The Need for Vizier

April 2025

1 Mini-Lit Review

https://notebooklm.google.com/notebook/939a7ebd-89ea-40dd-8c62-134f047f09d6

2 Goal

Mini lit review showcasing that current SOTA LLMs fail to actually capture the "meat" of most technical papers (insights, etc.) and focus on basically chucking in the abstract + methodology on 0-shot prompting.

3 Introduction

Modern LLMs have demonstrated incredible capabilities in a wide variety of NLP tasks including the extraction and summarization of information from textual data. From needle in a haystack problems to generating structured summaries, user have found much success applying these models in both academic and industrial settings. A common and straightforward approach to leverage the utility offered by LLMs is to use zero-shot prompting, whereby one instructs the LLM to perform a task based purely on its pre-trained knowledge. This approach is particularly appealing due to the simple user interaction and lack of need for task specfic examples. Despite modern advancements in the pre-training of large scale models, a critical limitation arises when we try to apply zero-shot prompting to the comprehension of long, complex technical documents (i.e research papers).

While LLMs can indeed process and generate text that is relevant to the document without fail, typically covering key points mentioned in the abstract and the paper's relevance from the introduction, they often struggle to achieve in-depth, genuine, and useful insights presented in the documents that extend beyond surface-level summaries. Moreover, technical papers require a nuanced grasp of domain-specific knowledge and reasoning to interpret results and understand the significance of the findings in a broader, relevant context. This mini-literature review will focus on understanding the limitations of zero-shot prompting and showing that LLMs indeed have the capability required to comprehend and articulate the "meat" of the content effectively, but are constrained by the particular user interaction.

4 Limitations of Zero-Shot Prompting

In the goal of presenting a technical paper or any paper for that matter properly, one of the first things we look for in the report is accuracy. If a paper delves into a highly specialized area of research (almost niche) that is not well represented in the LLMs training corpus, the model capacity to generate the relevant information and insights through a simple zero-shot prompting technique can be constrained. This is because the LLM tends to rely on heiristics and techniques that were effective for understanding papers it had seen during pre training, but these same strategies may not transfer well or apply effectively to underrepresented domains. Furthermore, the quality of the response in contegent upon the precise wording of the prompt. A poorly worded prompt can easily lead to outputs that are factually incorrect and/or irrelevant to what the user is looking for. Due to the precise language employed in technical areas, this problem is further exacerbated.

> "As shown by our §5 results, LLMs face notable challenges that pervade the computa-tional social sciences. The first challenge comes from the subtle and non-conventional language of expert taxonomies. Expert taxonomies contain technical terms like the dialect feature copula omission (§3.1.1), plus specialized or nonstandard definitions of colloquial terms, like the persuasive scarcity strategy (§3.1.8), or white grievance in im-plicit hate (§3.1.4). LLMs may lack sufficient representations for such technical terms, as they may be absent from the pretraining data (Yao et al. 2021). How to teach LLMs to understand these social constructs deserves further technical attention. This is especially true for novel theoretical constructs that social scientists may wish to define and study in collaboration with LLMs" (Can Large Language Models Transform Computational Social Science?)

> "Zero-shot learning often struggles with generalization, especially for domain-specific or low-resource languages, leading to lower accuracy and unintended biases. Few-shot learning, while mitigating some of these issues, remains sensitive to prompt design and data selection, affecting model reliability and consistency" (Challenges and limitations of zero shot and few shot learning)

The meaning of certain findings or statements in a technical paper can be heavily dependent on the surrounding context, which zero-shot prompts might not adequately capture. The limited context window and the inability of basic zero-shot prompting to maintain and utilize long-range dependencies within a technical paper can hinder the capture of nuanced meanings. Technical papers often build arguments and present evidence across multiple sections. Zero-shot prompting, typically focusing on a limited portion of the text, might fail to grasp how different parts of the paper relate to and influence each other, leading to a fragmented understanding

> "Moreover, we find that models struggle to identify the salient information and are more error-prone when summarizing over longer textual contexts." (Evaluating large language models on medical evidence summarization)

> "Multi-document summarization presents a multifaceted challenge in the realm of natural language processing, requiring a sophisticated approach to distill relevant information from numerous source documents. The intricacy lies in the necessity of intelligently fusing diverse content, accommodating varied perspectives, and handling potential redundancies. One of the primary challenges involves determining the relevance of each document, factoring in considerations such as doc-ument length, thematic alignment, and the presence of critical terms. The summarization system must effectively weigh the significance of each document to create a cohesive summary that avoids repetition and provides a comprehensive overview." (Survey of text summarization: Techniques, evaluation and challenges)

Bias represents another significant challenge. LLMs are trained on large datasets of text and code, which may inadvertently contain societal biases. Consequently, zero-shot prompting can sometimes lead to outputs that amplify these biases, potentially resulting in skewed responses. In the context of analyzing technical papers, this could lead to a tendency to favor certain perspectives, methodologies, or research outcomes that were more prevalent in the training data, potentially overlooking and/or misinterpreting findings that deviate from these viewpoints.

> "Researchers should weigh the benefits of applying prompting methods to CSS, along with the limitations and risks of doing so. Most notably, LLMs are known to amplify social biases and stereotypes (Sheng et al. 2021; Abid, Farooqi, and Zou 2021; Borchers et al. 2022; Lucy and Bamman 2021; Shaikh et al. 2022), as well as viewpoint biases in subjective domains (Santurkar et al. 2023b). These biases can emerge in open-ended generation tasks like the explanation and paraphrasing (Dhamala et al. 2021). The performance of LLMs as tools for classification and parsing may vary systematically as a function of demographic variation in the target population (Zhao et al. 2018)" (Can Large Language Models Transform Computational Social Science?)

Real-world, useful reports leverage domain specific knowledge in accurately interpreting and presenting technical papers. These papers essentially embed key insights into dense mathematical formulations (minimal explanations due to assumed familiarity), nuanced methodologies, and specialized language. Without familiarity, even a capable LLM may fail to discern the significance of results and/or misinterpret terminology. As demonstrated in fields such as law, medicine, and finance, summarizing documents effectively requires a deep grasp of domain context and historical precedent. This lack of embedded expertise often results in semantic drift, loss of critical detail, or biased interpretations favoring familiar but irrelevant concepts from pretraining.

> "Issues such as the loss of crucial information, semantic drift in longer summaries, domain-specific knowledge re-quirements, and the intricacies of handling multi-document summa-rization pose ongoing hurdles" (survey of text summarization: Techniques, evaluation and challenges)

> "In the medical field, summarizing research papers and clinical reports requires an in-depth understanding of intricate medical concepts, treatment modalities, and the ever-evolving landscape of healthcare. For example, a summarization system must grapple with the complexities of drug interactions, disease mechanisms, and the latest medical advancements to produce summaries that accurately convey critical information" (survey of text summarization: Techniques, evaluation and challenges)

> "Legal document summarization encounters a distinct set of challenges, demanding expertise in legal jargon, case law, and statutory intricacies. The summarization of legal briefs, court decisions, or con-tracts necessitates a keen awareness of legal precedents and the historical context of cases. A system without the requisite domain-specific knowledge might struggle to navigate the complexities of legal language, risking misinterpretations that could have significant consequences. In the financial domain, summarizing reports demands a comprehensive understanding of financial terms, accounting principles, and industry-specific metrics. Summarization tools should adeptly interpret balance sheets, income statements, and cash flow reports to distill key financial insights accurately. The absence of financial domain expertise may lead to oversights in crucial financial data, diminishing" (survey of text summarization: Techniques, evaluation and challenges)