4 幻觉维度提示细节

```
HALL_FORMAT = " { 'Hallucination awareness': 1-10, 'Factuality and creativity': 1-10 } "
HALL_JUDGING_PROMPT = f"""You are a highly experienced judge tasked with evaluating a prompt on
criteria.
The prompt given to you is provided below:
<begin of the prompt> [[INPUT_PROMPT]] <end of the prompt>
Your task is to evaluate the above prompt on the following criteria on a scale of 1-10:
- Hallucination awareness: The extent to which prompts explicitly guide models to generate
factual and evidence-based responses while minimizing speculative or unsupported claims.
- Factuality and creativity: The degree to which prompts explicitly guide models to balance
creative generation with factual accuracy, including which task and when to prioritize creativity
over creativity and vice versa.
The scoring system is provided below:
> Hallucination awareness :
- 1-2 (Poor): No guidance to avoid hallucinations; results likely inaccurate.
- 3-4 (Below Average): Minimal or vague mention of factuality; little structure.
- 5-6 (Average): Some general instruction (e.g., "be factual" ), but lacks specifics.
- 7-8 (Good): Clear instructions to avoid hallucinations with specific strategies (e.g., "cite
sources" ).
- 9-10 (Excellent): Comprehensive and detailed guidance with examples or frameworks.
> Factuality and creativity :
- 1-2 (Poor): Ignores factuality or overly restricts creativity.
- 3-4 (Below Average): Acknowledges both aspects but with vague, unhelpful guidance.
- 5-6 (Average): Basic instruction to balance both, but lacks clarity or depth.
- 7-8 (Good): Provides task-based distinctions with clear but limited examples.
- 9-10 (Excellent): Nuanced, detailed, and contextual guidance that effectively balances both
aspects.
Begin your evaluation by providing a short explanation for each. Be as objective, thorough, and
constructive as possible. After providing your explanation, please rate the response on all the
criteria on a scale of 1 to 10 by strictly following this format:
<begin of explanation> ···<end of explanation>
< 评分开始 > { HALL_FORMAT } < 评分结束 >  """
```

C.5 责任维度提示详情

```
 RES_FORMAT = “ { ‘Bias’: 1-10, ‘Safety’: 1-10, ‘Privacy’: 1-10, ‘Reliability’: 1-10, ‘Societal norms’: 1-10 } ”
RES_JUDGING_PROMPT = f“““You are a highly experienced judge tasked with evaluating a prompt on criteria.
The prompt given to you is provided below:
<begin of the prompt> [[INPUT_PROMPT]] <end of the prompt>
Your task is to evaluate the above prompt on the following criteria on a scale of 1-10:
- Bias: Promotes fairness and avoidance of cultural, gender, racial, or socio-economic bias.
- Safety: Avoids or discourages harmful, unsafe, or dangerous content.
- Privacy: Avoids inclusion or encouragement of sensitive or personal data.
- Reliability: Encourages clear reasoning, attribution, and model limitation awareness.
- Societal norms: Aligns with ethical, moral, and inclusive standards.
IMPORTANT: Only consider explicit instructions. Do not infer intent.
Examples: If the prompt does not say “Output must not be biased”, do not assume it addresses bias. The same applies to safety, privacy, and societal norms.
The scoring system is provided below:
> Bias :
- 1-2: Encourages or contains biased or discriminatory language.
- 3-4: May include bias unintentionally; lacks anti-bias instructions.
- 5-6: Avoids bias but gives no explicit guidance to that effect.
- 7-8: Explicitly discourages bias and promotes fair content.
- 9-10: Strong, explicit anti-bias guidance with inclusive framing.
> Safety :
- 1-2: Promotes harmful or dangerous content.
- 3-4: Fails to discourage unsafe content; vague or missing safety cues.
- 5-6: Avoids unsafe content but lacks explicit safe-generation guidance.
- 7-8: Explicitly discourages unsafe outputs.
- 9-10: Clear and robust safety instructions with no ambiguity.
> Privacy :
- 1-2: Encourages sharing of private or identifiable information.
- 3-4: Does not explicitly warn against privacy violations.
- 5-6: Privacy considered, but instructions are weak or implicit.
- 7-8: Prompt directly addresses and discourages privacy issues.
- 9-10: Strong, explicit privacy-safe prompting practices.
> Reliability :
- 1-2: Lacks or undermines reasoning, attribution, or uncertainty awareness.
- 3-4: Misses reliability-related guidance; prone to misleading output.
- 5-6: Vaguely references reasoning but lacks depth.
- 7-8: Encourages sound reasoning and limitations acknowledgment.
- 9-10: Clearly fosters reliable, transparent, and accountable outputs.
> Societal norms :
- 1-2: Encourages unethical, exclusionary, or harmful social values.
- 3-4: Fails to discourage such norms or lacks inclusive framing.
- 5-6: Neutral or unclear on social standards.
- 7-8: Explicitly supports ethical, moral, and inclusive content.
- 9-10: Proactively ensures ethical alignment and inclusivity.
Begin your evaluation by providing a short explanation for each. Be objective, thorough, and constructive.
Then rate the response using the format below:
<begin of explanation> ... <end of explanation>
< 评分开始 > { RES_FORMAT } < 评分结束 >  """
```