Designing LLM-Based Support for Homelessness Caseworkers

Whitney Nelson[1], **Victor Wang**[1, 2], **Eunsol Choi**[2], **Min Kyung Lee**[1]

[1]School of Information, The University of Texas at Austin
[2]Department of Computer Science, The University of Texas at Austin
whitney.nelson@utexas.edu

Abstract

Our research explores the use of Large Language Models (LLMs) to assist social service delivery for people experiencing homelessness. We follow a human-centered design approach and work with caseworkers who provide homelessness-intensive case management. They identified summarization of case notes as an opportunity to better understand their clients. We then asked caseworkers to generate summaries in order to understand what information summaries need to contain. However, instead of resulting in "gold standard" summaries that can be used to train LLMs, we found that there were diversities in summaries that they wanted to see depending on their case management philosophies and roles. (We report our ongoing exploration of LLM summaries to support the diversities.) We also discovered that enabling workers to verify the summary is a key issue. We share implications that summaries should be able to support caseworkers' diverse goals, which requires future work on prompting engineering, evaluative metrics, and summary verification mechanisms.

Introduction

The integration of Large Language Models (LLMs) has opened new avenues for automating complex tasks like text summarization. However, the application of these general-purpose models to specialized domains presents unique challenges. While the general perception of LLMs is often limited to their applications in commercial and research settings, their potential in the public sector, particularly in social work, is both significant and under explored.

One of the key applications of LLMs in social work is in the domain of text summarization. Caseworkers, who often find themselves inundated with extensive case notes and records, can leverage the advanced summarization capabilities of LLMs to streamline their workflow. The ability of these models to condense large volumes of text into concise, coherent summaries presents a unique opportunity to enhance the efficiency and effectiveness of casework. This is particularly relevant in scenarios where caseworkers need to quickly assimilate and act upon critical information from various sources.

This paper investigates the potential of LLMs to generate domain-specific summaries for caseworkers. Our research question delves into the practicality and methodology of using LLMs for this purpose: How can general-purpose LLMs be utilized to create domain-specific summaries for caseworkers through prompt engineering? In addressing this question, the study explores the intersection of LLM capabilities with the nuanced demands of casework. It considers how prompt engineering can be employed to guide these models in producing summaries that are not only accurate but also contextually relevant to the varying roles and philosophical approaches inherent in casework. The paper seeks to understand the complexities of adapting a general-purpose LLM to the specific needs of caseworkers, who require reliable and precise summaries to make informed decisions in high-stakes environments. Through this exploration, we aim to contribute to the broader discourse on the application of human-centered AI technologies in specialized fields.

Related Work

AI for Public Sectors

In the public sector, the integration of AI and automation is becoming increasingly vital, particularly in high-stakes domains such as legal, child welfare, and case management. These technologies are transforming public services by automating tasks like text analysis and enhancing decision-making processes through predictive analytics (Kanapala, Pal, and Pamula 2019; Saxena et al. 2023). This is instrumental in sectors where professionals often face overwhelming workloads.

Casework is one such role in the public sector (Kuo et al. 2023). However, aligning these AI models with the actual needs and values of users in the public sector remains a challenge (Saxena et al. 2023). It is critical to utilize user-centric AI design in the public sector, ensuring that these technologies not only advance in technical capabilities but also align with the values and needs of those in high-stakes, impactful domains (Barale 2022). This approach underscores the potential of AI not just to streamline administrative processes in the public sector, but also to support more informed and empathetic decision-making, thereby contributing to the overall enhancement of public services.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Prompt Engineering and HCI

Prompt engineering is a new way to interact with LLMs through natural language. Users control the LLM output by using natural language as prompts to guide the model. While prompt engineering makes LLMs accessible to non-technical users, users still struggle with how to best interact with LLMs (Zamfirescu-Pereira et al. 2023b). Best practices and strategies for prompt engineering are still being developed. The main challenge is that prompts are “fickle” (Zamfirescu-Pereira et al. 2023a). It is difficult to determine what specific parts of the prompt influence the output.

To address the fickle nature of prompts, research focuses on interfaces for guiding users, persona based approaches to prompt engineering, chatbot development, and creating design methods to evaluate and test prompts. Interfaces for prompt guidance aim to assist non-technical users to generate the best possible output from the LLM (Mishra et al. 2023). Many interfaces use visuals to compare different prompts and model outputs, or provide templates for users (Arawjo et al. 2023; Dang et al. 2022b). However, there is still uncertainty about what parts of the prompts matter most to the model output (Yin et al. 2023; Dang et al. 2023).

While the fickle nature of prompts remains a challenge, broad guidelines for prompts have emerged. These include few-shot, zero-shot, instruction based, and persona based prompts. Few-shot involves providing one or more examples of the desired output within the prompt for the model to generalize from. In comparison to zero-shot prompting, which does not provide input or output examples to the model, few-shot is used in domain-specific use cases, such as summarizing medical records (Chuang et al. 2023; Ma et al. 2023). It is considered a best practice for knowledge representation for domain-specific information. Persona based prompts are another approach, which involves assigning the model a particular persona.

Prompt Engineering and Summary

Recent research in text summarization using LLMs has showcased advancements, particularly in addressing domain-specific challenges. Studies like (Ma et al. 2023) leverage dynamic prompts in summarizing legal documents, revealing the need for nuanced, context-aware approaches in LLM applications. This is echoed in software engineering research evaluating ChatGPT’s efficacy in code summarization, which identified its limitations in generating precise summaries. (Sun et al. 2023). These insights underline the importance of developing domain-specific summarization methods.

AI-powered tools have also been developed for tasks like real-time text editing utilizing LLMs (Dang et al. 2022a). These tools aid in structuring and revising text, providing writers with an external perspective for critical reflection and restructuring. Similarly, the NEWTS dataset introduces topic-focused summarization, pushing the boundaries of general-purpose models (Bahrainian, Feucht, and Eickhoff 2022). The FROST model further innovates by using entity chains for content planning, enhancing summary specificity and controlling hallucinations (Narayan et al. 2021). Likewise, the SPeC framework’s soft prompts mitigate performance variability in clinical note summarization, ensuring consistent and reliable outputs (Chuang et al. 2023). Collectively, these advancements demonstrate the transformative potential of LLMs in text summarization, underscoring the ongoing need for innovation in prompt engineering and model adaptation for diverse applications.

Formative Studies with Caseworkers

Method

The study was performed using a combination of in-depth interviews and interactive summary generation activities with users in order to understand caseworker’s information needs and potential challenges for automatic text summarization.

Organization: Homelessness Intensive Case Management We worked with a government-led nonprofit organization that has been serving people experiencing homelessness to achieve long-term stability through an Integrated Case Management (ICM) program over the past two decades. They provide services aimed at long-term living and housing stability, such as creating housing search plans, counseling services, identifying appropriate programs or treatments for medical health needs, etc. ICM clients have dedicated caseworkers with a case management plan. Our prior research interviews (Slota et al. 2023) as part of the bigger engagement with the organization revealed an ongoing expansion of responsibilities beyond what is strictly necessary for ICM, such as participation in emergency response, coordination of transportation for cold weather shelters, protective lodges, and serving as guides to the overall system of services available to people on the homelessness continuum. They also provide short-term, on-demand aid and advice to other homeless individuals as walk-in services, also known as triage. These walk-in services include activities such as holding mail, obtaining bus tickets, assistance with obtaining documents, etc. Walk-in clients are served based on the immediate needs of the client and have a limited meeting time of about 30 minutes. Each caseworker deals with as many as 50-60 clients per day.

The caseworkers at the organization primarily collect data in two forms: they write detailed free-text notes that describe the caseworker’s interaction with a client and assign labels from a predefined list of data labels to characterize the outputs and outcomes of the interaction for funding and performance reports. Caseworkers first write the case notes during or after a client interaction and then assign relevant data labels.

Case Notes Case notes are free-text records of caseworker and client-related interactions, written during or after such interactions. They assist caseworkers in tracking clients’ case history, informing their decisions on the next steps for client service. Caseworkers consult past case notes before or during client meetings to access pertinent information such as contact details, the clients’ last known location, and pending applications for housing programs.

The dataset shared by the organization contains case notes that span from October 2016 to September 2022. There are a total of 63,485 case notes across 1,691 clients written by 30 caseworkers. These notes vary in length, ranging from a few words to 130 sentences. The content includes contemporary functions of documentation such as assessment and planning, service delivery, and continuity and coordination of services (Reamer 2005). They chronicle events related to clients, encompassing not only direct interactions between clients and caseworkers but also other interactions such as email exchanges with service providers, records of received mail, phone calls, etc. Both Intensive Case Management (ICM) and walk-in (triage) clients have case notes, though the nature of these notes may slightly differ. Walk-in clients' case notes often focus on immediate actions and services rendered, while ICM clients' notes may cover longer-term plans and more personal details. We conducted a thematic analysis (Braun and Clarke 2006) of case notes aimed to understand the data collected and their implications for case management practices. Case note content can be represented in six major themes: 1) action items such as updating an application, renewal, or housing options; 2) client status updates such as job, application fill up, or caseworker task updates such as received mail; 3) requests for passes/-cards/services; 4) scheduling and tracking meetings or appointments; 5) emotional/general conversation snippets; 6) potential next steps.

Interviews In order to understand the challenges and the information needs of caseworkers at the organization, we conducted 60-minute semi-structured interviews with caseworkers. Our prior collaboration with the organization established familiarity and trust in our research team for the participants. The interviews focused on topics such as the potential purpose and utility of a summary of case notes. We explored how caseworkers use case notes to guide their interactions with clients and walked through their information gathering workflow when preparing to meet with a client.

During the second part of the interviews, we shared a sample of case notes and had each participant share out loud what information would be useful for a summary and why. Participants each reviewed about five to ten case notes per client and reviewed notes for up to three clients depending on time.

Summary Goal Feedback To better understand case note summary content and format, five interview participants completed a review of summary goals and a summary writing session. Prior to the session, participants were provided with five summary goals based on the information needs described in the interviews. Caseworkers provided feedback on the goals including if the goal made sense, what specific information they would want to see associated with each goal and why, and when they would use the information from each summary. The summary goals provided were: 1) inform the interaction with a client to inform the next steps, 2) to use the knowledge when you are interacting with a client for the first time, 3) help understand the longitudinal patterns of the client that can promote self-reflection on your own practices, 4) help assess the outcomes/success of case management over time, and 5) help to have better documentation of the kinds of provided services.

Summary Writing Activity Caseworkers then participated in two summary writing activities, the first with two of the same clients and the second with one of their own clients. Caseworkers chose which goals to write their summary for based on their personal needs and workflow. All participating caseworkers chose Goal 1, while no participants chose Goal 3. Table 1 shows which goals were chosen by which participants. For the first summary writing activity, caseworkers were given the same two clients and generated summaries for both sets of case notes. The participants read through the case notes and highlighted text they wanted included in the summary. Caseworkers generated summaries for one of their own clients for the second summary writing activity. The summary formats included a dated or bulleted list and free form paragraphs. Participants also reviewed the generated summaries from other caseworkers to compare content and format. Each participant mentioned they preferred the bulleted or dated list format, even if they wrote their summary in paragraph format.

Findings

General Attitude Towards Automatic Summaries Overall there were mixed responses to potential automation from caseworkers. However, almost all caseworkers interviewed acknowledged that a summary of case notes would be helpful for better serving their clients. Key themes related to the benefits of an automatic summary tool were saving time and providing more targeted services to clients, especially for walk-in triage clients. Concerns included the potential for the summary to miss important information and the summary biasing the caseworker by removing the human focused element of casework.

Benefits – Saving Time and Providing Targeted Services: Improving time constraints was heavily emphasized by caseworkers as a potential benefit of automatic summarization, especially for triage clients. Casework is split into two categories, triage and intense case management (ICM). Triage case management provides services to walk-in clients. The goal of triage is to provide the most immediate service such as mental health help or shelter referrals. ICM are clients that are assigned a caseworker. The caseworker works closely with the client to establish long term, stable, housing and achieve other goals, such as medical and mental health. All but one (P5) caseworker is responsible for both types of casework, with (P5) fully dedicated to triage.

Caseworkers have an ICM caseload of about 20-30 clients at a time, and may see up to 60 clients in one day through triage. Many triage clients are waitlisted for ICM and have several months, or even years, worth of case note history. Triage clients may see a different caseworker at each visit. Due to this heavy caseload and client inconsistency, caseworkers do not have time to read through each case note for triage clients. All caseworkers we interviewed stated that a summary would be helpful in better serving triage clients, as they do not have time to thoroughly read through more than 1-5 of the most recent notes.

Goal 1	Inform the interaction with a client to inform the next steps	P1, P2, P3, P4, P5
Goal 2	To use the knowledge when you are interacting with a client for the first time	P1, P3, P5
Goal 3	Help understand the longitudinal patterns of the client that can promote self-reflection on your own practices	None
Goal 4	Help assess the outcomes/success of case management over time	P1
Goal 5	Help to have better documentation of the kinds of provided services	P3, P4

Table 1: Participants' chosen goals for summary writing activity

Providing more targeted services also emerged as a potential benefit of an automatic summary tool. Caseworkers also heavily emphasized its utility for triage clients. The goal of triage is to assist the walk-in client with the most urgent and immediate need. These are typically mental health, substance abuse, or temporary housing services. Due to the high volume of triage clients each day, caseworkers have limited time with each client and need to make quick judgment calls. Caseworkers described spending a few minutes reading through at most five case notes in order to understand the client's history. Caseworkers stated that a summary could assist in providing more targeted services by giving a more holistic view of the client's needs. One caseworker gave the example of not recommending services that have already been referred and were ineffective (P4). For example, if a client expresses interest in substance abuse rehabilitation a summary can ensure the caseworker does not refer the client to a service that has already been attempted unsuccessfully.

Concerns – Trust and Bias: While the caseworkers could imagine the benefits of an automatic summary tool, they also shared concerns related to trust and bias. Ensuring the summary is accurate and relevant was expressed by each of the caseworkers as a major concern. Caseworkers stated that they would worry that the summary may miss something important potentially leading to a decline in the quality of service (P1). Caseworker service recommendations are high stakes because they affect basic needs, such as housing and health. If a summary is not accurate or missing relevant information, a client could miss out on important services that greatly affect their quality of life. Caseworkers expressed they may not trust the summary and end up going back through the notes anyway. Therefore an automatic summary must consider how to show accuracy and relevance, potentially by referencing case notes directly in the summary.

Another key concern was the potential for caseworkers to over rely on the summary therefore removing the human focus of casework. Caseworkers expressed worry that the summary could bias the interaction with a client. One caseworker emphasized that she doesn't like to rely on the notes but instead focuses on the client's needs that day (P1). She described that it was unfair to judge a client on something that happened months ago that may no longer be an issue. Another caseworker described that providing too much information defeats the purpose of a summary and could cause the caseworker to focus on something that is not important or look back through the notes anyway (P2). What and how much information is provided by summary is crucial to limiting caseworker bias.

Diversities in "Gold Standard" Summaries We observed that there was a variance in the gold standard summaries that caseworkers wanted to use.

Philosophical Diversity: Caseworkers have diverse philosophical approaches to case management, which can significantly impact how caseworkers engage with clients and manage their cases. A caseworker's role and experience can affect their approach to information gathering and the amount of context they seek out about a client. These differences can lead to variations in how caseworkers assess, plan, and intervene with their clients. An effective data intervention must balance the needs of different case management approaches.

The interview participants brought up their approaches to case management while discussing what they would like to see in a summary solution and when describing their workflow. (P5) mentioned that she wants to read as much as possible about a client's history so the client does not have to repeat their story. Many of their clients have a trauma and it can be triggering for a client to repeat it during each visit. This approach is different from (P1) who emphasized focusing on how the client is presenting that day. She described that emphasizing a client's history can bias how a caseworker interacts with the client, when that information may not be relevant to the client's current situation. These different approaches can result in variations in how each caseworker conceptualizes and addresses the root causes of clients' challenges.

The different approaches also affect how caseworkers record their notes and what information is included. (P4) discussed how her background in medical case management affected the level of detail she would record in her notes. She would record as much information as possible. After learning best practices for homeless case management, she now records less information but acknowledges that her notes tend to be more detailed. This approach to case note writing is different from (P1) and (P2), who described focusing on only the most important information, such as housing status, mental and medical diagnoses, upcoming appointments, and safety issues. They stated that their case notes can include additional details about the client's behavior and the interaction, but emphasized limiting their opinion as much as possible.

Role and Responsibility Diversity: There is also variance in a caseworker's role. A caseworker interacts with clients on a walk-in basis, triage, and with a set number of clients more closely, intensive case management (ICM). Intensive case management is a client-centered approach in social services and healthcare that entails a comprehensive and ongoing support system. It involves a dedicated caseworker who works closely with individuals, developing personalized care plans, and coordinating a range of services to address complex needs. The primary goal is to enhance the individual's well-being and self-sufficiency through intensive, long-term support and collaboration.

Triage case management, on the other hand, is a system designed for efficient and rapid assessment and prioritization of cases. Caseworkers quickly evaluate cases to determine their urgency and severity. Triage case management is particularly valuable in situations where there is a high volume of cases or limited resources. All but one, (P5), of our interview participants are responsible for both triage and ICM clients. The workflow for gathering information is different for each type of case management. (P2) discussed when in triage you can see up to 60 clients in one day and you do not have time to go back in a client's history. Therefore, the emphasis during triage is the most recent notes. (P4) described that she will read the last one to three notes depending on the client when in triage, but will go back much further when working with a client in ICM.

There was agreement among the participants for the usefulness of a summary and the included content for clients in triage. Summaries for triage clients should include any safety concerns, mental health diagnoses, housing status, upcoming appointments, and referrals. This information is useful because it identifies if there are any safety concerns a caseworker needs to be aware of and directs the caseworker to immediate needs that can be addressed within one meeting. There was less agreement for what should be included in the summary for ICM clients. (P2) mentioned that the caseworkers should know their ICM client's history intimately and be very familiar with their goals and services. This was echoed by the other interview participants. The participants still believed a summary for ICM could be helpful, but were less sure what information to include.

Exploration of LLM summary for caseworkers

Method

To test the design of LLM generated summaries we followed two research stages. Both stages were informed by interviews with caseworkers and their feedback on what would be useful in a summary. The first stage defined the specific purposes of the summary based on caseworker goals and the creation of a human annotation guide. The second stage included prompt engineering and iterating on summary requirements. The second stage of research is ongoing.

Formalizing Caseworker Goals as Prompts to Support Diverse Use Cases To generate effective prompts, we created a human annotation guide based on the four summary goals identified in previous study. The purpose of the guide was to provide a comprehensive overview of the goal of the summaries, the necessary background information on the organization and the role of a caseworker, and what should be included in the output of the summary. The annotation guide was used to create the sample summaries and draft the prompts. This document was foundational to ensuring human annotators had consistent expectations for the purpose of the summary, what should be included in the summary output, and informing iterations of prompts

For summary prompt experimentation, the number of goals was reduced to Goal 1 and Goal 2. These goals were chosen because they were identified as the most useful by the previous interview participants. The purpose of Goal 1 is to identify next steps in the progression towards stable housing, and the purpose of Goal 2 is to initiate a productive first meeting with a client.

Goal Summary Prompt Engineering To produce the summaries, we utilized prompt engineering without additional technical layers or fine tuning to the models. Literature informing prompt experimentation includes expert systems and knowledge representation along with current papers exploring prompt engineering and summarization.

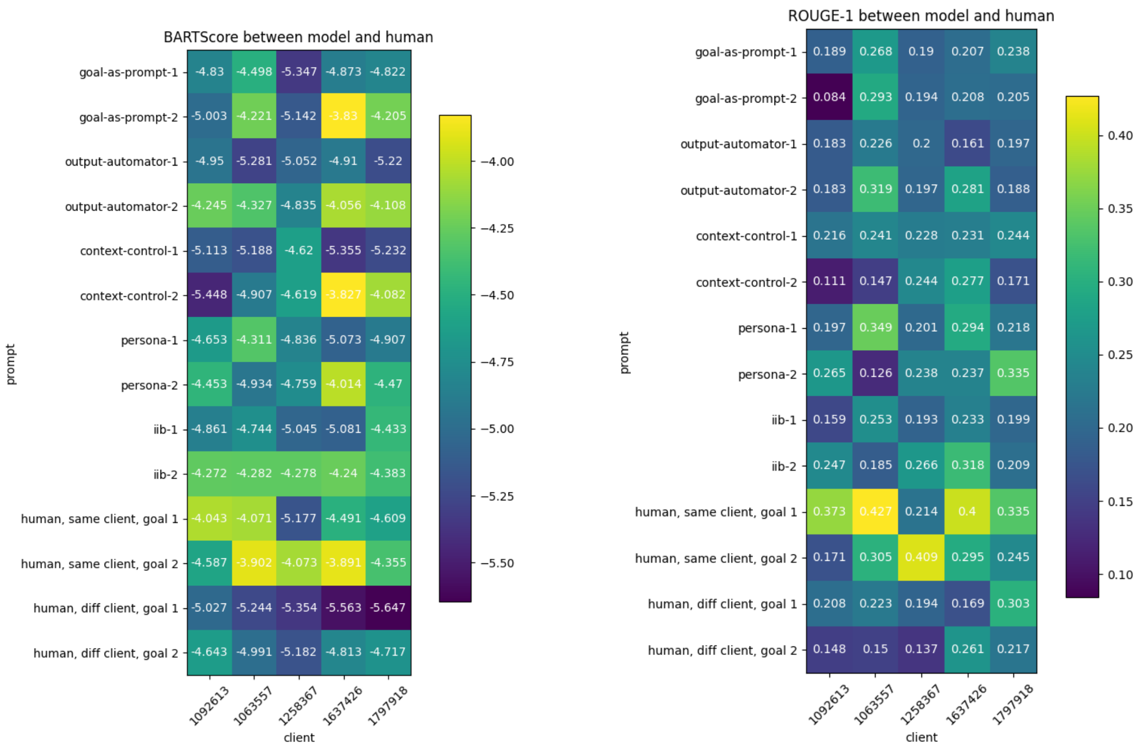
Five sets of prompts were tested based on the human annotation guide and current literature on prompt engineering. Output Automator prompts specifically stated what should be in the model output (e.g. "The output should contain the recommended next action for the caseworker."). Context Control prompts included statements about what the model should consider and not consider (e.g. "Only consider open or incomplete tasks and ignore information about client behavior."). Persona based prompts asked the model to take on the persona of a caseworker and define the outputs based on certain personality characteristics. Identity, Intent, Behavior prompts further expanded on Persona Based prompts by telling the model to take on a identity with specific intents and behaviors. Lastly, Goal as Prompt directly input the goal and requirements from the human annotation guide into the model. (White et al. 2023; Kumar et al. 2022)

Model The model we used to generate summaries is Llama 2 13B chat-fine-tuned (Touvron et al. 2023). We chose this model because it was the best-performing open-source model that our compute could handle. In preliminary experiments, we also considered models like Mistral and Alpaca (Jiang et al. 2023; Taori et al. 2023), but found them to do poorly compared to the Llama 2 model. We use the publicly released model with no additional fine-tuning. We decode with top_p=0.1, temperature=0.1.

Hierarchical Summarization For clients with especially many case notes, the case notes may not all fit within the model context size. In such cases, we used a simple sliding window mechanism to generate subsummaries for sections of consecutive case notes. We used the minimum number of sections such that each section fit into the context size (after accounting for prompt and response length) and the sections covered all the case notes.

Evaluation To evaluate the similarity between two summaries, we use BARTScore and ROUGE. BARTScore measures the predictability of a summary X given a summary Y. ROUGE computes lexical similarity. All scores reported are F1 (harmonic mean of scores in each direction). A higher score (less negative, in the case of BARTScore) means higher similarity.

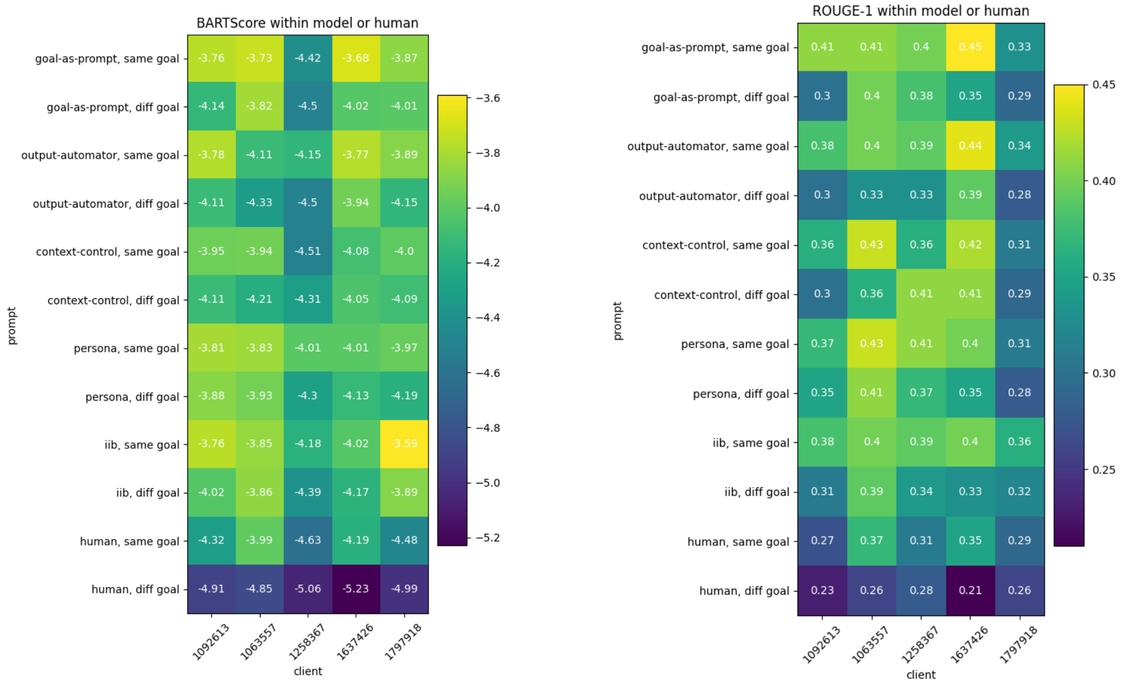
We obtained human summaries for 5 clients annotated by



(a)

(b)

Figure 1: Scores between model and human summaries on 5 prompts over 2 goals. Last 2 rows are a lower baseline; second last 2 rows are an upper baseline.



(a)

(b)

Figure 2: Scores with respect to goal (within model or human summaries) on 5 prompts or 2 humans.

2 humans[1]. The clients to be annotated and evaluated were chosen randomly among clients with 10-30 case notes, as a balance between scale in number of clients and number of case notes per client.

We were interested in 1) the summary similarity between model and human, and 2) the summary similarity with respect to the goal. Similarity between summaries written by a model and by a human is a measure of how consistently a model captures information that a human considers pertinent to a goal. Similarity between summaries for the same goal and distinction between summaries for different goals are a measure of the specificity and distinctiveness of the summary goals.

Results

Figure 1 shows the results for summary similarity between model and human. The top 10 rows (5 prompts, 2 goals) were computed as the scores between model and human summaries for the same goal and client. The last 2 rows are a lower baseline – the scores between human summaries for the same goal but different clients. The second last 2 rows are an upper baseline – the scores between human summaries for the same goal and client. We observe that the model scores (scores between model and human summaries) often lie between the two baselines but significantly behind the upper baseline.

Figure 2 shows the results for summary similarity with respect to the goal. For every run (of score computation), two summaries for the same client were used. For the model summaries (top 10 rows), the score was computed between a summary for the current prompt and each summary for a different prompt, and the average over these 4 scores was taken. For "same goal", the summaries in each run were for the same goal; for "diff goal", the summaries in each run were for different goals. The same procedure is done for the human summaries (bottom 2 rows), where instead of 5 prompts we have 2 humans. As expected, we observe that each "same goal" row is generally greater than "diff goal" row below. Although this occurs less distinctly in the model summaries, it shows that model summaries reflect the specificity and distinctiveness of summary goals.

Summary-Aided Visualization

In light of the Figure 1 results in which model performance has a long way to go from human performance, we have begun to consider a visualization that incorporates both the summary and the input case notes. Specifically, the visualization would place the summary and case notes side by side, and highlight and link parts of the summary and the case notes that are related. This approach would remove the dichotomy between reading a summary and reading case notes, and in the process, address several of the concerns expressed by the interviewed caseworkers. In particular, the evidence supporting a summary sentence would be readily accessible, removing the risk of bias from missing context or of inaccurate information. In this way, the role of the summary moves from the obligation to include all relevant information to a position of facilitating a caseworker skimming the case notes.

Related work for this approach includes improving readability of research papers, but often requires the training of additional models, such as LLMs fine-tuned to a domain, sentence classifiers, or question-answering models (Lee et al. 2019; Fok et al. 2023; August et al. 2023). However, in our study context as well as many others in the public sector, the labeled or unlabeled data necessary to train such models is unavailable. Thus, we seek to detect the evidence relationships connecting summary sentences with case note sentences by making use of pre-trained models. The two approaches we have attempted are sentence-embedding similarity (Gao, Yao, and Chen 2021) and natural language inference (Schuster et al. 2022). Our naive, preliminary results have not yielded satisfactory results, but we nevertheless believe this to be a promising avenue for addressing the challenges with summarization in low data settings.

Discussion and Conclusion

The diversity in philosophical approaches and role-specific needs to casework greatly influences the type of information necessary for a summary. An LLM's ability to tailor its summarization process to align with these philosophies and roles becomes crucial. The summary cannot be a generic extraction of data, but should be a reflection of the nuanced needs and perspectives of individual caseworkers. This distinction underscores the importance of a flexible, user-guided summarization process. Designing LLM interfaces that allow caseworkers to specify or adjust the focus of summaries can significantly enhance the utility of these AI tools in diverse case management roles.

This research highlights the potential and challenges in adapting general-purpose LLMs for the creation of domain-specific summaries, particularly in the context of casework. Our findings demonstrate that while LLMs can process and summarize extensive text data, the key to their effective application is in carefully crafted prompt engineering and an in-depth understanding of the domain-specific nuances. Developing frameworks and best practices for representing the knowledge of domain-experts through prompt engineering are still ongoing. While this gap remains, alternative visuals and interfaces may aid in establish user trust and limiting bias. Additional research and evaluation is necessary to determine how to best apply LLMs toward text summarization in public sector, high-stakes domains.

References

Arawjo, I.; Swoopes, C.; Vaithilingam, P.; Wattenberg, M.; and Glassman, E. 2023. ChainForge: A Visual Toolkit for Prompt Engineering and LLM Hypothesis Testing. *arXiv preprint arXiv:2309.09128.*

August, T.; Wang, L. L.; Bragg, J.; Hearst, M. A.; Head, A.; and Lo, K. 2023. Paper Plain: Making Medical Research

[1] This does not constitute the entirety of our human summaries, but due to iterative revision of the summarization requirements and incomplete overlap in the set of clients annotated by each human, not all human summaries collected could be used in this experiment.

Papers Approachable to Healthcare Consumers with Natural Language Processing. *ACM Trans. Comput.-Hum. Interact.*, 30(5).

Bahrainian, S. A.; Feucht, S.; and Eickhoff, C. 2022. NEWTS: a corpus for news topic-focused summarization. *arXiv preprint arXiv:2205.15661*.

Barale, C. 2022. Human-centered computing in legal NLP-An application to refugee status determination. In *Proceedings of the Second Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, 28–33.

Braun, V.; and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2): 77–101.

Chuang, Y.-N.; Tang, R.; Jiang, X.; and Hu, X. 2023. Spec: A soft prompt-based calibration on mitigating performance variability in clinical notes summarization. *arXiv preprint arXiv:2303.13035*.

Dang, H.; Benharrak, K.; Lehmann, F.; and Buschek, D. 2022a. Beyond text generation: Supporting writers with continuous automatic text summaries. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, 1–13.

Dang, H.; Goller, S.; Lehmann, F.; and Buschek, D. 2023. Choice over control: How users write with large language models using diegetic and non-diegetic prompting. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–17.

Dang, H.; Mecke, L.; Lehmann, F.; Goller, S.; and Buschek, D. 2022b. How to prompt? Opportunities and challenges of zero-and few-shot learning for human-AI interaction in creative applications of generative models. *arXiv preprint arXiv:2209.01390*.

Fok, R.; Kambhamettu, H.; Soldaini, L.; Bragg, J.; Lo, K.; Hearst, M.; Head, A.; and Weld, D. S. 2023. Scim: Intelligent Skimming Support for Scientific Papers. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, 476–490. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701061.

Gao, T.; Yao, X.; and Chen, D. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6894–6910. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. arXiv:2310.06825.

Kanapala, A.; Pal, S.; and Pamula, R. 2019. Text summarization from legal documents: a survey. *Artificial Intelligence Review*, 51: 371–402.

Kumar, H.; Musabirov, I.; Shi, J.; Lauzon, A.; Choy, K. K.; Gross, O.; Kulzhabayeva, D.; and Williams, J. J. 2022. Exploring the design of prompts for applying gpt-3 based chatbots: A mental wellbeing case study on mechanical turk. *arXiv preprint arXiv:2209.11344*.

Kuo, T.-S.; Shen, H.; Geum, J.; Jones, N.; Hong, J. I.; Zhu, H.; and Holstein, K. 2023. Understanding Frontline Workers' and Unhoused Individuals' Perspectives on AI Used in Homeless Services. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–17.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240.

Ma, C.; Wu, Z.; Wang, J.; Xu, S.; Wei, Y.; Liu, Z.; Guo, L.; Cai, X.; Zhang, S.; Zhang, T.; et al. 2023. ImpressionGPT: an iterative optimizing framework for radiology report summarization with chatGPT. *arXiv preprint arXiv:2304.08448*.

Mishra, A.; Soni, U.; Arunkumar, A.; Huang, J.; Kwon, B. C.; and Bryan, C. 2023. PromptAid: Prompt Exploration, Perturbation, Testing and Iteration using Visual Analytics for Large Language Models. *arXiv preprint arXiv:2304.01964*.

Narayan, S.; Zhao, Y.; Maynez, J.; Simões, G.; Nikolaev, V.; and McDonald, R. 2021. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9: 1475–1492.

Reamer, F. G. 2005. Documentation in Social Work: Evolving Ethical and Risk-Management Standards. *Social work (New York)*, 50(4): 325–334.

Saxena, D.; Moon, E. S.-Y.; Chaurasia, A.; Guan, Y.; and Guha, S. 2023. Rethinking" Risk" in Algorithmic Systems Through A Computational Narrative Analysis of Casenotes in Child-Welfare. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–19.

Schuster, T.; Chen, S.; Buthpitiya, S.; Fabrikant, A.; and Metzler, D. 2022. Stretching Sentence-pair NLI Models to Reason over Long Documents and Clusters. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2022*, 394–412. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Slota, S. C.; Fleischmann, K. R.; Lee, M. K.; Greenberg, S. R.; Nigam, I.; Zimmerman, T.; Rodriguez, S.; and Snow, J. 2023. A feeling for the data: How government and nonprofit stakeholders negotiate value conflicts in data science approaches to ending homelessness. *Journal of the Association for Information Science and Technology*, 74(6): 727–741.

Sun, W.; Fang, C.; You, Y.; Miao, Y.; Liu, Y.; Li, Y.; Deng, G.; Huang, S.; Chen, Y.; Zhang, Q.; et al. 2023. Automatic Code Summarization via ChatGPT: How Far Are We? *arXiv preprint arXiv:2305.12865*.

Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.

White, J.; Fu, Q.; Hays, S.; Sandborn, M.; Olea, C.; Gilbert, H.; Elnashar, A.; Spencer-Smith, J.; and Schmidt, D. C. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Yin, F.; Vig, J.; Laban, P.; Joty, S.; Xiong, C.; and Wu, C.-S. J. 2023. Did You Read the Instructions? Rethinking the Effectiveness of Task Definitions in Instruction Learning. *arXiv preprint arXiv:2306.01150*.

Zamfirescu-Pereira, J.; Wei, H.; Xiao, A.; Gu, K.; Jung, G.; Lee, M. G.; Hartmann, B.; and Yang, Q. 2023a. Herding AI Cats: Lessons from Designing a Chatbot by Prompting GPT-3. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, DIS '23, 2206–2220. New York, NY, USA: Association for Computing Machinery. ISBN 9781450398930.

Zamfirescu-Pereira, J.; Wong, R. Y.; Hartmann, B.; and Yang, Q. 2023b. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–21.