What Can Prompt Engineering Learn From Questionnaire Design?

A Path Towards a Framework

October 24, 2024

Lilian Huang
Elizabeth Dean
Brandon Sepulvado
Joshua Lerner
Ipek Bilgen
Leah Christian

What is prompt engineering?

- Crafting input text to interact with Large Language Models (LLMs)
- Goals of a prompt:
  - Optimize LLMs' performance on the desired task
  - Customizing output to meet specifications

Primary LLM use cases

Single-turn

- A single call to get one piece of data or instruct it to perform a specific task
- Analogous to a one-off survey question with no follow-up

Multi-turn conversation

- Back-and-forth conversation with multiple rounds of user inputs
  - "Chatbot" style
  - Providing extra information
  - Or asking LLM to iterate on and refine its output
- Analogous to a survey question with follow-up probes, or a qualitative interview

What might questionnaire design have to teach us when using LLMs?

- Once viewed as a black box process
  - Carried out by one or few people
  - Little time or resources for prior testing
- Progress has been made in demystifying it
  - Development of quality evaluation frameworks
- Can we achieve the same with prompt engineering for LLMs?

Why **not**?

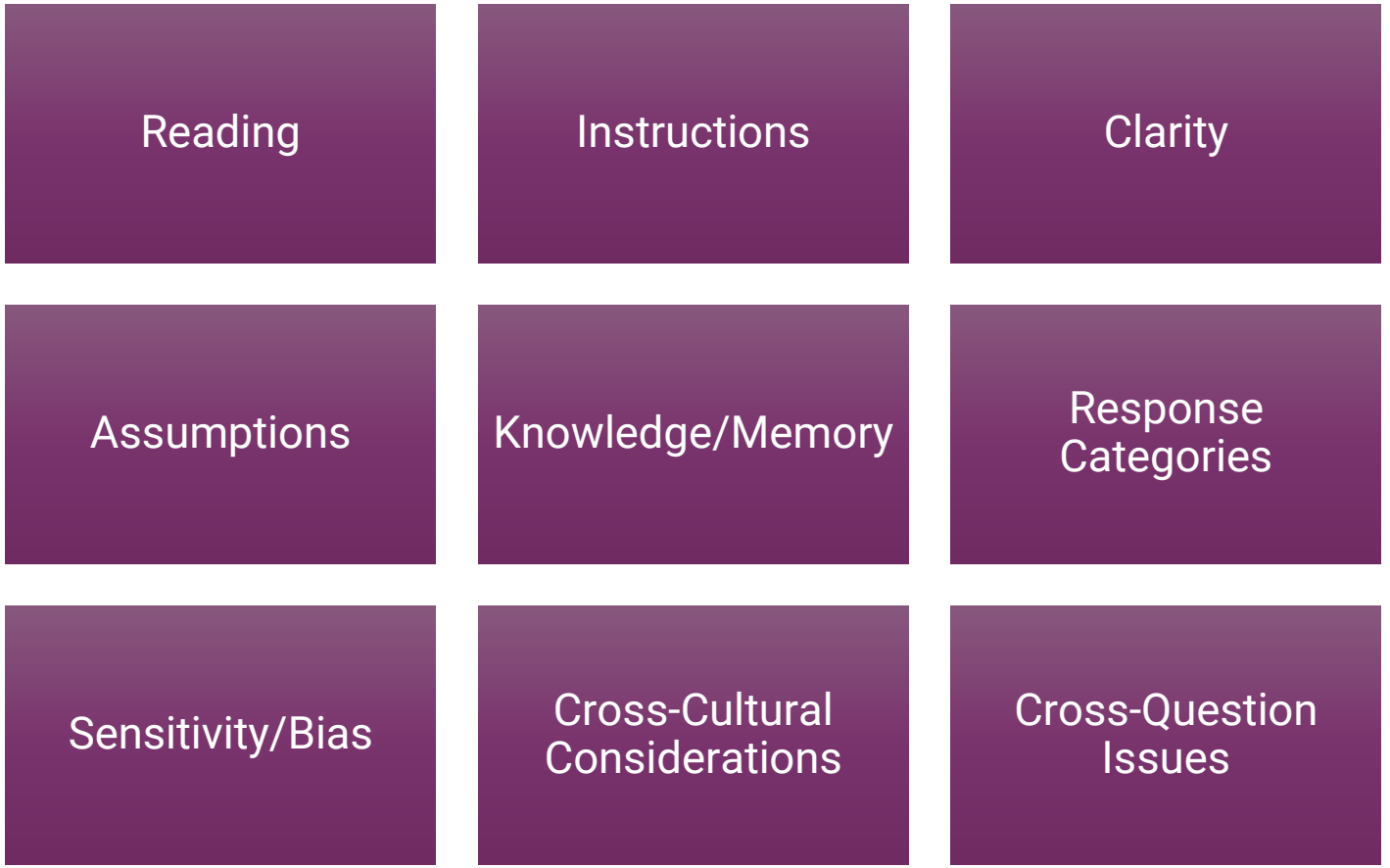
- LLMs are **not** survey respondents
  - Do not engage in human thought
    - Learn and generalize statistical relationships between textual tokens
  - No underlying construct validity
  - The barriers to optimal responses are different from the barriers humans face
    - e.g. LLMs do not experience response burden or social desirability effects
  - Survey researchers are **not** LLM users
    - Goal is not to elicit a specific response

Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., & Farajtabar, M. (2024). GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. *arXiv preprint arXiv:2410.05229*.

So why **should** we even try to apply survey research principles?

- Survey researchers are experts in nudging respondents to improve response **quality**
  - e.g. clarity, response length
  - Likewise, LLM users want to nudge features of LLM output text
- Prompt engineering is largely trial-and-error at present
  - e.g. manual revisions, using other LLMs to iterate on LLM prompts
  - A clear framework for prompt development can help us understand the impact of prompt variations on LLM output
  - Can enable us to conserve time and resources

Question Appraisal System (QAS)



Willis, G. B., & Lessler, J. T. (1999). Question appraisal system QAS-99. *National Cancer Institute.*
Dean, E., Caspar, R., McAvinchey, G., Reed, L., & Quiroz, R. (2007). Developing a low-cost technique for parallel cross-cultural instrument development: The question appraisal system (QAS-04). *International Journal of Social Research Methodology*, *10*(3), 227-241.

Can the components of this framework provide a map for how we approach prompt engineering?

QAS Component: Instructions

Survey Question

From the respondent's point of view, check that the introductions, instructions, or explanations are not:

- Conflicting
- Inaccurate
- Overly complicated

LLM Prompt

Make sure instructions are clear and accurate, e.g.:

- Be specific as to desired output formatting and length
- Give several concrete examples of desired output ***(few-shot prompting)***

QAS Component: Clarity

Survey Question

Check for potential problems communicating the question's intent to the respondent

- Awkward wording and syntax
- Undefined terminology
- Vagueness allowing for multiple interpretations

LLM Prompt

Black-box nature of LLMs makes it especially difficult to ensure clarity

- Ask LLM to explain its reasoning ***(chain-of-thought prompting)***
  - Give example of a thought-out response
  - Add "let's think step by step" to prompt
- Can improve results but, more importantly, gives users clarity on what LLM is doing

QAS Component: Assumptions

Survey Question	LLM Prompt
Are inappropriate assumptions made? • About survey respondent and their living situation • About whether they have certain experiences	Explicitly ask LLM to adopt a specific persona ***(role prompting)*** • Instruct it to “think like a…” or “You are a…” • For some models (e.g. OpenAI’s GPT models), this can be set in a separate system message • But can have unexpected results

QAS Component: Knowledge/Memory

Survey Question

Respondent may:

- Not be able to **recall** information on the spot
- Not have factual **knowledge** in the first place

LLM Prompt

LLMs have documented issues with memory retrieval (Chauvet, 2024)

- Stronger on recognition than recall
- Stronger on immediate retrieval than delayed retrieval
- Sometimes provide information not directly relevant to what is being requested
  - Pretrained information overrides specific context of that conversation

Chauvet, J. M. (2024). Memory GAPS: Would LLM pass the Tulving Test?. *arXiv preprint arXiv:2402.16505.*
Chauvet, J. M. (2024). Memory Traces: Are Transformers Tulving Machines?. *arXiv preprint arXiv:2404.08543*.

QAS Component: Knowledge/Memory (2)

Survey Question

Respondent may:

- Not be able to **recall** information on the spot
- Not have factual **knowledge** in the first place

LLM Prompt

LLMs have been known to hallucinate factually incorrect statements and nonexistent sources

- Need to ensure LLMs cite sources which are **real** and **support the claim being made**
  - Provide LLM with exact documents you want it to use
  - Emphasize the importance of citation, e.g. give a minimum number of required citations
  - May have to provide summaries or snippets instead of full documents

Gao, T., Yen, H., Yu, J., & Chen, D. (2023). Enabling large language models to generate text with citations. *arXiv preprint arXiv:2305.14627.*

QAS Component: Response Categories

Survey Question

Response categories should be:

- Mutually exclusive
- Exhaustive
- In a logical order
- Not overwhelming

LLM Prompt

LLMs should be explicitly provided with options to express uncertainty

- Especially important for avoiding hallucinations
- LLM should be allowed to say “I don’t know” or refuse to answer (Xu et al., 2024)
- LLM can be asked to state its level of confidence in its own response (Tian et al., 2023)

Xu, Z., Jain, S., & Kankanhalli, M. (2024). Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.
Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., ... & Manning, C. D. (2023). Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*.

QAS Component: Sensitivity/Bias

Survey Question

Need to ensure that question is not:

- Biased
- Asking about content of a sensitive nature (embarrassing, private, illegal)
- Incentivizing a socially desirable response

LLM Prompt

LLMs are not emotionally sensitive, but have encoded bias that may be triggered

- Prompting techniques to reduce bias:
  - ***Debiasing*** (Ganguli et al., 2023)
  - Prompting for "slow and thoughtful" or "step by step" responses (Kamruzzaman & Kim, 2024)
  - Instruction to avoid gender-specific pronouns, etc
  - Giving LLM explicit examples of bias so it can recognize them
  - ***Adversarial triggers*** (Venkit et al., 2023)

Ganguli, D., Askell, A., Schiefer, N., Liao, T. I., Lukošiūtė, K., Chen, A., ... & Kaplan, J. (2023). The capacity for moral self-correction in large language models. *arXiv preprint arXiv:2302.07459*.

Kamruzzaman, M., & Kim, G. L. (2024). Prompting techniques for reducing social bias in LLMs through system 1 and system 2 cognitive processes. *arXiv preprint arXiv:2404.17218*.

Venkit, P. N., Gautam, S., Panchanadikar, R., Huang, T. H. K., & Wilson, S. (2023). Nationality bias in text generation. *arXiv preprint arXiv:2302.02463*.

QAS Component: Cross-Cultural Considerations

Survey Question

Question should not contain **unsuitable assumptions** or **overly specific references** that would not be effectively understood across cultures

LLM Prompt

LLM output is often not culturally sensitive or tailored to user's background (Kharchenko et al., 2024)

- Even when output is customized, it is often according to stereotypes or hallucinations
- Due to lack of diversity in training data
- Possible countermeasures in prompting:
  - Write prompt in a different language rather than English
  - Ask LLM to cite sources to justify its cultural understanding

Kharchenko, J., Roosta, T., Chadha, A., & Shah, C. (2024). How well do LLMs represent values across cultures? Empirical analysis of LLM responses based on Hofstede cultural dimensions. *arXiv preprint arXiv:2406.14805*.

QAS Component: Cross-Question Issues

Survey Question

Look across the whole questionnaire to ensure:

- Questions are ordered in a logical way
- Terminology is consistent
- Response options are consistent
- Skip pattern is logical and comprehensive

LLM Prompt

- LLMs may struggle with following a multi-turn conversation
  - Be explicit when referring to something mentioned earlier, or else LLM may infer incorrectly (Sun et al., 2024)
- Even a one-off LLM prompt should be structured in a way that makes relative importance and relevance clear
  - If there is a long preamble introducing the context, place the actual question at the very end (Sun et al., 2024)

Sun, Y., Liu, C., Zhou, K., Huang, J., Song, R., Zhao, W. X., ... & Gai, K. (2024, August). Parrot: Enhancing Multi-Turn Instruction Following for Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 9729-9750).
Sun, S., Zhuang, S., Wang, S., & Zuccon, G. (2024). An Investigation of Prompt Variations for Zero-shot LLM-based Rankers. *arXiv preprint arXiv:2406.14117*.

Conclusion: Towards a framework

- Guiding principles are needed when developing LLM prompts
  - Not all elements of questionnaire design are directly applicable to LLM prompts
  - But there is enough in common to serve as the basis for an overarching framework
- Especially useful for junior analysts/researchers
  - Increases accessibility
  - Helps them internalize overall principles of survey research

A final reminder…

- **LLMs are not survey respondents**
  - The aim of this prompt engineering process is to nudge the features of output text
  - We are not interacting with human thought processes
- LLMs have many ingrained limitations; there is only so much we can do to address them **at the prompting stage**
- Even when using LLMs to support our work, we should **think like actual researchers**

Thank you.

Lilian Huang
Statistician at NORC
huang-lilian@norc.org

Research You Can Trust™

NORC at the University of Chicago

Questions?