What Makes a *Good* Natural Language Prompt?

Do Xuan Long[1,2], **Duy Dinh**[1*], **Ngoc-Hai Nguyen**[1*],
Kenji Kawaguchi[1], **Nancy F. Chen**[2], **Shafiq Joty**[3,4], **Min-Yen Kan**[1]

[1]National University of Singapore, [2]Institute for Infocomm Research (I[2]R), A*STAR,
[3]Salesforce AI Research, [4]Nanyang Technological University, Singapore

xuanlong.do@u.nus.edu, {dinhcongduy131200, haibeo2552001}@gmail.com,
{kenji,knmnyn}@nus.edu.sg, sjoty@salesforce.com, nfychen@i2r.a-star.edu.sg

Abstract

As large language models (LLMs) have progressed towards more human-like and human–AI communications prevalent, prompting has emerged as a decisive component. However, there is limited conceptual consensus on what exactly quantifies *natural language* prompts. We attempt to address this question by conducting a meta-analysis surveying 150+ prompting-related papers from leading NLP and AI conferences (2022–2025), and blogs. We propose a *property- and human-centric* framework for evaluating prompt quality, encompassing 21 properties categorized into six dimensions. We then examine how existing studies assess their impact on LLMs, revealing their imbalanced support across models and tasks, and substantial research gaps. Further, we analyze correlations among properties in high-quality natural language prompts, deriving prompting recommendations. We then empirically explore multi-property prompt enhancements in reasoning tasks, observing that single-property enhancements often have the greatest impact. Finally, we discover that instruction-tuning on property-enhanced prompts can result in better reasoning models. Our findings establish a foundation for property-centric prompt evaluation and optimization, bridging the gaps between human–AI communication and opening new prompting research directions[1].

1 Introduction

Pre-trained LLMs (Brown et al., 2020; Chowdhery et al., 2022; OpenAI, 2022; Touvron et al., 2023a; Team et al., 2023; Guo et al., 2025), renowned for their ability to generate human-like text, have exhibited exceptional performance across various natural language processing tasks. While their effectiveness is profoundly influenced by the quality of *natural language* prompts (Sahoo et al., 2024),

the art and science of effective prompts remain underexplored. As human–AI interactions become ubiquitous, developing a deeper understanding of these natural language prompts is crucial since they serve as the primary communication interface between humans and AI systems.

Despite the importance of understanding natural language prompts, there remains limited consensus on how to quantify them. Current approaches rely predominantly on *outcome-centric* measurements, such as model-specific performance metrics (Deng et al., 2022; Lin et al., 2024; Shi et al., 2024) and iterative trial-and-error testing (Pryzant et al., 2023; Long et al., 2024a) possibly resulting in prompts optimized for machine interpretation rather than human understanding. This can lead to challenges in interpreting and verifying them, potentially introducing adversarial behaviors in LLMs (Zou et al., 2023; Zhu et al., 2023) and raising concerns about alignment, transparency, overall reliability, and even human–AI communications.

Several prompting studies (Bsharat et al., 2023; Lin, 2024) and guidelines (OpenAI, 2024b; Anthropic, 2024) recently introduce recommendations enhancing certain *properties* of prompts such as "Specify the desired length of the output". These *property-centric* recommendations, focusing on prompt quality rather than model performance, offer interpretable strategies and can complement outcome-centric approaches. However, they have key limitations. First, there is no unified or theoretical property-centric framework that abstractly encompasses such practical recommendations, hindering systematic understanding, analysis, and comparison of these strategies. Second, it is unclear whether these recommendations offer universal benefits across models and tasks or are more model- or task-specific. Third, the interactions among these recommendations and their combined effects on model performance remain understudied.

To address these limitations, we present a meta-

---

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

[1]Our codes and data will be made publicly available at here.

analysis to systematically study natural language prompts. We survey prompting papers from top NLP and AI conferences in 2022–2025 and blogs written by top-tech companies (see §B for the full list) and identify 21 prompt-level properties across six evaluation dimensions offering a novel *property- and human-centric* perspective (§3). Building on this, we examine how prior studies assess which models and tasks benefit from enhancing each property, uncovering significant imbalanced distributions in the #papers supporting each property across models and tasks, and research gaps (§4). Next, we analyze correlations among these properties in a subset of high-quality natural language prompts, deriving practical recommendations for prompt design (§5). We then conduct a case study on reasoning tasks to understand the impact of enhancing multiple prompting properties on model performance (§6). Notably, we observe that different prompting properties influence models differently across tasks, and enhancing multiple properties does not always lead to greater improvements; a single property is often the most effective, and fine-tuning models on property-enhanced instructions further boosts such effectiveness. Our contributions are summarized below:

1. We introduce a novel property- and human-centric framework for evaluating the quality of natural language prompts, identifying 21 key properties across six evaluation dimensions to shift the focus from outcome-centric to property-centric assessment.

2. We conduct a meta-analysis of prior studies from 2022–2025 NLP/AI conferences and blogs to investigate how these properties affect model performance, revealing significant research imbalances and gaps.

3. We analyze correlations among these properties in a curated set of high-quality prompts, deriving practical recommendations to guide effective prompt design.

4. We study prompting and fine-tuning models for reasoning tasks, finding that optimizing a single prompting property often outperforms combining multiple ones, with effects varying across tasks and models.

2   Related work

Prompt analysis.   Prompting plays a key role in harnessing the full potential of LLMs (Liu et al.,

2023a; Sahoo et al., 2024), driving significant prompt analysis research interest. Existing studies primarily focus on two key directions. The first analyzes the structural components of prompts, highlighting how their variants in terms of formatting (Long et al., 2025a) and phrasing (Yin et al., 2023) can lead to substantial performance differences, and their appearance rates (Ma et al., 2024). These studies aim to understand prompt components and their impact on model performance. The second analyzes prompts through practical experiments, providing design recommendations such as chain-of-thought prompting (Wei et al., 2022; Kojima et al., 2022), being polite with LLMs (Bsharat et al., 2023) and even sets of general guidelines (Anthropic, 2024; OpenAI, 2024b). However, these prompt analysis studies are often task-specific or focus on particular properties of prompts. In this work, for the first time, we introduce a unified property-centric framework that abstractly composites these practical recommendations, facilitating systematic understanding, analysis, and comparison of prompting strategies.

Prompt engineering and optimization.   Prompt engineering (Wei et al., 2022; Zhang et al., 2023; Zhou et al., 2023c) and optimization (Deng et al., 2022; Pryzant et al., 2023; Long et al., 2024a) aim to find prompts that maximize a language model's performance for a given task. While much of the existing research focuses on enhancing benchmark performance, there are emerging recent efforts emphasizing broader prompt properties such as clarity (Lin, 2024; Anthropic, 2024), politeness (Bsharat et al., 2023; Yin et al., 2024), structured formatting (OpenAI, 2024b), and even fairness in output generation (Ji et al., 2023; Yuan et al., 2023). However, it remains unclear whether these properties yield universal benefits across models and tasks or if their effects are model- or task-specific. Furthermore, their interactions and combined influence on model performance remain largely unexplored. We address these gaps in Sections 4 to 6.

3   Prompt quality evaluation

We begin our study by conducting a comprehensive survey of over 150 papers and blogs. Our methodology is straightforward: we first examine papers published in ACL, EMNLP, NAACL from ACL Anthology[2], and ICLR, and NeurIPS on

---

[2] https://aclanthology.org/

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		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 OFA	1 OFA	1	1 OFA	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: ⑤: ChatGPT / Codex; ∞: LLaMa / OPT / RoBERTa / BART; ⅇ: Qwen; M: Mistral; ⅃: Alpaca; Yi: Yi; G: 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. The distribution of papers supporting various properties is highly imbalanced across models and tasks. We discuss the findings in detail in §4.

OpenReview[3] from 2022 to 2025. Relevant papers are further identified through keyword searches on Google. While striving for thoroughness, we acknowledge the possibility of inadvertently omitting some related papers. We then manually identify prompting objectives and recommendations from these papers that influence model performance, and conceptualize them as prompt **properties**. These properties are defined below along with its evidence (denoted by abbreviation **e.b.**).

I. Communication and language. Prior studies highlight the importance of specific communication properties for desired LM outcomes. For example, Yin et al. (2024) find that impolite prompts degrade model results across tasks and languages, while Shi et al. (2023) discover that irrelevant contexts can distract LLMs, and more explicit prompts enhance model performance (Bsharat et al., 2023; Lin, 2024). Inspired by these and LLMs being more humanoid, prompt evaluation should consider human-like communication properties. We introduce four for evaluation, partially motivated by Grice's Maxims of Conversation (Grice, 1975):

- **Token quantity:** The extent to which prompts provide optimal and relevant information while minimizing token usage, balancing in-

formation completeness with efficiency (e.b. Shi et al. (2023); Jiang et al. (2023b)).

- **Manner:** The degree to which prompts are clear and direct (across turns) while minimizing unnecessary ambiguity, complexity, and confusion (e.b. Anthropic (2024)).

- **Interaction and engagement:** The extent to which the prompts explicitly encourage the models to gather the necessary details and requirements by asking questions of clarification or confirmation (e.b. Deng et al. (2023)).

- **Politeness:** The degree to which prompts maintain respectful, professional, and context-specific politeness, including the use of courteous language (e.g., "please", "thank you") (e.b. Yin et al. (2024)).

II. Cognition. Wei et al. (2022); Zhou et al. (2023a) pioneer in introducing prompting methods that decompose complex reasoning tasks into simpler steps, enhancing LLM performance. Subsequent studies extensively investigate strategies that optimize the subtasks to further align them with model capabilities (Khot et al., 2023; Suzgun and Kalai, 2024). In addition, Sun et al. (2022) show that integrating self-generated knowledge improves question answering performance of LLMs.

---

[3] https://openreview.net/

5837

Philosophically, these works imply that maximizing LLMs' learning and problem-solving requires meticulous management of their cognitive loads.

Sweller and Chandler (1991) introduce Cognitive Load Theory, categorizing cognitive loads into intrinsic (task complexity), extraneous (unclear or poorly designed instructions), and germane (efforts to understand, memorize, and organize information). Motivated by this, prompt evaluation should concern three loads on LLMs:

- **Manage intrinsic load:** This evaluates the prompts in explicitly guiding models to break complex tasks into actionable steps aligned with LM skills (e.b. Zhou et al. (2023a)).

- **Reduce extraneous load:** The extent to which prompts minimize unnecessary complexity via simplifying language and removing redundant or irrelevant information to reduce unnecessary load (e.b. OpenAI (2024b)).

- **Encourage germane load:** The degree to which prompts explicitly engage models with their prior knowledge or deep working memory (e.g., "ask itself" (Press et al., 2023)) to integrate it with existing and new knowledge for problem-solving (e.b. Sun et al. (2022); Mialon et al. (2023); Fan et al. (2024)).

III. Instruction. The instructional values of prompts are crucial for achieving the desired output (Sahoo et al., 2024). Drawing on Gagne's Nine Events of Instruction (Gagné, 1985) and the Metacognitive Theories (Schraw and Moshman, 1995), we present instructional criteria to evaluate them non-overlapping with other dimensions:

- **Objective(s):** How well prompts explicitly communicate the task objectives, including expected personae, outputs, formats, constraints, audiences, and other applicable criteria (e.b. Chang (2023); Long et al. (2025b)).

- **External tool(s):** The extent to which prompts explicitly guide models to identify when specific external tools or knowledge resources are needed that go beyond task objective(s), and perform corresponding external calls (e.b. Yao et al. (2023)).

- **Metacognition:** This assesses prompts in explicitly guiding models to reason, self-monitor, and self-verify outputs to meet ex-

pectations and enhance reliability (e.b. Wang and Zhao (2024)).

- **Demo(s):** The extent to which the prompts explicitly include examples, demonstrations, and counterexamples to illustrate the desired output (e.b. Dong et al. (2024)).

- **Reward(s):** How well prompts explicitly establish feedback and reinforcement mechanisms that encourage the models to achieve desired outputs (e.b. Bsharat et al. (2023)).

IV. Logic and structure. Coherent structural prompts are shown to be effective across various tasks (Wang et al., 2024a; Huang et al., 2024a). Moreover, prompting guidelines (Guide, 2024; OpenAI, 2024b) also recommend structuring input and output to obtain better performing prompts. For logic, recent studies (Wang et al., 2024g; Pham et al., 2024) highlight the importance of contextual consistency where knowledge conflicts within prompts substantially degrade LM performance. Building on these insights and the established human logic criteria for effective communication (Grice, 1975; Mercier and Sperber, 2011), we introduce two logical criteria:

- **Structural logic:** This evaluates the logical clarity and coherence of prompts' structure, and the progression between components (e.b. Wang et al. (2024a); Zhou et al. (2024b)).

- **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 (e.b. Pham et al. (2024)).

V. Hallucination. Prompting can lead to hallucination where models generate plausible but non-factual content (Huang et al., 2024b). While it remains challenging to anticipate whether and when a prompt triggers hallucination (Farquhar et al., 2024), prompts can be designed to encourage models to be aware of this critical issue. We propose that prompt evaluation should address two hallucination-related criteria:

- **Hallucination awareness:** The extent to which prompts explicitly guide models to generate factual and evidence-based responses while minimizing speculative or unsupported claims (e.b. Gao et al. (2023)).

- **Balancing factuality with 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. We have yet observed prompting methods designed for this criterion to date. However, Sinha et al. (2023) propose a training approach to balance these aspects for LMs.

In this dimension, we do not evaluate hallucination within prompts as it partially overlaps with the "Quantity" of Communication.

VI. Responsibility. This dimension emphasizes responsible prompting that mitigates concerns related to inclusion, privacy, safety, bias, reliability, fairness, transparency, and societal norms (Stahl and Eke, 2024; Hua et al., 2024), especially tasks involving sensitive topics or diverse audiences:

- **Bias:** The extent to which prompts are devoid of biases and explicitly encourage models to generate content that is free from cultural, gender, racial, or socio-economic biases and avoids stereotypes (e.b. Si et al. (2023b)).

- **Safety:** The degree to which prompts are free from unsafe content and explicitly encourage models to generate safe outputs, avoiding harmful content such as guidance on hazardous activities or weapon creation (e.g., Zou et al. (2023); Zheng et al. (2024a)).

- **Privacy:** The extent to which prompts do not contain sensitive privacy information and explicitly encourage the models to generate content free of personally sensitive or identifiable information (e.b. Edemacu and Wu (2024)).

- **Reliability:** How well prompts explicitly encourage explicit reasoning processes and attribution, including acknowledgment of model limitations and uncertainties (e.b. Si et al. (2023b); Long et al. (2024b)).

- **Societal norms:** The degree to which prompts exclude harmful norms and explicitly encourage models to generate inclusive and appropriate content aligning with widely accepted cultural, ethical, and moral standards (e.b., Yuan et al. (2024b)).

4 How do properties impact model performance?

To assess how the properties in §3 impact model performance, we analyze surveyed papers up to date to determine if these aspects were studied. We categorize the tasks explored into six groups: *(1) Real-world chat*, comprising benchmarks collected from real users such as AlpacaEval (Li et al., 2023c) and ShareGPT (ShareGPT, 2023); *(2) Evaluation suite*, which have multiple evaluation tasks such as MMLU (Hendrycks et al., 2021) and C-Eval (Huang et al., 2023c); *(3) Reasoning/QA*, covering reasoning and question-answering tasks like GSM8K (Cobbe et al., 2021) and HotpotQA (Yang et al., 2018); *(4) Generation*, focusing on text generation benchmarks such as summarization (Nallapati et al., 2016), and translation; *(5) NLU*, encompassing natural language understanding tasks like GLUE (Wang et al., 2018) and CommitmentBank (De Marneffe et al., 2019); and *(6) Others*, which include safety, personalization, judgment, and retrieval tasks. For each property, we gather three information: the number (#) of papers supporting the property, tasks that improving the property enhances their performance, and models. We discuss our findings in Table 1 below as actionable prompting recommendations.

Across tasks. There is logical alignment between task requirements and emphasized properties, with notable variations in the #papers supporting them across tasks. Firstly, **in real-world chats, communication properties emerge as the most supported, followed by instruction and cognition properties.** This arises from the practical use of LLMs, where users often craft rich and informative prompts to handle complex and varied tasks. These prompts can extend to tens of thousands of tokens and may sometimes include redundant details (Jiang et al., 2023b) or lack focus (Pan et al., 2024), particularly in multi-turn interactions (Ferron et al., 2023; Bsharat et al., 2023). Additionally, the significance of instruction properties reflects the interactive nature of chat, while cognition properties are essential for achieving desired outcomes. Secondly, **for evaluation suites, cognition, instruction, and communication properties are studied the most, with logic additionally emphasized in reasoning/QA tasks.** This aligns with the nature of these benchmarks, where well-cognitive instructions are crucial to strengthen LLM reasoners (Wei et al., 2022; Sun et al., 2022; Qin et al., 2023; Bhuiya

et al., 2024). Additionally, logic and structure logic also highlight the importance of systematic solving approaches for such tasks (Liu et al., 2024b; Cheng et al., 2024b). Thirdly, **for generation tasks, communication properties receive the most support, followed by the instruction.** This observation reflects the critical importance of efficient token management in generation tasks (Jiang et al., 2023b; Li et al., 2023e; Pan et al., 2024). Interestingly, several studies underscore the effectiveness of incorporating politeness (Mishra et al., 2023; Xu et al., 2024; Mishra et al., 2024; Yin et al., 2024), potentially reflecting the inherent biases of LLMs in processing benign rather than informal queries. Fourthly, **there are limited prompting studies for NLU tasks, and instruction properties appear to be the most explored, followed by cognition properties.** This can be explained by the fact that NLU tasks require models to accurately interpret prompts to reason deeply over language meaning or implications that go beyond surface-level understanding. Finally, **lower extraneous and better safeguard prompts have been shown to be effective for enhancing safety** (Xiao et al., 2024; Zheng et al., 2024a); **better intrinsic for personalization** (Lyu et al., 2024; Do et al., 2025); **better intrinsic and lower bias for judging** (Liu et al., 2023b; Zheng et al., 2023); and **lower extraneous for retrieval** (Liu et al., 2024a). While these findings highlight the nuanced alignment between task requirements and the properties shown, significant research gaps remain in exploring how enhancing other properties can further improve model performance on these tasks.

Across models and properties. We observe that the distribution of model explorations across properties is highly imbalanced. Specifically, OpenAI's proprietary models (CodeX (Chen et al., 2021), InstructGPT (Ouyang et al., 2022), ChatGPT (OpenAI, 2022), GPT-4/4o (OpenAI, 2023, 2024a)) have been the most extensively studied, followed by open-source LLaMa models (Touvron et al., 2023a,b; Dubey et al., 2024), and Google's models (FLAN (Chung et al., 2024), PaLM (Chowdhery et al., 2022), Gemma (Team et al., 2024)). This raises concerns regarding the transferrable effectiveness of these properties across models. We hypothesize that different properties benefit models differently and that these benefits may also differ across tasks, and validate it in §6.

Our analysis reveals task-specific versus universal properties: while better intrinsic load management, demonstrations, and external tools emerge as being universally effective, hallucination-awareness and responsibility appear to be more task-specific. Better intrinsic load highlights the current LLM weaknesses in implicitly and effectively decomposing complex tasks into more manageable subtasks without explicit guidance. Moreover, demonstration property underscores the value of learning from examples, while using external tools indicates that even with reduced cognitive load and good demonstrations, LLMs still benefit from tools for certain tasks.

Open questions (Oq). (**Oq1**) The effectiveness of properties varies across models due to differences in their inherent knowledge, thus, it is an open question whether and when a property beneficial to one model is useful for another. In addition, the missing entries in Table 1 highlight several critical yet unexplored properties. For instance, (**Oq2**), while reasoning is fundamental for humans to address tasks (Pearl, 1998), it is yet studied whether fostering deeper reasoning (improved germane load), reflective behavior (enhanced metacognition), or responsibility can enhance outcomes of LLMs in real-world chat, evaluation suits, and NLU tasks. Moreover, (**Oq3**), despite creativity's intuitive importance for multiple tasks such as generation, its effectiveness on LLMs remains an open question. Additionally, significant gaps remain in understanding property dynamics, particularly (**Oq4**) the conditions under which certain relevant or even task-irrelevant properties (Taveekitworachai et al., 2024) become effective and why. Lastly, (**Oq5**), the observation regarding task-specific and universal properties raises important questions about whether prompt engineering and optimization should prioritize one over the other and which is more significant. Studying (**Oq1**)-(**Oq5**) holds huge potential for advancing the efficiency, reliability, and alignment of LLMs. Future research could pursue comparative studies across diverse LLMs and tasks, develop quantifiable metrics to evaluate prompts across multiple dimensions, and explore hybrid strategies blending task-specific and universal prompt properties.

5 How do these properties appear and correlate in high-quality prompts?

We study high-quality natural language prompts to investigate the correlations between these proper-

ties to derive prompting recommendations. We manually collect our test set consisting of 765 single-turn prompts from prompt engineering papers, ChatGPT Prompts Collections[4], Awesome ChatGPT Prompts[5], Alpaca (Taori et al., 2023), Natural Instructions (Mishra et al., 2022), Complex Instructions (He et al., 2024), and 50 real-world multi-turn ($>$ 2 turns) conversations from LMSYS-Chat-1M (Zheng et al., 2024b) having 204 prompts, totaling 969 prompts in Appx.-Table 4. We evaluate these prompts across 21 proposed properties using GPT-4o-2024-11-20 (OpenAI, 2024a) with Self-consistency (Wang et al., 2022) as the judge. We also test open-source models, including DeepSeek R1 Distill Qwen 32B (Guo et al., 2025) and Mistral Small 24B It 2501 (Jiang et al., 2023a), as judges. However, we do not use them ultimately since we face significant evaluation format following issues (Long et al., 2025a) with DeepSeek and Mistral achieving only 65.42% and 71.19%. In addition to GPT-4o, we supplement our correlation results with findings from Gemini-2.0-flash (Team et al., 2023) in Appendix §E.

Methods. Automatic evaluations using LLMs can be unreliable, especially given the variability in evaluation prompts (Doostmohammadi et al., 2024). This creates a significant challenge in deriving reliable correlation conclusions from these evaluations. To mitigate this, we first manually label 50 random prompts in 21 properties and then design evaluation prompts to closely align with human judgments. Each annotation is agreed upon by our three prompting researchers with bachelor's degrees and at least six months' experience.

For each evaluation dimension, we begin with a prompt similar to the reference-free judging prompt on a scale of 1-10 proposed by Zheng et al. (2023). However, we find that this method results in drastically low Cohen's Kappa agreement (Cohen, 1960) with human raters; 15/21 topics achieved scores below 0.15, see Appx.-Fig. 2, "Ori. eval.". We then supplement an incremental grading system for each criterion, "Ori. eval. + Inc.", similar to (Yuan et al., 2024a), which significantly enhances agreements. Nevertheless, the germane load, objectives, rewards, and responsibility properties continue to score low. This is because the evaluator tends to score them higher than human based on implicit instructions rather than explicit cues as expected. To mitigate this issue, we explicitly instruct the evaluator to judge explicit signals, resulting in significantly better agreements ("Ours" in Appx.-Fig. 2). We evaluate all prompts with "Ours".

Findings. For this specific set of prompts, the property correlations are provided in Fig. 1. We do not consider correlations between properties if both have an average score below 5/10 (hatched by "\\") because low average scores naturally but may falsely suggest correlations. We observe 17/210 strong correlations ($\geq$ 0.7) among 21 properties. Some of them align with their real-world overlaps. For example, token quantity, manner, structural logic, contextual logic, and extraneous load reflect the natural correlations between token efficiency, clarity, directness, exclusion of irrelevant details, and logical coherence. Within dimensions, we notice structural logic strongly correlates with contextual logic; hallucination awareness with factuality and creativity; safety with societal norms. Surprisingly, we notice strong correlations between objectives and intrinsic load; objectives and germane load; hallucination awareness and reliability. These can be attributed to the nature of effective human prompting: as we optimize intrinsic and/or germane loads, we tend to articulate objectives more clearly. Similarly, enhancing hallucination awareness inherently contributes to reliability awareness.

We learn prompting recommendations from the analysis of this set of prompts. Firstly, optimizing prompts for directness, clarity, and conciseness may potentially improve token efficiency, and logical coherence, and reduce extraneous cognitive load. Secondly, clear objectives naturally emerge when prompts are logically structured guiding models to self-monitor their generation or execute tasks step-by-step. Thirdly, explicitly incorporating hallucination awareness in prompts may result in better reliability awareness. Lastly, since these prompts were carefully selected by humans, certain non-obvious correlations, such as those between structural logic, contextual logic, token quantity, and manner, suggest that these properties should be optimized jointly.

Open questions (Oq). While our analysis reveals certain correlations among prompt properties, several open questions remain for future investigation. First, **(Oq6)** we hypothesize that correlations may vary across different pools of prompts especially those that are task-specific, potentially leading to distinct prompting recommendations. We leave

---

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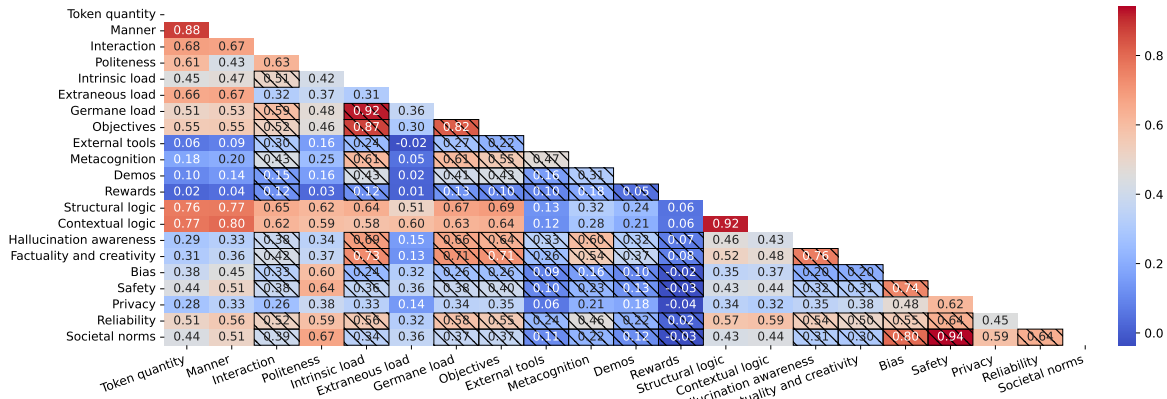[5] `https://github.com/f/awesome-chatgpt-prompts`

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.

this for future research. Secondly, **(Oq7)** when two properties exhibit a strong correlation, it remains to be determined whether enhancing prompts in one property causally enhances the other or if these properties merely co-occur within our dataset. Finally, **(Oq8)**, understanding how these correlations influence model performance is critical for advancing prompt optimization methods. The investigation of **(Oq6)**–**(Oq8)** offers a pathway to optimize LLM prompts by analyzing property correlations and eliminating optimization redundancies. Future work could use causal inference tools, such as structural equation modeling, to distinguish mere co-occurrence from influence, and conduct diverse model- and task-specific experiments to quantify these effects more precisely.

6 Should we enhance properties of prompts during experiments?

We perform a preliminary investigation into the impact of combining these properties on the performance of model reasoning. Our experiments are performed under two settings: *prompting* (§6.1) and (2) *fine-tuning* (§6.2), and conducted on the MMLU (Hendrycks et al., 2021), CommonsenseQA (Talmor et al., 2019) and ARC-Challenge (Clark et al., 2018), and GSM8K datasets.

6.1 Property-enhanced Prompting

Our prompting experiments are performed with Llama-3.1-8B-it (Dubey et al., 2024), Qwen2.5-7B-it (Qwen Team, 2024), and OpenAI o3-mini (OpenAI, 2025) focusing on three dimensions: communication, cognitive loads, and instruction. We exclude demonstrations, objectives, and external tools, as prior work extensively explored these properties. We begin with the zero-shot CoT

		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: Performance of models (%) on various tasks under different configurations. Arrows indicate changes relative to Zero-shot CoT.

prompt (Kojima et al., 2022) "Answer the following question step-by-step.". We then introduce the following modifications: (1) Add "Please" to promote Politeness; (2) "Reflect on your prior knowledge to gain a deeper understanding of the problem before solving it." to encourage Germane load; (3) "Self-verify your response thoroughly to ensure each reasoning step is correct." to promote Metacognition; (4) "You will be awarded 100 USD for every correct reasoning step." to improve the Rewards.

Findings. Our results in Table 2 reveal that different prompting properties influence models in varying ways, with their impact differing across tasks. Overall, most of the property combinations benefit Llama-3.1 but negatively impact other models. Moreover, we observe that combining multiple positive properties does not necessarily

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: Performance of two fine-tuned Qwen-2.5-7B-it models (%) on **polite data / non-polite data** under different settings.

yield stronger improvements; instead, a single property often proves most effective. Specifically, politeness yields the best results for Llama on the Comm.QA and ARC-C datasets, whereas metacognition achieves the highest performance for Qwen across all tasks. Regarding combining properties, while both politeness and germane load individually enhance Llama's performance on MMLU and ARC-C, combining them results in lower performance than politeness alone. A similar pattern is observed when combining metacognition with rewards for Llama on the CommQA dataset. Surprisingly, for the o3-mini model, we observe most properties result in negative effects. We hypothesize that this could be due to the model being excessively trained on chain-of-thought data, causing the properties to push the prompts out of distribution. Finally, we also note that in cases where we do not observe any improvement, this does not imply that these properties lack impact. Instead, more sophisticated or optimized prompting methods that better foster these properties may yield improvements. We leave these explorations for future research.

6.2 Property-enhanced Fine-tuning

To better understand how model-specific factors, particularly instruction tuning, affect the effectiveness of prompt properties, we conduct a targeted fine-tuning experiment on the Qwen-2.5-7B-It model. We choose it as it does not show better reasoning with more polite prompts. We fine-tune two variants of Qwen-2.5-7B-It using data either enriched with politeness or left in its original form. Specifically, we sample 2,500 examples from the Alpaca-GPT-4o dataset[6], and create two fine-tuning sets: one with "Please" added to each instruction, and one unchanged.

---

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

Findings. As shown in Table 3, firstly, fine-tuning Qwen-2.5-7B-It on polite prompts leads to notable performance gains when appending "Please" to the inputs. This suggests that instruction-tuning on data with explicit politeness markers enhances the model's sensitivity to polite prompt styles, enabling performance improvements that simple prompt-level politeness alone could not achieve (§6.1). Second, surprisingly, instruction-tuning with polite-enhanced data achieves better results compared to original data across almost all property-enhanced experiments. This suggests that incorporating politeness, or more broadly, certain properties, during instruction tuning can lead to more effective and robust reasoning models.

7 Conclusion

This paper explores natural language prompts and their impact on model performance through a novel property-based perspective. We survey over 150 prompting studies and introduce a taxonomy of 21 key properties for assessing prompt quality and their influence on model performance. Our analysis reveals an uneven emphasis on different properties across models and tasks, exposing significant research gaps in property-based prompt optimization. We further identify correlations among properties within a pool of good natural language prompts, leading to actionable prompting recommendations. In a reasoning task case study, we find that enhancing single prompt properties often outperforms multi-property combinations, and fine-tuning on these improves reasoning, challenging the assumption that combining properties always yields better results. As the field continues to evolve, we hope this work will inspire researchers to pursue deeper investigations into the relationships between prompt properties and model behaviors and advance prompt evaluation methods and their implications in diverse applications.

Limitations

Despite our best efforts to conduct a rigorous and comprehensive study, we acknowledge several limitations inherent to our methodology.

First, our study is constrained by the scope of the literature we survey. Due to limitations in human resources, we are unable to cover all relevant papers in the field. While we make diligent efforts to mitigate this by surveying a diverse set of publications from various conferences and topics, it is possible that some relevant studies are omitted. This may affect the comprehensiveness of our findings and, consequently, the conclusions we draw.

Second, our correlation property analysis is limited to a predefined set of properties. While these properties are carefully chosen to represent diverse and meaningful dimensions, analyzing alternative properties can produce different results. To address this, we ensure that the collected prompts are diverse and verified through human review. However, the inherent variability in property selection introduces potential limitations to the generalizability of our findings, and caution should be exercised when extrapolating these results to other contexts.

We also 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.

Finally, our multi-property prompt enhancement experiments are conducted using supplementary prompts in their simplest form, without optimization for specific models. While this approach establishes a foundational analysis, it may lead to suboptimal handling of certain properties and neglect the potential advantages of more refined prompts regarding these properties for individual models. This limitation affects the robustness of our findings and highlights the need for future research into prompt optimization techniques.

In summary, while we take significant steps to mitigate these limitations, they reflect the inherent challenges in conducting a study of this scope and complexity. We hope that our work serves as a foundation for further exploration and refinement in this area.

Ethical considerations

Our analysis could potentially be misused to optimize prompts for harmful purposes, such as generating misinformation, hate speech, or privacy violations. While our research is not intended for such applications, preventing all potential misuse is inherently challenging. Although our study may improve the effectiveness of adversarial applications and malicious actors, we do not expect it to be inherently more advantageous for harmful purposes than for positive applications. Lastly, we compensate our annotators at an hourly rate of $20, which exceeds the local minimum wage.

Acknowledgement

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 Meth-*

ods 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*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

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*.

Shangying Hua, Shuangci Jin, and Shengyi Jiang. 2024. The limitations and ethical considerations of chatgpt. *Data Intelligence*, 6(1):201–239.

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.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023a. Pretrain, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

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. 2023b. 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*.

Hanjia Lyu, Song Jiang, Hanqing Zeng, Yinglong Xia, Qifan Wang, Si Zhang, Ren Chen, Chris Leung, Jiajie Tang, and Jiebo Luo. 2024. LLM-rec: Personalized recommendation via prompting large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 583–612, Mexico City, Mexico. Association for Computational Linguistics.

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.

Bernd Carsten Stahl and Damian Eke. 2024. The ethics of chatgpt – exploring the ethical issues of an emerging technology. *International Journal of Information Management*, 74:102700.

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*.

Fan Yin, Jesse Vig, Philippe Laban, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2023. Did you read the instructions? rethinking the effectiveness of task definitions in instruction learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3063–3079, Toronto, Canada. Association for Computational Linguistics.

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*.
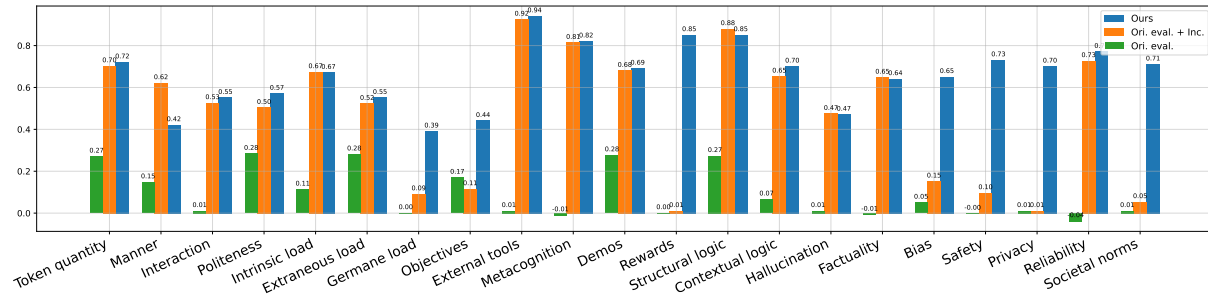
A Supplementary Results

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 Surveyed papers

Table 5: Table with Automatic Index Increasing

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 Prompt Gallery Dataset for Text-to-Image Generative Models (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., 2023b)	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 List of papers supporting properties in Table 1

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., 2023b; 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.

D Prompting for Dimension Evaluation

D.1 Communication Dimension Prompt Detail

```
COM_FORMAT = "{'Token quantity': 1-10, 'Manner': 1-10, 'Interaction': 1-10, 'Politeness': 1-10}"
COM_JUDGING_PROMPT = f"""You are a highly experienced judge tasked with evaluating a prompt on the following criteria.
The prompt for you to evaluate 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 and rate each criterion on a scale of 1-10:
- Token quantity: The extent to which prompts provide optimal and relevant information while minimizing token usage, balancing information completeness with efficiency.
- Manner: The degree to which prompt is clear and direct (across turns) while minimizing unnecessary ambiguity, complexity, and confusion.
- Interaction: The extent to which the prompts explicitly encourage the models to gather the necessary details and requirements by asking questions of clarification or confirmation.
- Politeness: The degree to which prompt maintains professional and context-specific politeness.
The scoring system is provided below:
> Token quantity
- 1-2 (Poor): The prompt is highly inefficient with token usage. It includes excessive, redundant details or is overly wordy without adding meaningful information. It either lacks critical information or includes irrelevant details, making it difficult for the model to understand or respond effectively.
- 3-4 (Below Average): The prompt is either too long or too short, with noticeable inefficiencies in token usage. It may include some unnecessary information or omit key details, reducing its effectiveness.
- 5-6 (Average): The prompt is moderately efficient in token usage but could be improved. It includes most necessary information but may have minor redundancies or omissions.
- 7-8 (Good): The prompt is efficient in token usage, providing a good balance between information completeness and conciseness. It includes all necessary details without significant redundancy.
- 9-10 (Excellent): The prompt is highly efficient in token usage, providing optimal and relevant information with minimal redundancy. It is concise yet comprehensive, enabling the model to respond effectively.
```

> **Manner**:
- 1-2 (Poor): The prompt is unclear, ambiguous, or overly complex, leading to significant confusion. It lacks directness and may require multiple interpretations.
- 3-4 (Below Average): The prompt has noticeable issues with clarity or directness. It may contain unnecessary complexity or ambiguity, making it harder for the model to understand.
- 5-6 (Average): The prompt is generally clear but could be more direct or simplified. It may have minor ambiguities or complexities that do not severely hinder understanding.
- 7-8 (Good): The prompt is clear and direct, with minimal ambiguity or complexity. It is easy for the model to understand and respond to.
- 9-10 (Excellent): The prompt is exceptionally clear, direct, and free of ambiguity or complexity. It is straightforward and easy for the model to interpret.
> **Interaction**:
- 1-2 (Poor): The prompt does not encourage interaction or clarification. It assumes all necessary information is provided and does not prompt the model to ask questions.
- 3-4 (Below Average): The prompt minimally encourages interaction but lacks explicit guidance for the model to ask clarifying or confirming questions.
- 5-6 (Average): The prompt somewhat encourages interaction but could be more explicit in guiding the model to ask questions or seek clarification.
- 7-8 (Good): The prompt effectively encourages interaction, explicitly guiding the model to ask clarifying or confirming questions when necessary.
- 9-10 (Excellent): The prompt excellently encourages interaction, clearly and explicitly prompting the model to gather all necessary details through questions or confirmation.
> **Politeness**:
- 1-2 (Poor): The prompt is unprofessional, impolite, or inappropriate for the context. It may use offensive or overly casual language.
- 3-4 (Below Average): The prompt lacks consistent politeness or professionalism. It may have moments of appropriateness but fails to maintain a respectful tone throughout.
- 5-6 (Average): The prompt is generally polite and professional but could be more consistent or context-specific in its tone.
- 7-8 (Good): The prompt maintains a professional and polite tone throughout, with minor room for improvement in context-specificity.
- 9-10 (Excellent): The prompt is exceptionally polite, professional, and context-specific. It maintains a respectful and appropriate tone at all times.
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>
<begin of ratings> {COM_FORMAT} <end of ratings> """

D.2 Cognition Dimension Prompt Detail

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:
<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:
- 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>
<begin of ratings> {COG_FORMAT} <end of ratings>

D.3 Instruction Dimension Prompt Detail

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:
<begin of explanation> ... <end of explanation>
<begin of ratings> {INS_FORMAT} <end of ratings> """

5870

D.4 Logic and Structure Dimension Prompt Detail

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>
<begin of ratings> {LOGIC_FORMAT} <end of ratings> """

D.5 Hallucination Dimension Prompt Detail

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>
<begin of ratings> {HALL_FORMAT} <end of ratings> """

D.6 Responsibility Dimension Prompt Detail

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>
<begin of ratings> {RES_FORMAT} <end of ratings> """

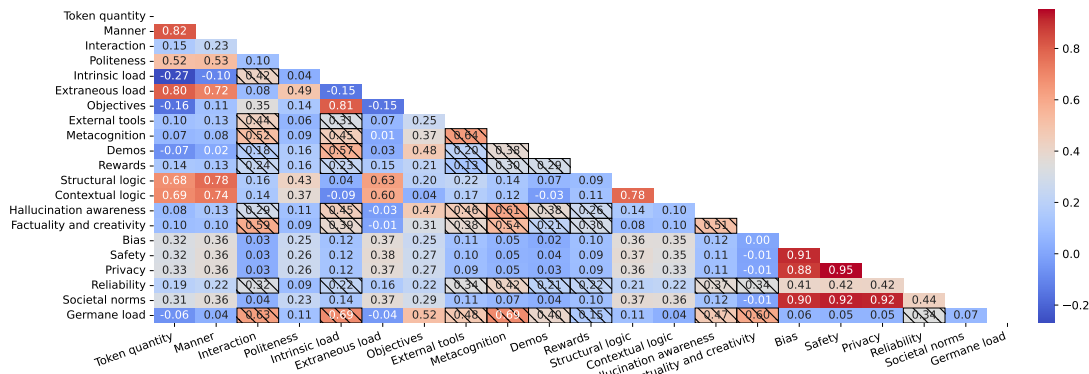# E Correlation results with findings from gemini-2.0-flash

Figure 3: Correlations of properties evaluated by gemini-2.0-flash. 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.

We observed that most of the strong correlations identified in our previous analysis remain consistent, including (token quantity; manner; structural logic; contextual logic; and extraneous load), (objectives; intrinsic load), (structural logic; contextual logic), and (safety; societal norms), with two correlations being slightly not as strong as before (now 0.6 by Gemini-2.0-flash versus 0.7 by GPT-4o): (hallucination awareness; factuality and creativity) and (objectives; germane load). These additional results further support the (almost) generalizability of the observed correlations across different high-performing LLMs, rather than being restricted to specific model groups (e.g., OpenAI models).