

The Promises and Perils of using LLMs for Effective Public Services

Erina Seh-Young Moon
 University of Toronto
 Toronto, Canada
 erina.moon@mail.utoronto.ca

Nuzaira Habib
 University of Toronto
 Toronto, Canada
 nuzaira.habib@mail.utoronto.ca

Devansh Saxena
 University of Wisconsin-Madison
 Madison, United States
 devansh.saxena@wisc.edu

Matthew Tamura
 University of Toronto
 Toronto, Canada
 matthew.k.tamura@gmail.com

Behnaz Shirazi
 Child Welfare Institute, CAS of
 Toronto
 Toronto, Canada
 bshirazi@torontocas.ca

Angelina Zhai
 Georgia Institute of Technology
 Atlanta, United States
 angelina.zhai@gatech.edu

Altaf Kassam
 Child Welfare Institute, CAS of
 Toronto
 Toronto, Canada
 akassam@torontocas.ca

Shion Guha
 University of Toronto
 Toronto, Canada
 shion.guha@utoronto.ca

ABSTRACT

Governments are the primary providers of essential public services and are responsible for delivering them effectively. In high-stakes decision-making domains such as child welfare (CW), agencies must protect children without unnecessarily prolonging a family's engagement with the system. With growing optimism around AI, governments are pushing for its integration but concerns regarding feasibility and harms remain. Through collaborations with a large Canadian CW agency, we examined how LocalLLM and BERTopic models can track CW case progress. We demonstrate how the tools can potentially assist workers in opportunistically addressing gaps in their work by signaling case progress/deviations. And yet, we also show how they fail to detect case trajectories that require discretionary judgments grounded in social work training, areas where practitioners would actually want support to pre-emptively address substantive case concerns. We also provide a roadmap of future participatory directions to co-design language tools for/with the public sector.

CCS CONCEPTS

- Human-centered computing → Empirical studies in HCI; • Applied computing → Computing in government.

KEYWORDS

Large Language Models, Human-LLM Collaboration, child welfare, public sector

ACM Reference Format:

Erina Seh-Young Moon, Matthew Tamura, Angelina Zhai, Nuzaira Habib, Behnaz Shirazi, Altaf Kassam, Devansh Saxena, and Shion Guha. 2026.

Please use nonacm option or ACM Engage class to enable CC licenses  This work is licensed under a Creative Commons Attribution 4.0 International License. *CHI '26, April 13–17, 2026, Barcelona, Spain*
 © 2026 Copyright held by the owner/author(s).
 ACM ISBN 979-8-4007-2278-3/2026/04
<https://doi.org/10.1145/3772318.3790297>

The Promises and Perils of using LLMs for Effective Public Services. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain*. ACM, New York, NY, USA, 21 pages. <https://doi.org/10.1145/3772318.3790297>

1 INTRODUCTION

With growing optimism around the potential of AI to improve public operations and services, governments in North America are aggressively experimenting with and adopting AI tools within public agencies [48, 89, 90, 130]. For example, the US Department of Homeland Security recently released a guide on how government officials can improve public service delivery through the responsible use of generative AI tools [114]. Additionally, Canada was the first country to implement a national AI strategy [8] and has recently partnered with a private company that develops LLMs to identify opportunities for AI to improve public sector services [54]. Underlying this movement lies an assumption that AI tools such as LLMs can deliver significant efficiency and tangible productivity gains for public agencies while reducing wasteful governmental spending [31, 49, 54]. However, what remains insufficiently explored is how specific applications of AI can improve current public sector practices.

The HCI community has a long history of researching the public sector using computational text analysis tools and following the emergence of LLMs, interest in this area has grown significantly [92]. Previously, Saxena et al. [110] applied LDA topic modeling to deconstruct how risk is conceptualized within child welfare narratives, and Field et al. [35] examined whether natural language processing tools can exacerbate racial biases within these systems. Recently, Nelson et al. [85] also explored the feasibility of using LLMs to summarize casenotes to assist homelessness caseworkers in their jobs. Over the years, SIGCHI research has moved from applying computational text analysis tools on narrative documents to understand street-level realities of the public sector to exploring how LLMs can be used on these texts to support workers [42, 81, 85].

Motivated by the growing interest of North American governments in adopting AI tools for the public sector [89, 90, 130], our study builds on prior research that has examined the potential of applying computational text analysis tools on social work narrative texts [42, 81, 85, 110, 111]. In this study, through collaborations with a large Canadian child welfare agency, HCI scholars and child welfare practitioners who work at the agency collaborated together to explore how AI tools can be used to address a key operational challenge for the agency, ensuring that clients do not remain engaged with the agency for longer than necessary. The agency was concerned that caseworkers were engaging with some families for unnecessarily long periods and wanted to explore how computational text analysis tools could be applied to casenote narratives to improve their operational practices. To this end, we sought to track a family's progress in meeting predefined Service Plan goals in Regular casenotes using BERTopic [45] and LocalLLMs (specifically, Llama 3.1 [3]). Due to the highly sensitive nature of the data, we employed computational text analysis tools that can be run locally [94]. In this study, we asked the following research questions:

- **RQ1:** How can language models be used to understand how case management goals are defined and tracked in social work?
- **RQ2:** How can LLMs assist caseworkers to track case progress goals?
- **RQ3:** How can HCI researchers engage with the public sector to design LLM tools that support public sector work processes?

This paper makes the following unique research contributions.

- We showcase an example of how HCI researchers and child welfare professionals can come together to investigate how language models can be used to address an important operational challenge for a large child welfare agency.
- Drawing on two types of child welfare documents (Service Plans and Regular casenotes), we illustrate how caseworkers can use AI tools based on language models to track how child welfare case management goals are defined and are being met.
- We outline the opportunities and limitations of using LLMs within child welfare systems. We reveal how child welfare practitioners can take advantage of the scalable capabilities of LLMs to support the delivery of effective services to clients. However, we also highlight how social work is necessarily inherently interpretive and context-dependent, which means that LLMs cannot and should not replace street-level decision-making [67].
- We outline a roadmap for how HCI scholars can collaborate with the public sector to design AI tools for and with the public sector.

To address our RQs, we structured our study as a three-stage process, which we visualize in Figure 1. To address RQ1 (Section 5), we applied BERTopic on Service Plan goals and Regular casenote narratives. We extracted topics and elicited shared themes between the two document types to understand how the types of child welfare case management goals mentioned in the Service Plan are tracked and worked on throughout a case. We addressed RQ2 in two steps (Sections 6 and 7). We took a subset of our casenote dataset

to apply a LocalLLM to identify which Regular casenotes contain information relevant to Service Plan goals. We manually labeled the casenotes as well to evaluate the LocalLLM's performance. Then, on the same subsetted casenote dataset, we used a LocalLLM to inquire which themes we had identified in the first step of the study (Section 5) appeared in Regular casenotes that had been labelled as containing Service Plan-relevant/irrelevant goals. Through these two stages, we could identify case progress-relevant documents and examine their narrative content. We then reflected on RQ3 in the Discussion section.

2 REFLECTION ON RESEARCH ETHICS

We received approval from the Research Ethics Board of our university to use child welfare narrative documents for this research. We are deeply aware that beyond the REB guidelines, research using child welfare documents is morally complex and entangled with issues of oppression and surveillance. Considering that Canadian governments are pushing for greater AI adoption within the public sector, we hope our study can surface opportunities for and limitations on using AI within the child welfare space. When conducting this study, we considered the principles of public interest and closely collaborated with our co-authors at the child welfare agency to ensure that we were conducting research in the interest of the agency's clients. We anonymized all personal information from the data and do not make any raw data public.

3 RELATED WORK

3.1 SIGCHI research on public sector data-driven tools

SIGCHI researchers have a long history in critically researching public sector sociotechnical systems, including examining issues around algorithmic governance and public data [51, 57, 73, 80], introducing novel participatory HCI methodologies [43, 46, 62], and designing human-centered computational tools that promote human agency and empowerment [42, 63, 106, 109]. Most relevant to this study, in recent years, the SIGCHI community has turned its attention to researching algorithmic decision-making tools that are built using the vast quantities of personal data collected by public agencies [16, 23, 66, 116]. Systematic literature reviews on the technical components of these tools have found that these decision-making tools are largely configured to solve prediction problems [53, 74, 79, 104, 117]. Learning patterns in historical data using statistical and machine learning techniques, the majority of these tools are intended to classify people according to some value or risk they present to society so that those deemed ‘deserving’ can be prioritized for public service delivery [53, 69]. The overarching impetus for using these tools is to support or supplant human decision-making such that public services are delivered to clients in a more consistent, objective, and evidence-based manner [69, 100]. However, recent work by SIGCHI scholars has raised concerns on the validity and limited utility of these tools when deployed on the ground. For example, in studying predictive algorithms designed for adults and homelessness social service provision, Showkat et al. [117] and Reimund et al. [99] found that cost-saving and efficiency are central values promoted in these algorithms, which can result

in the de-prioritization of important human values and an abstraction away from critical contextual client information. Furthermore, Moon and Guha [79] and Saxena et al. [105] found that such predictive algorithms often fail to account for their resource-constrained deployment context. Accurate algorithmic prediction outputs do not necessarily translate to the successful delivery of public services because public resources, such as good foster homes or housing opportunities, are often very limited [33, 68, 79, 99, 107].

Recognizing the implementation gap between decisions rendered by predictive decision-making algorithms and the delivery of public services, researchers have called for the need to shift from a prediction-focused paradigm towards an intervention-focused paradigm – I.e., where public sector data should be used to effectively support public sector workers' decision-making processes rather than designing predictive decision-making algorithms that solely focus on the efficient delivery of services [37, 42, 57, 69, 79, 105]. Social service literature has long found that decision-making and the quality of public services delivered can be improved when organizations focus on how well an organization meets its pre-defined objectives [25, 65, 119, 131]. In line with this social work scholarship, SIGCHI researchers have proposed applying novel co-design methodologies such as comic boarding and voting to elicit stakeholder needs and preferences to design effective technical tools that center stakeholder (and in particular) client needs [42, 43, 62, 63, 123]. Furthermore, HCI scholars have also begun exploring the viability of applying computational text analysis techniques on narrative casenotes written by caseworkers to support public sector workers [42, 85, 110, 111]. Nelson et al. [85] recently found that caseworkers found value in computational tools that can summarize casenotes to provide customized services for clients and Saxena et al. [110] unveiled that casenotes can reveal critical and contextual case information, highlighting their potential as a valuable data source for developing decision-support tools for public sector workers.

3.2 SIGCHI research on child welfare

Recent SIGCHI research on child welfare systems has centered on understanding how technologies can empower and uplift the values of different groups engaged in CW services. Work in this domain has been wide-ranging. Some studies have focused on how CWS members engage with- and mediate- online technology. For example, Ammari et al. [4] examined how individuals in the foster care system find community on online Reddit communities. Furthermore, Caddle et al. [18], Badillo-Urquiola et al. [10], and Badillo-Urquiola et al. [9] have inquired how collaborative sociotechnical systems that engage foster parents and caseworkers can mitigate adolescent online safety risks. In recent years, a plethora of research has critically examined CWS algorithmic decision-making tools, which have grown in prominence across North American child welfare agencies [103, 104]. Through a review of these tools [104] and community design workshops with stakeholders [16, 120], the SIGCHI community has consistently found that CW algorithms narrowly focus on case deficits by using the target outcome – child maltreatment risk – rather than focusing on family strengths. The researchers [16, 104, 120] have argued this is problematic as the algorithmic targets diverge from the overarching goal to improve

child welfare and instead lead to systemically biased and punitive outcomes for families. SIGCHI researchers have also conducted audits examining the technical underpinnings of these algorithms, which support the above-mentioned literature. For example, in reviewing different CW decision-making algorithms that predict child maltreatment risk, Gerchick et al. [39] and Moreau et al. [82] found CW models used protected attributes and racially biased features that reinforce systemic discrimination and stigmatization. Additionally, Saxena et al. [112] found that the tools would use a family's cooperativeness with caseworkers to predict child maltreatment risk rather than assess CW interventions' effectiveness. On-the-ground studies examining how workers engage with these tools found that caseworkers are, in fact, aware of the tools' limitations and accordingly detect/mediate erroneous algorithmic outputs by gaming inputs going into the tools [105] or overriding erroneous recommendations [21, 30, 57]. With increasing awareness of the limitations of current algorithms in improving child welfare outcomes, SIGCHI researchers have increasingly emphasized the need to move away from predictive, deficit-focused models [81, 110, 120] and instead towards designing fair, human-centered, holistic technical tools that support child welfare workers in making informed, strength-based decisions for families [22, 57, 106, 111].

3.3 SIGCHI Computational text analysis research on the public sector

With the growing affordances from computational text analysis techniques, the SIGCHI community has increasingly explored how language tools can be used to study sociotechnical systems and elicit contextual information [2, 4, 19, 42]. For example, Abebe et al. [2] and Chancellor et al. [19] found that applying topic models on social media data can provide predictive information on maternal mortality rates and mental illness severity. Relatedly, Antoniak et al. [5] found running topic models on birth stories on Reddit can reveal aggregated patterns of typical narrative sequences and also rich and unique information on individual birth experiences and power relationships between personas [5]. A growing body of scholarship has also explored the viability and opportunities of applying computational text analysis on public sector narrative casenote data [35, 81, 85, 110, 111]. In many public sector domains, caseworkers document copious unstructured narratives that record their observations, relevant details, and interactions with relevant parties [88, 111, 125]. By applying topic models on narrative casenotes, researchers have shown the viability of applying computational text analysis on these documents to elicit patterns of invisible work conducted by caseworkers that are unaccounted for in ethnographic studies [111] and opportunities to infer how procedural and transitory factors impact bureaucratic street-level decisions [110]. Furthermore, Field et al. [35] have explored the possibility of examining racial disparities in casenotes using word statistics and named entity recognition analysis, and Moon et al. [81] have identified limitations in using casenotes to predict future outcomes through predictive validity assessments.

Most recently, there has been growing interest within and beyond SIGCHI in exploring how LLMs can be integrated within the public sector to support public sector workers [17, 85, 92–95, 118]. For example, within HCI literature, through design sessions with

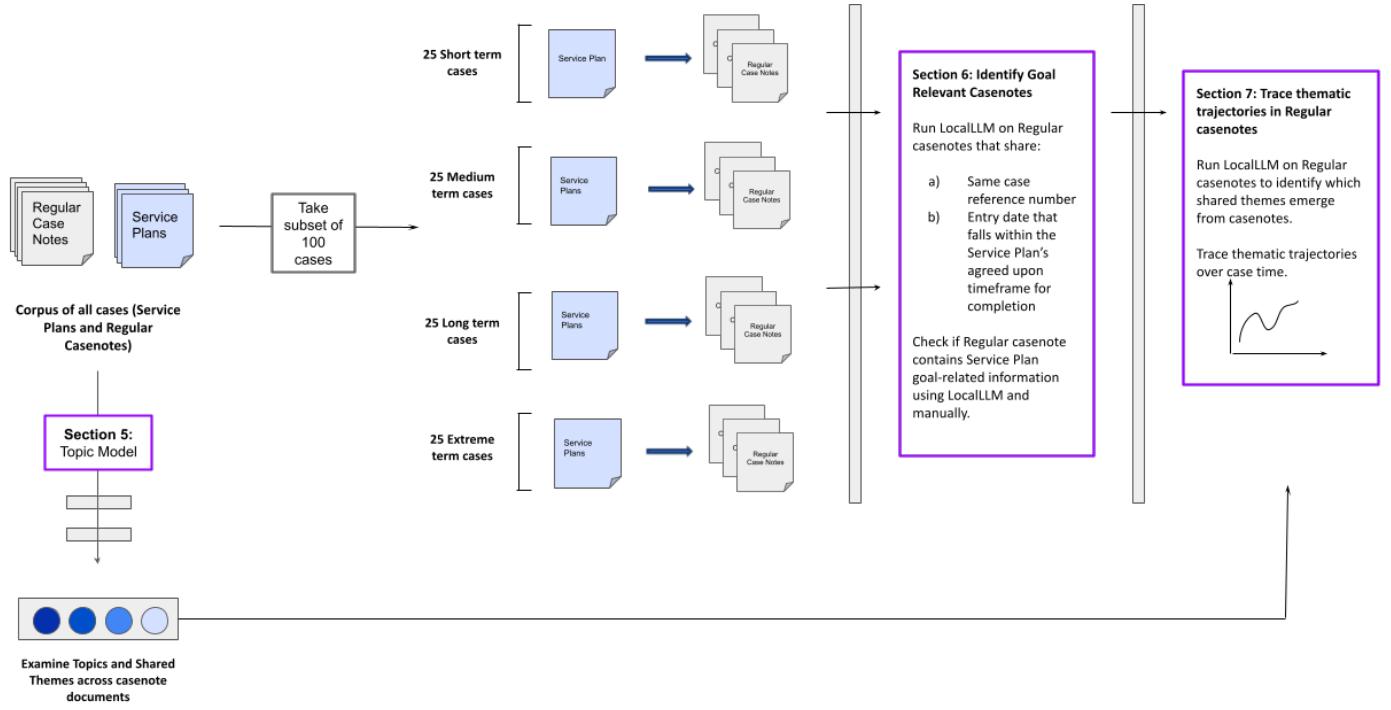


Figure 1: Visualization of this study's analysis steps

homelessness caseworkers, Gondimalla et al. [42] and Nelson et al. [85] found that workers are interested in the development of tools, including LLMs, that allow them to effectively use information from casenotes to support informed decision-making for clients. While public agencies collect vast quantities of narrative data on clients on electronic information systems, social work scholarship has consistently found that caseworkers struggle to parse through dense casenote documents to identify core issues impacting clients and quickly access relevant information [50, 96, 102]. LLMs appear to offer new technical opportunities to distill and surface key information. Beyond SIGCHI, Perron et al. [94] and Stoll et al. [122] evaluated the ability of LocalLLMs in extracting child welfare case factors from unstructured child welfare narratives, and Báez et al. [17] and Singer et al. [118] examined how social work educators can leverage LLMs to enhance social work education.

Our literature review shows there has been growing interest in SIGCHI and beyond to utilize computational text analysis techniques on public sector data - particularly casenote data - to support caseworkers in their work. However, there remain immense concerns around using LLMs on highly-sensitive data, the subjectivity of their outputs, and the risk of human over-reliance on them in high-stakes contexts [6, 14, 94]. As governments aggressively explore how public agencies can integrate LLMs within the public sector, we applied BERTopic and local language models on child welfare casenotes to examine how computational text analysis tools can and cannot aid caseworkers in effectively delivering child welfare services for families.

4 RESEARCH CONTEXT

In this study, we partnered with a non-profit child welfare agency that serves a large metropolitan area in Canada. The organization is regulated and governed by the provincial child welfare ministry with the mandate to protect children and youth from abuse and neglect by providing risk assessment, family guidance and counselling, and permanency planning services. The organization is always interested in improving its operational decision-making processes to enable caseworkers to deliver timely, targeted services while minimizing unnecessary family involvement with the system. While acknowledging that families may require CW support to mitigate the risk of future harm to children, the organization also noted that prolonged involvement with the system can impose an intrusive and stressful burden on families and divert caseworkers from providing support to other families in need [27, 77]. Therefore, through examinations of child welfare casenotes, the organization wanted to gain an in-depth understanding of how child welfare case goals were defined, tracked, and achieved, as well as why some cases exceeded average timeframes for case resolution. Following provincial child protection guidelines, caseworkers at the agency document all client-related information, including internal and external discussions, services delivered, and observations about the family, in casenotes. As a result, child welfare casenotes provide a rich source of contextual insights into a family's experience within the child welfare system [81, 110, 111]. For this study, we focused on two types of narrative documents documented by caseworkers, Service Plans and Regular Casenotes for families receiving Ongoing

Case Management Services. We provide further detail on this in the following paragraphs.

4.1 Data Overview

For this project, we focused on casesnotes written for child welfare cases that received Ongoing Case Management Services (“Ongoing Services”) from the agency. When concerns regarding a child are reported to the agency, and substantiated through initial assessments and investigations, and it is determined that the family requires Ongoing Services to protect the child from future harm, the case is formally transferred to Ongoing Services. In Figure 2, we show a simplified case flow diagram showing how cases are transferred to Ongoing Services (yellow boxes in the Figure).

The first 30 days in receiving Ongoing Services are a critical period where the foundation for subsequent casework decisions is made. During the first month, caseworkers gather and holistically assess detailed information on family functioning, strengths, and unique needs. Caseworkers then work collaboratively with family members to draw up a **Service Plan** that outlines: (1) objectives, (2) activities, (3) the rationale for each objective and activity, (4) agreed-upon timeframes for completion, and (5) the family’s unique case reference number. Objectives are designed to depict higher-level, longer-term goals for a case while Activities are intended to depict shorter-term, actionable goals that will allow the family to achieve Service Plan Objectives. A short Reason is also often included to provide brief contextual rationale for setting a particular Service Plan’s Objectives and Activities. For every Service Plan Objective, there can be more than one corresponding Activity and Reason. Below we show a paraphrased and condensed example of a Service Plan’s Objective, Activity, and Reason.

Objective: “For [name] to be raised in a household whose parents demonstrate healthy conflict resolution”

Activity: “[name] to acquire effective methods for controlling his anger.”

Reason: “The children have observed conflicts between their parents, including physical aggression and objects being thrown.”

The Service Plan is an important document that links caseworker assessments with interventions. Per ministry guidelines, the Plan’s goals and reasons are written in standardized and easy-to-understand language because the goal of the document is to provide an actionable framework against which progress can be measured over time for families and workers. After the first month, caseworkers are tasked with meeting the family regularly (at a minimum once a month) and providing services that support the achievement of the Service Plan’s terms. Formally, the Service Plan is reviewed every six months with families, and a new plan may be drawn up reflecting a family’s progress and change in the child or family’s circumstances.

In addition to the Service Plan, throughout a family’s engagement with the agency, caseworkers record contemporaneous **Regular casenotes** that document any information pertaining to the case, including detailed observations and email/text/phone/in-person conversation records with families and other service providers such as doctors or lawyers. Each Regular Casenote includes a unique family reference number and the date, time, and location at which the recorded event took place. Documenting casenotes is a critical

responsibility for caseworkers, as it informs workers about current efforts made to support children and informs workers on the client’s history [38, 85]. Service Plan Objectives and Activities are closely tied with Regular Casenotes because the Service Plan sets the goals that guide caseworkers’ efforts in supporting families, while Regular Casenotes provide real-time updates on how these goals are being achieved by families and facilitated by workers. Based on observable changes in family functioning and parenting, the agency will close a case when there is no longer evidence of safety threats to the child and improvements have been made in the terms outlined in the most recent Service Plan.

4.2 Dataset Description

Percentile	Case Duration (Days)	N
0-25%	5 - 145	180
25-50%	146 - 232	180
50-75%	233 - 387	181
75-100%	388- 840	179

Table 1: Dataset Case Duration Percentiles and Counts

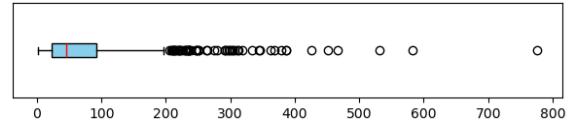


Figure 3: Distribution of the Number of Regular Casenotes Per Family

Dataset For this study, we obtained Regular and Service Plan casenotes from 720 families (i.e., 720 ‘cases’) who received Ongoing Services at the agency between January 2022 and June 2025 and were discharged by June 2025 (henceforth called the ‘dataset’). Cases ranged in duration with the shortest case being 5 days and the longest being 840 days, averaging 275 days, overall. In Table 1, we show the distribution of case durations for our dataset and a breakdown of the number of families within each duration quartile range. As seen in the Table 1, we found that the agency engaged with an approximately equal number of families within the four duration quartile ranges, suggesting that the agency was handling an equal distribution of cases of varying complexities. Our collaborators in this study expressed a particular interest in understanding why some cases were falling within the 75% to 100% percentile of case duration range.

Regular Casenotes Our dataset had a total of 52,748 regular casenote records for the 720 families. On average, a family had a mean number of 73 regular casenote entries but as seen in Figure 3, the number of regular casenotes entered for each family varied significantly.

Service Plans Not all families had Service Plans, especially if a case was closed within 30 days. In our dataset, we had a total of 1,213 Service Plans for 654 families, which consisted of a total of 1677 Objectives, and 2,288 Activities. As seen in Table 2, each family typically only had 1 to 2 Service Plans drawn up during their engagement with the agency although longer cases had multiple Service Plans which were drawn up during bi-annual Service Plan review meetings (see Figure 2).

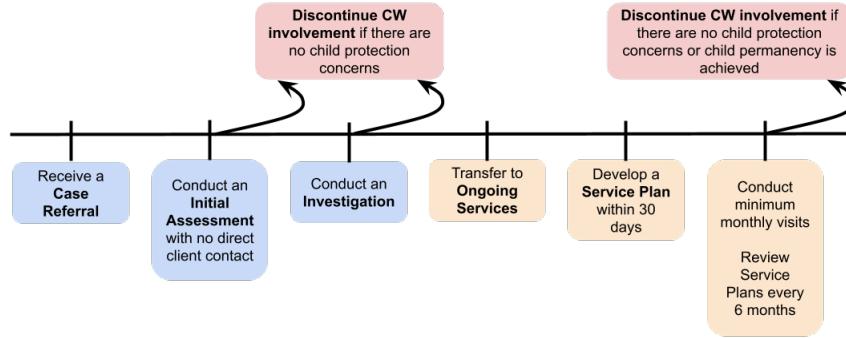


Figure 2: Simplified case flow diagram within the agency (In this study, we focus on cases that receive Ongoing Services depicted in the orange boxes)

Number of Service Plans	Number of Cases
1	320
2	188
3	97
4	31
5 - 7	18

Table 2: Distribution of the Number of Service Plans Per Family

5 ELICITING THEMATIC CONNECTIONS BETWEEN REGULAR CASENOTES AND SERVICE PLAN OBJECTIVES AND ACTIVITIES (RQ1)

In Section 4.1, we explained how the goals set in the Service Plan are closely related to Regular casenotes: Service Plan goals guide the delivery of services to clients; Regular casenotes then document family progress in relation to the goals while also capturing dynamic, changing circumstances of clients, which in turn, can reshape Service Plan goals. Considering the close, symbiotic relationship between the two documents, we were interested in identifying thematic similarities and dissimilarities between the two documents in this study. To this end, we applied topic modeling (specifically BERTopic [45]) on the 1) Service Plan’s Objectives, 2) Activities, and 3) Regular casenotes. In the following subsections, we describe our data preprocessing steps, analysis approach, and findings.

5.1 Data Cleaning, Preparation, and Anonymization

For each document type (i.e., Service Plans and Regular casenotes), we extracted and saved core information from the document into a tabular format, including a case’s reference number and date of record. For Service Plans, we also extracted information on the time period for which the goals will be worked on, as well as the Plans’ objectives, activities, and reasons; and for Regular casenotes, we extracted narrative records.

Following extraction, we cleaned the data, removing extraneous whitespace, line breaks, and special characters. To prevent distinct clustering based on language, we removed French words by passing each sentence through the langdetect library [40], and removing sentences that have been detected as French. An additional manual cleaning step was taken afterwards to remove any repetitive French sentences that langdetect missed. Following cleaning, we anonymized personal information to protect the privacy of the persons within the cases. The anonymization process was conducted with the following 3 steps. First, we removed personally-identifying information, namely email addresses and phone numbers, replacing them with [email] and [phone], respectively. Next, we replaced client names with the label [client name], so that information pertaining to the client could remain identifiable and not be lost in anonymization. Lastly, we used the spaCy English language model “English transformer pipeline” (en_core_web_trf) to identify words within the texts that correspond to a person and replace that section of text with the label [person]. We used a transformer-based language model to anonymize because, unlike strict string-matching methods, it can identify uncommon names and names that have other meanings as words.

5.2 Data Analysis Approach

To generate topics, we ran BERTopic [45], then conducted a manual review to determine the types of topics and themes within the casenotes. BERTopic was selected for its ability to extract nuanced semantic information from text with its transformer-based language model embeddings, and for its superior performance in topic coherence and diversity compared to LDA and LMF [45]. We ran BERTopic using all-MiniLM-L6-v2 embeddings, UMAP, and HDBSCAN. We used UMAP for dimensionality reduction due to its ability to preserve local structure, which facilitates clustering, while also approximately preserving global structure, aiding in the interpretation of semantically similar clusters [75]. We used HDBSCAN to accommodate topic clusters of varying sizes. We set a low minimum cluster size to promote classification and reduce the number of unclassified data points, helping to ensure that the generated topic model was representative of the corpus. Hyperparameters of the models were selected based on two criteria: number of topic clusters, and topic coherence. Upon each adjustment to the

model hyperparameters, we first examined the number of topics. If there were too many topics to feasibly manually examine, the model was retrained with adjustments. Otherwise, the representative documents of each topic were reviewed to determine if there is a coherent and identifiable theme; if not, the model was retrained.

For each of the topic models trained on the Objectives, Activities, and Regular Casenotes, four of the co-authors of the paper employed an open-coding process and manually examined documents and keywords assigned to each topic to manually assign a label to each topic and group the topics into high-level topics [15]. Two other co-authors of the paper who work at our collaborating child welfare agency then checked and corrected the high-level topics based on their expertise [29]. After iterative discussions, we agreed on the high-level topics that emerged in the Services Plan's Objectives, Activities, and Regular casenotes. Through these interpretive steps, we found the three different narrative types shared similar thematic topics, albeit with some differences.

5.3 Results: Thematic Connections Between Regular Casenotes and Service Plan Objectives and Activities

Training BERTopic models on the Regular casenotes generated 60 topics, the Service Plan Objectives yielded 38 topics, and Activities produced 82 topics. Manual inspection of the topics showed that having a high number of topics for each narrative document type allowed for a comprehensive and granular understanding of the wide range of topics covered in the dataset, but these topics could also be grouped into higher-level themes as they contained thematic similarities. As the aim of identifying topics in the Regular Casenotes and Service Plan goals was to understand thematic convergences and divergences between the narrative text types, through iterative and collaborative discussions, we identified high-level themes from the Regular casenotes, Service Plan Objectives, and Service Plan Activities. See Appendix A for examples of how topics were grouped for thematic similarity.

In Table 3, we provide an overview of the themes that emerged across the different narrative types, highlighting those that were shared and those unique to each narrative text type. As seen in the table, the three types of narrative texts shared many thematic similarities as the themes point to core child welfare-related issues such as parenting, health, child custody, and safety. At the same time, thematic differences emerged between the three narrative types due to intrinsic differences in the different narrative text types. The following bullet points explain some of the most shared and unique themes that emerged. See Appendix B for exemplar sentences for the below listed themes.

- **(Shared Theme) Family Relationship/ Visits & Parenting** This theme was about improving relationships between bio-parents and children by promoting healthy family communication, constructive discipline practices, participating in parenting programs, and establishing access visits with parents and children when a child is not placed with the bio-parent.
- **(Shared Theme) Child Custody & Criminal/Legal** This theme was centered on topics related to child custody or legal and criminal justice related issues impacting a family. When

bio-parents were no longer together, child welfare workers were tasked with connecting parents to mediators who can resolve child custody disagreements and caseworkers also supported family members navigate court proceedings. This theme emerged across all narrative document types.

- **(Shared Theme) Medical/Mental Health** This theme appeared in all goal types and regular casenotes and focused on any medical, health, or substance abuse related issues. The theme could include the facilitation and scheduling of medical appointments, checkups, mental health, and substance abuse treatment appointments as well as families working on developing coping strategies when faced with mental health challenges.
- **(Shared Theme) Anger Management Conflict & Safety** This theme appeared across all narrative types and encompassed topics such as child or parent aggression and intimidation, parental conflicts occurring in front of children, and parenting-related safety concerns.
- **(Shared Theme) Administration related tasks, resources, and scheduling** This theme appeared most frequently in the regular casenotes and also emerged in the Objectives and Activities. The focus of this theme was largely on caseworkers facilitating the delivery of child welfare support services to families, interpretation services, referrals, transportation, material resources, and other appointments. Under this theme, caseworkers also conducted administrative tasks, putting together consent forms, obtaining signatures, and providing financial aid to families.
- **(Unique Theme) Attempts to Contact** This theme only appeared in regular casenotes because it was about caseworkers attempting to reach out to various parties, including family members and other service providers.
- **(Unique Theme) Support Network** This theme focused on workers helping families connect and build support systems so that when challenges arise, families can draw on friends, relatives for support and are knowledgeable of available community resources. This theme did not emerge as a distinct category in the Regular Casenotes, as building a support system typically involved workers connecting clients to community service providers and encouraging parents to engage with friends and relatives. As a result, in Regular casenotes, this theme appeared under other themes such as "Family Relationship/ Visits & Parenting" and "Daycare & Equipment" where workers assisted families connect with child supplies and daycare programs.
- **(Unique Theme) Child Development** The theme of ensuring that a child was developing age-appropriately emerged as a unique category only in the Service Plan Objectives. Similar to the "Support Network" theme, this occurred because achieving this Objective required caseworkers to connect children with schools, counseling services, and health professionals. Consequently, the operationalization of this theme appeared under other themes such as "Family Relationship/Visits & Parenting," "Medical/Mental Health," and "School" in the Regular Casenotes and Service Plan Activities.

Super Theme	Regular Casenote	Service Plan - Objective	Service Plan - Activity
Family Relationship/ Visits & Parenting	x	x	x
Child Custody & Criminal / legal	x	x	x
Daycare & Child Equipment	x	x	x
Attempts to Contact	x		
Medical/Mental Health	x	x	x
Anger Management Conflict & Safety	x	x	x
Housing, Home Environment & Adoption	x	x	x
School	x	x	x
Kinship	x	x	x
Administration related tasks, resources, and scheduling	x	x	x
Support Network		x	x
Child Development		x	x

Table 3: Super themes identified in Regular casenotes, Service Plan Objectives, and Activities (The last three columns indicate whether the theme emerged for the specific narrative text type)

Upon manual examination of the 38-topic model solution for Service Plan Objectives and the 82-topic model for Activities, we found that high-level goals (i.e., Objectives) in child welfare at the agency are more standardized, while activities tend to be more customized to the case. Activities provided concrete, actionable steps to achieve a case's Objective. Guided by the Service Plan goals, the narrative information in Regular casenotes directly reflected how caseworkers supported their clients guided by pre-established Service Plan goals while also adapting to clients' unique circumstances. See Appendix C for a more in-depth analysis and distinctions between Service Plan Objectives and Activities.

6 MAPPING ACTIVITY-RELEVANCE IN REGULAR CASENOTES (RQ2)

In the second part of this study, we sought to examine how computational text analysis approaches (i.e., a LocalLLM) could be used to track Service Plan goal progress in Regular casenotes. We employed a LocalLLM because they are open-source LLMs that run in offline environments and do not upload any data to the cloud, unlike their cloud-based counterparts. The local deployment of LocalLLMs ensures the protection and full control of sensitive private data, while still offering sufficient computational capabilities of modern LLMs [94]. Our results from Section 5 revealed that Activities provided actionable guides for families to achieve overarching Service Plan Objectives. We hypothesized that Activities could better track case progress than Objectives, so we employed a LocalLLM to label which Regular casenotes mentioned Service Plan Activities were being completed or worked on and manually reviewed the labels.

6.1 Data Analysis Approach

Llama 3.1 [3] was selected for the LocalLLM data analysis tasks. This model was selected for its lightweight, being 4.9GB in size and having 8 billion model parameters, allowing it to be stored in system RAM on inference on local devices, and for its superior run-time performance compared to state-of-the-art 7-10 billion parameter lightweight LLMs of similar size such as Mistral, PHI4 and deepseek-R1 [44]. Llama 3.1 was run on the Ollama platform

[41] due to its ability to work on multiple operating systems and devices.

To track Service Plan Activity progress within Regular casenotes, we sought to directly match Activities to Regular casenotes by including both the Activity and the casenote within the prompt, and querying if within that Regular casenote, if there is any indication of progress towards the Activity. To mitigate potential issues around prompt brittleness, where small changes in prompt formats can lead to inconsistent performance fluctuations [87], we experimented with multiple prompt formats and selected the prompt below, which generated responses most accurately and relevant to our needs. Specifically, to reduce the varying structures and formats of the text from impacting the Activity-casenote tracking, the LocalLLM was tasked with first generating a summary of the contents of the casenote, and using that generated summary to see if there is indication of progress. The length of the summary remained unspecified, allowing for longer summaries to be generated for content rich casenotes, helping preserve information. We provide the prompt below:

Activity-Regular Casenote progress tracking prompt:

You are analyzing a child welfare worker reviewing case notes. Do the following:

1. Read through the casenote and store it in summary.
2. Assess whether the summary indicates progress toward completing or working on the following activity: {activity_name}. Answer strictly 'Yes' or 'No'.
- Case Note: {narrative_text}

We ran the LocalLLM for a subset of 100 cases (out of the 720 cases in our dataset) where cases had Regular casenotes and at minimum, at least one Service Plan. To ensure we had a representative sample of cases, we randomly selected 25 cases within each case duration percentile as depicted in Table 1. Our subset of 100 cases included 25 cases that fall within the 5-145 days range ("Short Cases"), 25 cases that fall within the 146-232 days range ("Medium Cases"), 25 cases that fall within the 233-387 days range ("Long Cases"), and 25 cases that fall within the 388+ days range ("Extreme

Cases"). The first author also manually read through each regular casenote entry for the 100 cases ($N=6,031$) to compare the accuracy of the LocalLLM's labels with a human labeler and resolved ambiguous casenotes by consulting with two co-authors of the paper who work at the agency. In Table 4, we show the average number of words within each Regular casenote entry for cases of different duration types for our subset of 100 cases. As seen in the table, all cases, irrespective of case length, had approximately similar number of words in each casenote.

6.2 Results

6.2.1 Manually labeling shows approximately half of the Regular casenotes mentioned Activity relevant information except for Extremely long cases. In Figure 4, we show the proportion of regular casenotes that mention Activity-relevant information following our manual labeling. Along the x-axis, we show a normalized timeline of a case. Prior work by Saxena et al. [111] demonstrated that child welfare casenotes follow a structured sequence of events, and that segmenting cases into equal intervals can reveal temporal trends across cases of varying durations, illustrating the 'Life of a Case'. We adopted this methodology to depict temporal trends in cases such that if a case duration was 100 days, casenotes were segmented such that regular casenotes written during the first 10 days fell within the 0.1 normalized time period. On the y-axis, we show the proportion of regular casenote entries that mention Activity-relevant information for the specific normalized time period. As shown in the figure, across each normalized time period, Short, Medium, and Long cases generally included Activity-relevant information in approximately 46–66% of the Regular casenotes written for each period. However, Activity-relevance consistently dropped towards the latter half of Extremely long cases. Manual readings of the casenotes revealed that there were decreases in the proportion of Activity-relevant regular casenotes in Extremely long cases when new child welfare concerns not outlined in the Activities emerged. For example, a family Activity could be set to be "*[name] to engage in positive communication and interactions with [name]*". However, during the timeframe in which the Activity must be completed, a bio-parent may be placed under house arrest and unable to leave home to meet their child. In this case, regular casenotes would no longer include information related to the Activity. We only observed slight increases in Activity-relevance in casenotes during the 0.4 and 0.7 time period when new Service Plans would be typically drawn up for Extreme cases. New Service Plans would address the updated needs of families but even then the rebounds in Activity-relevance were shortlived for complex cases (see the round dotted circles in the Figure). In the following paragraphs, we compare how LocalLLMs labeled Activity-relevance in casenotes compared to manual labels.

6.2.2 Local LLMs can Indicate Case Progress and Deviations in Less Complex, Shorter Duration Cases. In Table 5 we present inter-rater reliability metrics comparing our manual labels with LocalLLM-generated labels for identifying Service Plan Activity content in regular casenotes. As shown in the Table, Short-length cases (5–145 days) achieved a higher Cohen's kappa of 0.604, but as cases became longer in duration, the kappa coefficient decreased to 0.402 for Long cases (233–387 days) and 0.470 for Extremely

long cases (388–840 days). The false positive rate (FPR) slightly increased as cases became longer in duration, while the false negative rate (FNR) remained relatively flat across case durations. This suggested that the LocalLLM tended to label regular casenote entries as Activity-relevant and could not correctly identify Activity-relevant narratives for more longer duration child welfare cases. Manual readings of the cases showed that cases that fell within the Long and Extremely long duration cases involved more topics related to access visit coordination, as children were often placed outside of the home and frequently involved court, legal issues, and more referrals to external service providers. Compared to Short and Medium cases, we observed that underlying issues that initially brought a family to the agency often escalated and triggered a chain of events (often unspecified in Service Plan Activities) that prolonged a family's engagement with the agency in Long and Extreme duration cases, which were then often incorrectly labeled as Activity-relevant by the LocalLLM.

Comparisons between manual and LocalLLM labels showed that the LocalLLM's tendency to predict regular casenote entries as Activity-relevant could be attributed to the model's prompt configuration limitations and dataset characteristics. For example, Activities often used language that could appear vague to the lay reader, but had specific meanings in child welfare literature. If an Activity stated, "*Parents to develop a safety plan in case [name] returns to family home*" or "*The family will work toward enhancing their support system*," the LocalLLM would identify any texts related to general safety or support as Activity-relevant. Furthermore, due to the highly sensitive nature of the data, we had anonymized all identifying information in the casenotes. This meant that when an Activity called on a specific person to carry out an action, the LocalLLM labelled any casenote that mentions any person conducting the action as Activity-relevant. When we manually labeled the texts, we could generally infer which bio-parent the Acitivity was relevant for based on their pronouns and infer who was carrying out the actions in the casenotes based on contextual details.

Table 5 shows that the LocalLLM produced approximately 13–14% false negative errors. We observed that these errors often occurred when the LocalLLM could not infer how acronyms or specific service provider names were related to an activity. For example, an Activity may state, "*[name] to engage in the recommended substance use/mental health related services and follow through with recommendations*." If a Regular casenote entry said, "*[name] from [hospital name acronym] advised that [name] is currently with them and that she came in on the 17th and that she was put on a Form 1 and that she is currently staying voluntarily*," the LocalLLM could not understand that the hospital acronym and specific mental-health related terms (i.e., Form 1) could indicate Activity-relevance.

When manually comparing LocalLLM labels with our manual labels, we also considered whether prompt brittleness could be leading to the misclassifications of Activity-relevance [87]. In most cases, however, we could trace the false positive and false negative errors back to the reasons detailed above, and to underlying ambiguities in how case progress is defined, which we elaborate on below.

6.2.3 Ambiguities arise on what is considered relevant Activity-relevant and require discretionary judgments grounded in

Case Duration	Average Number of Words			Total Count	
	Regular Casenote Entry	Service Plan - Objectives	Service Plan - Activities	Regular Casenote Entry	Service Plans
Short	230.1	9.44	13.57	530	27
Medium	256.6	10.79	12.59	925	26
Long	279.5	9.41	12.52	1485	42
Extreme	241.0	9.38	12.14	3095	76
All cases	251.9	9.63	12.53	6031	171

Table 4: Average number of words and total counts of the cleaned and anonymized Regular casenotes and Service Plans for the subset of 100 cases for progress tracking.

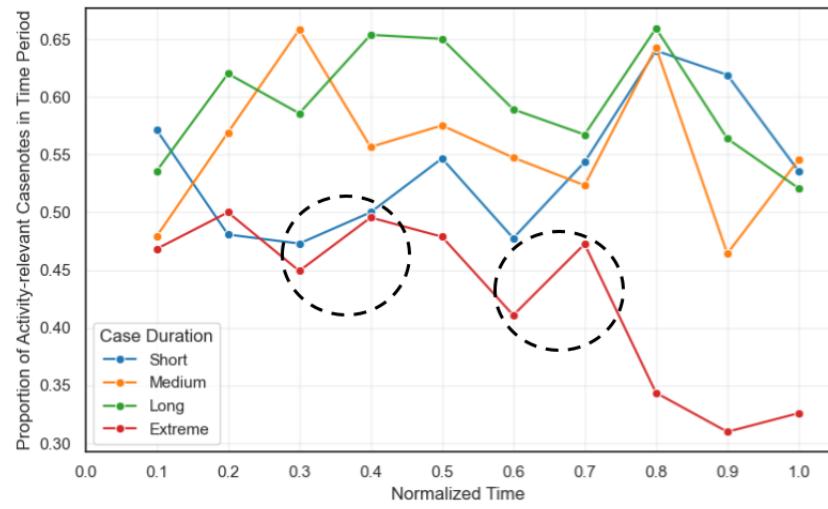


Figure 4: Proportion of Regular casenotes mentioning Activity-relevant information for each normalized time period. New Service Plans are typically drawn up during the 0.3 and 0.6 time period for Extremely long cases (denoted in black dotted circles).

Case Duration	Agreement	Cohen's κ	FPR	FNR
Short	0.804	0.604 [0.535, 0.670]	0.265	0.135
Medium	0.782	0.550 [0.493, 0.602]	0.324	0.135
Long	0.727	0.402 [0.352, 0.447]	0.470	0.146
Extreme	0.730	0.470 [0.440, 0.500]	0.382	0.134

Table 5: Inter-rater reliability between manual and LocalLLM progress tracking, with 95% bootstrapped confidence intervals, and LocalLLM classification performance by case duration category

social work training. It is important to emphasize that the exercise of using a LocalLLM to track Activity progress was not to build a perfect and deployable progress tracker for workers. Instead, the aim of employing a LocalLLM and manually labeling regular casenotes was to uncover challenges and opportunities for applying LocalLLMs within the child welfare domain. Critical limitations to applying LocalLLMs to track Activity progress surfaced when there were casenotes that the first author could not easily label as Activity-relevant and required consultations with two co-authors of the paper who work at the agency to draw on their social work expertise. For example, there was one particular case where the biological father and biological mother were separated and often at odds with each other. An Activity for the family was, “The society

will engage the parents in working with a family mediator to resolve custody and access arrangements.” Manual reading of the casenotes showed that due to parental conflict, the caseworker unwillingly acted as a mediator for the bio-parents to schedule visits with the children, even though parental mediation is not part of their main job duties. In this case, it was unclear if the caseworker acting as an (unwilling) mediator would be considered Activity-relevant. To see a more full picture of the diversity of ambiguous cases encountered, see Appendix D.

Discussions to resolve Activity-relevance ambiguities with our co-authors from the agency demonstrated that determining which casenotes are relevant to a Service Plan’s Activities requires discretionary judgments grounded in social work training and is not a

straightforward task. Instead, temporal and contextual factors surrounding a case must be considered, which can be particularly challenging to parse, define, and operationalize through LLM prompting.

7 THEMATIC TRACKING OF ACTIVITY-RELEVANT REGULAR CASENOTES (RQ2)

Our findings from Section 6 indicated which Regular casenote entries contained Activity-relevant information but did not provide information on the narrative content of the texts. To delve into content-specific, thematic trends within the casenotes, we built on the Super Themes identified in Table 3 in Section 5.3 and applied a LocalLLM to compare thematic narrative trends in Regular casenote entries that had been manually labeled as Activity-relevant and Activity-irrelevant.

7.1 Data Analysis Approach

To track the thematic content of Regular casenotes, we prompted a LocalLLM to identify themes that are present in each Regular casenote entry. We provided the LocalLLM a list of Themes, drawing on the Super Themes we had identified to be present in the Regular casenotes (see the second column Table 3). We used these themes because all of the Activities Super Themes are also present within these set of themes, all except for "Support Network", whose corresponding activities can be classified under the other themes such as "*Family Relationship/ Visits & Parenting*", "*Daycare & Equipment*", and "*Administration related tasks, resources, and scheduling*". By applying this approach, we could determine the themes of focus within each Regular casenote entry. Below, we provide a segment of the LocalLLM prompt. The prompt details the instructions for matching and the preset list of themes that the Regular casenotes should look out for. Similar to how our prompt was developed for Section 6.1, we adopted an iterative approach, experimenting with different prompt formats to mitigate potential prompt brittleness issues.

Regular Casenote Theme extraction prompt:

You are a social work assistant working in a child welfare agency. Your task is to identify within the casenote whether or not a provided theme is mentioned.

Indications must be: Match or No Match.

- "Match" means that the activity content falls within the theme or is directly related to it.

- "No Match" means the activity is not relevant to the theme.

casenote: {casenote}

Themes (with descriptions):

1. Family Relationship/ Visits & Parenting

Description: improving parent child relationship, such as communication in the family and discipline practices, parenting programs and establishing access and visits with parents and family members

2. Child Custody & Criminal / legal

Description: figuring out custody over child and

or going to court to deal with criminal legal...

Model generation parameters were adjusted to provide more consistent generated outputs, including adjusting the temperature parameter to a low value of 0.1, and specifying the json output format. To assess the reliability of the LocalLLM to correctly identify themes in the texts, two of the co-authors manually spot checked some of the thematic labels.

7.2 Results: LocalLLMs can Surface Thematic Divergences in Regular Casenotes

In this section, for improved readability, we present thematic narrative trends for Short and Extremely long cases for four Super Themes that are central to child welfare concerns.

The facet plots in Figure 5 show the proportion of Regular casenotes mentioning Super Themes across normalized time periods for Short cases, separated by casenotes manually labeled as Activity-relevant (figure on left) and Activity-irrelevant (figure on right). For example, the figure on the left shows that of all the Short-duration regular casenotes written during the 0.1 time-period, approximately 25% of Regular casenotes contained information that was 1) Service Plan Activity-relevant and 2) relating to the *Medical/Mental Health* theme.

From the left facet plot of Figure 5, we observe stepwise increases in the four Super Themes throughout the life of a case. This suggests workers generally closely follow Service Plan Activities in Short cases. The left line plot also shows that Activity-relevant casenotes that mentioned *Medical/Mental Health*, *Anger Management Conflict & Safety*, and *Attempts to Contact* themes become more prominent at the final stage of a case, i.e., at normalized time = 1.0. Through manual reading of the casenotes, we discovered this increase occurred at the end of cases because workers often contacted clients to inform them that the agency was closing the case due to a lack of safety concerns.

From the facet plot on the right in Figure 5, which shows the proportion of regular casenotes that do not contain Activity-relevant information for each time period, we observe comparatively fewer mentions of the four Super Themes. We observe that workers support clients on issues related to *Child Custody & Criminal/Legal matters* and *Medical/Mental Health* that are not relevant to Service Plan Activities during the latter half of a case, but their prominence decreases by the end of the life of a case. Manual reading of casenotes showed that, because Short-duration cases were often relatively straightforward, workers could quickly support clients in these areas by connecting them to available service providers.

Compared to Short cases, we observe different thematic trends for Extremely long cases. Figure 6 shows the proportion of Regular casenotes that mention Super Themes across normalized time periods for Extremely long cases, separated by casenotes manually identified as Activity-relevant (left) and Activity-irrelevant (right). From the left facet plot in Figure 6, we observe that Regular casenotes increasingly diverge from Service Plan Activities over the life of a case. For example, the prominence of casenotes that are Activity-relevant and mention themes, *Medical/Mental Health* and *Anger Management Conflict & Safety* decrease steadily over the life of a case, only temporarily rebounding during the 0.7 time

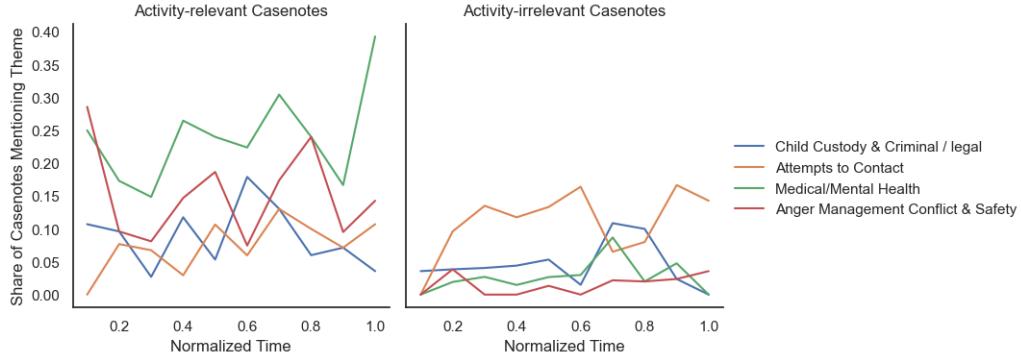


Figure 5: Proportion of Themes present in Regular Casenotes for each normalized time period for Short Duration Cases

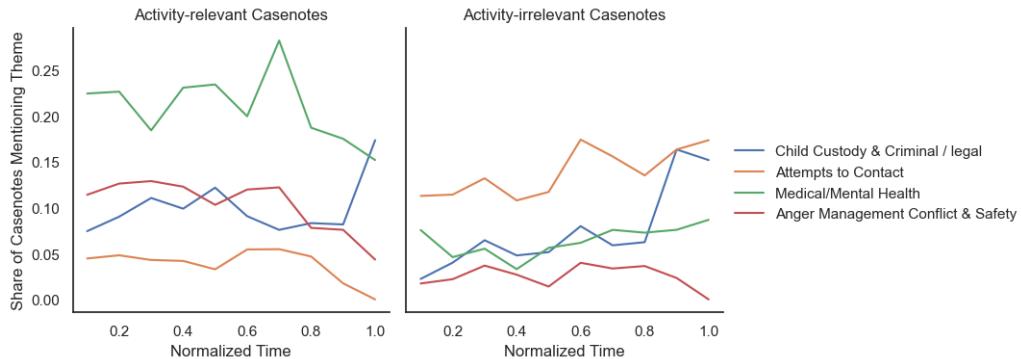


Figure 6: Proportion of Themes present in Regular Casenotes for each normalized time period for Extreme Duration Cases

period when new Service Plans are often drawn up. We also note that workers are making fewer *Attempts to Contact* clients/other related parties regarding Activity-relevant issues (Figure 6 left plot), and instead, workers are making more *Attempts to Contact* people regarding Activity-irrelevant issues (Figure 6 right plot) over time.

In the right facet plot of Figure 6, the increases we observe for the four Activity-irrelevant themes suggest that in Extremely long cases, new child welfare concerns, such as those related to a client's health, substance abuse usage, and criminal/legal issues, become more prominent over time and thus renders existing Service Plan Activities outdated due to emergent child welfare concerns. In fact, through manual readings of casenotes, we confirmed this often occurred in Extreme-duration cases. New concerns relating to a bio-parent's substance abuse issues or being implicated in legal or criminal-related matters would emerge, meaning that existing Activities centered on improving child-parent relationships could not be worked on. Instead, custody arrangements had to be made to establish a permanency plan for the child.

8 DISCUSSION

In this section, we discuss our key findings in relation to our research questions and extend our findings to broader implications for designing LLMs for the public sector.

8.1 Key Findings from our Results (RQ1 & RQ2)

In this section, we discuss our key findings concerning RQ1 and RQ2 to reflect on the opportunities and challenges of adopting LLMs for the public sector.

8.1.1 Promises of Applying LLMs in the Public Sector. The methodology we employed to track thematic trajectories of child welfare cases from Sections 5 and 7 suggests that **LLMs can provide case-workers with tools to track narrative content of casenotes with a high level of granularity and customization** [94], potentially supporting their clinical decision-making processes. In machine learning literature, prior to the induction of LLMs, traditional inferential statistical methods (such as generalized linear models) and topic modeling techniques have been useful in providing trends on average outcomes and aggregated perspectives [13, 45, 78, 86] but they have also exhibited poor performance against outliers [121]. In high-stakes domains such as child welfare, effectively adjudicating child welfare risks while avoiding misclassification of low- and high-risk cases, including rare but severe instances of abuse or neglect (i.e., outlier cases), is critical. By combining thematic clusters from BERTopic models with LLMs, our study provides potential opportunities for workers to use computational text analysis techniques as decision-aid tools, enabling birds-eye, exploratory content analysis of large volumes of unstructured text and then narrowing down into specific topics of interest that could

signal critical child welfare concerns. This approach could allow workers to identify specific thematic child welfare concerns and opportunistically address gaps in their work to better meet the needs of children and youth. Recent HCI scholarship has argued for moving away from a predictive risk-focused paradigm in child welfare tools built on reductive administrative data, noting that caseworkers view social work as a practice centered on generating knowledge about clients, with day-to-day interactions being central to their professional practice, and that they desire tools that support their knowledge-making practice on clients [64, 105]. Through our work, we show that nascent language models (such as BERTopic and LLMs) can surface contextualized client information that may help workers make better-informed child-welfare decisions.

8.1.2 Perils of Applying LLMs in the Public Sector. Despite the abovementioned opportunities, our findings also point to the dangers of applying LLMs blindly in domains such as child welfare, where there is high uncertainty and often no clear ground truths. Our findings in Section 6 show that tracking progress in child welfare cases is not straightforward due to 1) intrinsic uncertainties that underpin child welfare cases [110] and 2) mixed views between frontline workers and agency management on what constitutes a caseworker's duties. In this study, we tracked case progress by identifying which Regular casenotes made relevant references to Service Plan Activity goals manually and with a LocalLLM. **Comparing label agreements between the manual labels and the model showed that the LocalLLM performed increasingly worse for more complex, longer-duration cases, not only due to technical configuration limitations of the LocalLLM prompt¹ but also because of fundamental ambiguities on what is considered an Activity-relevant casenote entry.** As noted in Section 6.2.3, co-authors of the paper, including authors who work at the agency, needed to collaboratively discuss to reach a consensus on whether some casenotes would be deemed Activity-relevant. In cases, where caseworkers took on job roles that were outside of their job duties but were tangentially related to Service Plan Activity goals (e.g., when an Activity was for a caseworker to connect a family with a mediator but the worker took on the role of a mediator themselves albeit unwillingly), our co-authors from the agency argued that from a management perspective, such casenotes may be considered Activity-irrelevant. However, the co-authors also acknowledged that caseworkers often support families in ways that fall outside their official job duties to build rapport with clients, and therefore, could also be considered Activity-relevant [111]. Our agency co-authors explained that child welfare work involves balancing the dual objectives of reducing harm and preserving the family unit [36], meaning interpretations of Activity-relevance can vary between agency management staff and frontline caseworkers, as well as among frontline workers themselves.

Our findings in Section 6.2.1 also show that child welfare case goals are not a static construct, and tracking case progress based on documented Service Plan goals can be problematic. In Section 6, we found that over a life of a case, Regular casenotes made fewer mentions of activity-relevant content in Extremely long cases and through thematic tracking of casenotes in Section 7, we found

¹These limitations could be improved through further prompt engineering or incorporating domain knowledge through retrieval-augmented generation technology [93].

this was because caseworkers were addressing new emergent and pressing child welfare concerns that rendered existing Service Plan goals outdated. Our study shows that, similar to how caseworkers can adaptively adjust to the new needs of clients before new and updated Service Plans are drawn up for the family, LocalLLMs would need to draw on up-to-date child welfare goals to be able to track a family's progress in addressing child welfare-related concerns.

8.1.3 Practitioner-Grounded Methodological Takeaways for Building LLM-driven Decision-Supports. Our study's methodological steps illustrate how researchers can examine the ways language models might support social-work decision-making. Through participatory collaborations with public sector practitioners, the impetus for this study was motivated by the child welfare agency's commitment to improving decision-making processes. Our methodological decisions to categorize cases based on distributional patterns of a family's average engagement with the agency (i.e., short, medium, long, and extremely long cases) was guided by the agency's operational interests; we defined case progress in accordance with the provincial child welfare ministry's policy guidelines (as recorded in Service Plans); and traced case journeys through the documented lens of frontline workers. Through these research steps, our comparison of the LocalLLM's ability to identify progress against manual labels assessed how well the tools capture progress as articulated by the workers' documentation practices, which themselves are shaped by organizational mandates and provincial child welfare policies. Inspired by Antoniak et al. [5] and Saxena et al. [111], we also showcase a portable method to compare case trajectories over normalized time periods for different types of cases (Section 6.2.1).

8.2 The Role of LLMs in the Public Sector (RQ2)

8.2.1 LLMs as Diagnostic Tools that Center Human Discretion. In our study, we find LLMs can be useful for public sector practitioners as diagnostic tools to elicit case trajectories and relevant information for workers [1]. We also note the fallibility of ascribing adjudicative authority to LLMs when determining progress-relevance in documents. As seen from the decreasing Kappa coefficient in our results in Table 5 in Section 6, our LocalLLM could not correctly identify Activity-relevant casenotes in complex, longer-duration cases because Activity-relevance is dependent on temporally changing clients' circumstances and because what child welfare professionals consider to be Activity-relevant is based on contextual interpretations of a client's ecological environment [34, 105]. In the LLM research space, one could classify the work of child welfare caseworkers as subjective decision-making [6]; however, that would be an inaccurate description. **The crux of high-stakes public sector service work is that work in this space is inherently uncertain. The role of social workers is to 1) gauge and navigate these uncertainties based on best practices in social work scholarship and 2) follow established regulatory practices so that decision-making in this space is defensible and transparent [20, 38, 111].** In the case of child welfare, prior work by Saxena et al. [110] showed that a confluence of systemic, procedural, and risk factors often underpin complex child welfare cases, which can confound caseworkers' decision-making, and these uncertainties can persist even after case closure. Traditionally, when such

ambiguities arise, child welfare workers have engaged in collaborative meetings with other caseworkers (including their supervisors) to holistically evaluate a family's ecological environment following child protection regulatory guidelines [34, 105]. **Our findings show that LLMs, on their own, cannot and do not replicate child welfare caseworkers' iterative, discursive, and procedural decision-making processes.** And yet, we also note that our practitioner co-authors found value in the LLMs' potential as decision-support tools for inferring important child welfare signals and case trajectories from documents, particularly compared to contemporary predictive risk model that generate point predictions [72, 74, 79, 104, 117]. Therefore, we argue any bureaucratic decision-support tools that employ LLMs should be required to center human discretion (i.e., specialized expertise, value judgements, and heuristic reasoning) [105] whether that involves human labeling or oversight around the design and use of these tools.

8.2.2 Designing Public-Sector LLMs Around Local Governance Practices. Our findings underscore the importance of prioritizing and ensuring effective local governance when utilizing LLMs in the public sector. Currently, most welfare systems lack the resources to develop AI tools in-house and thus often acquire them from private companies [66]. Procuring AI technologies can result in power imbalances between public sector agencies and companies. Restrictive procurement contracts can make it difficult for workers to seek information about the model, and workers may come to over-rely on AI technologies, leading to private interests bleeding into public governance practices [56, 98]. These imbalances may be particularly severe for public administrations that lack infrastructural support to accommodate community and practitioner engagement in the AI design/deployment process. To mitigate the private sector co-opting public sector AI governance, we argue 1) local data stewardship should be prioritized by ensuring LLMs can be run locally; 2) impacted stakeholders and practitioners should be actively engaged with the AI development process early on [56]; and 3) the onus of justifying the impact and harms that arise from the AI tool should fall on developers [127].

We argue that ensuring local governance first practices is particularly critical in the adoption of LLMs for the public sector because these tools may be vulnerable to many of the same limitations as contemporary predictive risk algorithms, namely 1) the reductive representation of complex phenomena and 2) focus on individualizing societal issues while ignoring systemic constraints [58, 76, 97]. The same neoliberal austerity policies that gave rise to public sector algorithmic tools are driving the adoption of LLMs in public agencies, i.e., the desire to do more with fewer resources. For example, the US Department of Homeland Security's pilot projects that will apply LLMs in public agencies state these tools will promote efficiency, and improve the quality of public service work with less resources [113]. A similar rhetoric appears in public statements from the Canadian federal government when announcing its partnership with Cohere, a Canadian LLM developer [89]. If austerity policies drive the underlying impetus to apply LLMs, these tools could simply enable public agencies to better control access to public services for those deemed eligible, while overlooking individuals who fall through policy gaps and justifying further cuts to essential public programs [33]. While we acknowledge LLMs hold

remarkable potential to transform bureaucratic decision-making, we contend that recent interest in LLMs give rise to one of the oldest problems in HCI research. As Vold et al. [126] articulated over a decade ago, "even when core values align, tensions may still exist about how to achieve desired ends, or what these values mean in practice." Everyone likely wants to improve public service delivery, however how to do so, we argue, requires further inquiry through collaborations between practitioners and researchers in HCI, computer science, critical data studies, social work, public policy, and beyond.

8.3 Implications for HCI Research in the Public Sector (RQ3)

In this section, we propose recommendations for how HCI researchers can engage with public sector practitioners to design LLMs that support their work.

8.3.1 Building Ground-Up Participatory Partnerships with Public Agencies. We argue that **HCI researchers should work directly with public sector agencies on issues of key relevance for the organization.** The SIGCHI community has a long history of engaging in participatory methods to promote sustainable human-AI partnerships and supporting the responsible use and deployment of AI tools within the public sector [24, 47, 52, 59, 60, 84, 123, 128]. However, oftentimes these works have been conducted through short-term collaborations such as through design workshops that discuss hypothetical AI-related scenarios and interview studies to elicit stakeholder preferences and perceptions [26, 61]. While these studies are useful for understanding how AI tools should be used and designed, they are often extractive and do not necessarily lead to meaningful change for public sector stakeholders [12]. There is a greater need for HCI scholars to collaborate directly with public sector organizations to provide critical perspectives on how to design LLM-driven tools [7, 91, 115, 124], particularly because agencies cannot build LLMs in-house [66]. Due to the substantial computational resources, energy, and expertise required to build LLMs, most LLMs are built on foundational models developed by a handful of technology companies (in our study, we likewise relied on an off-the-shelf model, Llama 3.1, developed by Meta AI [129]). As North American federal governments aggressively seek to adopt AI tools to increase governmental efficiencies amidst austerity measures and techno-optimism, there lies a real concern that government services may be shaped by the interests of big technology companies as government workers increasingly rely on these tools [98, 130]. As such, there is renewed urgency for HCI scholars to collaborate directly with public sector organizations to provide critical perspectives on how to design AI tools that are centered on human experiences and needs [7, 91, 115, 124].

We also argue that **HCI researchers should strive to engage in sustainable long-term community-based research [26, 61]** as **short-term partnerships may be insufficient for fostering meaningful social change.** In this study, academics and child welfare professionals came together to investigate a question of interest for a large child welfare agency. Through our study, we showed that current off-the-shelf tools are unable to correctly identify progress-relevance in Regular casenotes for complex cases, which are precisely areas in which caseworkers would want the

most support. We also realized that our findings generated further research directions that required further examination. For example, our agency co-authors thought it particularly interesting that Extremely long cases had fewer Activity-relevant casenote entries towards the latter half of cases because new issues unrelated to Service Plan Activities emerged (Sections 6 and 7). Our co-authors reflected that the longer a family remains involved with the agency, the more likely it is that new child welfare concerns will arise. They suggested that future work could assess whether such cases should have been closed earlier or whether families required further services, and explore whether LLMs could detect Service Plan deviations in Regular casenotes to prompt workers to revisit case goals and assess their own practice. Moreover, our agency co-authors noted that for our study's methodological approach to generate actionable insights, the Super Themes of Table 3 needed to be broken down further and tracked using the methodology of Section 7. For example, currently, we have grouped substance abuse and counseling-related topics within the "*Medical/Mental Health*" theme (Section 5), but agency workers reflected that these two topics could be separated out because substance abuse involves varying levels of need and counseling encompasses a wide range of services. HCI scholarship has long emphasized the importance of empowering communities in participatory research by placing them in the driver's seat [26, 28, 61]. We provide an example of this approach and highlight the need for long-term collaborations to address substantive research questions of social relevance.

8.3.2 Operationalizing Ground-Up Partnerships for Public-Sector LLMs. In this section, we propose recommendations for how HCI researchers can engage with public sector practitioners to design LLMs that support their work.

- **We encourage collaborations with practitioners to identify areas of public sector work that can benefit from the technical capabilities of LLMs and areas where these technologies should *not* be used early on in the AI development process [11].** Oftentimes, AI harms and failures are attributed to idealistic problem formulations, where there is insufficient consideration of how AI tools can create value for practitioners [83, 108]. Through speculative co-design workshops [71], applications of innovative risk matrices [108], and exploratory research collaborations such as ours, practitioners and other stakeholders can identify contexts where resistance to automation is justified and where human discretion is necessary to augment transparent and accountable decision-making.
- **Through participatory methods, future research could identify and evaluate which forms of human–AI collaboration enhance workers' meaningfulness of public sector work while appropriately delegating tasks to LLM tools to improve overall performance.** Public sector and industry discourse on LLMs have largely emphasized its ability to improve the quality of public services delivered and efficiency gains but have paid less attention to workers' experiential aspect of working with AI tools [89, 101, 130]. Sadeghian et al. [101] recently found people prefer to work interactively with AI tools as it gives them a sense of meaningfulness in their work but that AI tools may be better at

some tasks than humans. Our agency co-authors have suggested that LLMs could be used as an interactive deliberative tool [70] in assisting workers document consistent and comprehensive details on cases (Section 8.3.1). Further work can examine how to design tools that enhance the experiential experience of workers and clients.

- **Train agency leaders and frontline staff on evaluating LLM tool performance, and work toward building consensus across managerial levels on the tool's evaluation criteria [56, 108].** For example, HCI researchers can co-develop customized toolkits with practitioners that allow workers to assess how LLM tools add value to their work. Instead of traditional methods of evaluating AI based on accuracy metrics, these toolkits should examine tradeoffs that can emerge between multiple evaluation metrics, such as model performance, data quality, and errors [108]. Greater literacy on how to identify AI issues will allow workers to contest AI-driven decisions and reduce over-reliance on these tools [55]. Furthermore, our study's findings showed that management and frontline workers can differ in what they consider to be Activity-relevant. Accordingly, such evaluation toolkits should be co-developed through organization-wide involvement, including leadership teams and frontline staff.

8.4 Limitations and Future Work

Our study only draws on casenotes from one child welfare agency in Canada, so our findings may not be generalizable across other child welfare systems in North America, which are governed and regulated by other laws and policies. Even so, we believe our methodological approach can be applied across other child welfare systems, as most agencies draw up case plans that outline goals that a family will work towards and document regular casenotes. Furthermore, it is important to remember that Regular casenotes contain caseworker perceptions and may reflect worker biases and omit information pertinent to signaling case progress. Future work can examine how our study findings align and differ across other child welfare systems, as well as other welfare systems. Furthermore, in our work, we applied Service Plan Activities to track case progress in Regular casenotes. Following Ministry guidelines, these goals are standardized in language and are intended to address child welfare concerns surrounding family/child safety, permanency, and well-being. Further research should investigate how comprehensively these goals address clients' child welfare concerns and if technical interventions can improve how goals are defined and documented by workers. Lastly, we only applied BERTopic and Llama 3.1 in our analysis. Although we experimented with other LocalLLM models, the final analysis was conducted using only Llama 3.1. Prior research has shown that LLMs and topic models may yield different outputs depending on the model [6, 32], and LLM performance may be affected by prompt brittleness issues [87]. Future work should validate our findings by comparing the performance of additional models and experimenting with different prompt parameters.

9 CONCLUSION

We conducted a study in collaboration with a large Canadian child welfare agency to examine how computational tools can be used

to infer child welfare progress. Using two types of child welfare documents, a Service Plan that outlines case management goals for families and Regular casenotes that detail day-to-day information pertaining to a family's case, we explored how AI tools can support caseworkers in tracking case progress. Our findings show that LLMs can assist caseworkers trace thematic trajectories in cases. However, we find LLMs cannot correctly identify which casenotes contain information relating to child welfare goals, especially as cases become more complex, because discretionary decision-making founded on social work practices is needed. Overall, we argue that LLMs should not be used to supplant caseworker decision-making but as decision-aid tools.

ACKNOWLEDGMENTS

This research was supported by the NSERC Discovery Early Career Researcher Grant RGPIN-2022-04570, MITACS Accelerate program, and the University of Toronto Data Science Institute Summer Undergraduate Data Science Opportunities Program. Opinions, findings, and conclusions expressed in this paper are those of the authors. We sincerely thank our collaborators and anonymous reviewers whose suggestions and comments helped improve this manuscript.

REFERENCES

- [1] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G. Robinson. 2020. Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) (FAT* '20). Association for Computing Machinery, New York, NY, USA, 252–260. <https://doi.org/10.1145/3351095.3372871>
- [2] Rediet Abebe, Salvatore Giorgi, Anna Tedijanto, Anneke Buffone, and H. Andrew Schwartz. 2020. Quantifying Community Characteristics of Maternal Mortality Using Social Media. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) (WWW '20). Association for Computing Machinery, New York, NY, USA, 2976–2983. <https://doi.org/10.1145/3366423.3380066>
- [3] Meta AI. 2024. *Introducing Llama 3.1: Our most capable models to date*. Retrieved Sept 2, 2025 from <https://ai.meta.com/blog/meta-llama-3-1/>
- [4] Tawfiq Ammari, Eunhye Ahn, Astha Lakhankar, and Joyce Y. Lee. 2025. Finding Understanding and Support: Navigating Online Communities to Share and Connect at the intersection of Abuse and Foster Care Experiences. *Proc. ACM Hum.-Comput. Interact.* 9, 2, Article CSCW087 (May 2025), 40 pages. <https://doi.org/10.1145/3710985>
- [5] Maria Antoniak, David Mimno, and Karen Levy. 2019. Narrative Paths and Negotiation of Power in Birth Stories. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 88 (Nov. 2019), 27 pages. <https://doi.org/10.1145/3359190>
- [6] Paula Akemi Aoyagui, Kelsey Stemmler, Sharon A Ferguson, Young-Ho Kim, and Anastasia Kuzminykh. 2025. A Matter of Perspective(s): Contrasting Human and LLM Argumentation in Subjective Decision-Making on Subtle Sexism. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (CHI '25). Association for Computing Machinery, New York, NY, USA, Article 529, 16 pages. <https://doi.org/10.1145/3706598.3713248>
- [7] Cecilia Aragon, Shion Guha, Marina Kogan, Michael Muller, and Gina Neff. 2022. *Human-Centered Data Science: An Introduction*. MIT Press.
- [8] Blair Attard-Frost, Ana Brandusescu, and Kelly Lyons. 2024. The governance of artificial intelligence in Canada: Findings and opportunities from a review of 84 AI governance initiatives. *Government Information Quarterly* 41, 2 (2024), 101929. <https://doi.org/10.1016/j.giq.2024.101929>
- [9] Karla Badillo-Urquiza, Zainab Agha, Denielle Abaqita, Scott B. Harpin, and Pamela J. Wisniewski. 2024. Towards a Social Ecological Approach to Supporting Caseworkers in Promoting the Online Safety of Youth in Foster Care. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 135 (April 2024), 28 pages. <https://doi.org/10.1145/3637412>
- [10] Karla Badillo-Urquiza, Xinru Page, and Pamela J. Wisniewski. 2019. Risk vs. Restriction: The Tension between Providing a Sense of Normalcy and Keeping Foster Teens Safe Online. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) (CHI '19). Association for Computing Machinery, New York, NY, USA, 1–14. <https://doi.org/10.1145/3290605.3300497>
- [11] Eric P.S. Baumer and M. Six Silberman. 2011. When the Implication is Not to Design (Technology). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) (CHI '11). Association for Computing Machinery, New York, NY, USA, 2271–2274. <https://doi.org/10.1145/1978942.1979275>
- [12] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the People? Opportunities and Challenges for Participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Arlington, VA, USA) (EAAMO '22). Association for Computing Machinery, New York, NY, USA, Article 6, 8 pages. <https://doi.org/10.1145/3551624.3555290>
- [13] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3, null (March 2003), 993–1022.
- [14] Jessica Y Bo, Sophia Wan, and Ashton Anderson. 2025. To Rely or Not to Rely? Evaluating Interventions for Appropriate Reliance on Large Language Models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (CHI '25). Association for Computing Machinery, New York, NY, USA, Article 905, 23 pages. <https://doi.org/10.1145/3706598.3714097>
- [15] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- [16] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. 2019. Toward Algorithmic Accountability in Public Services: A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-making in Child Welfare Services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) (CHI '19). Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3290605.3300271>
- [17] Johanna Creswell Báez, Arlene Bjugstad, Tae Kyung Park, Jacqueline L. Jones, Laurel N. Bidwell, Melanie Sage, and Laurel Iverson Hitchcock. 2025. Social Work Educators Innovating With Generative AI: An Exploratory Study. *Journal of Social Work Education* 61, 1 (2025), 14–29. <https://doi.org/10.1080/10437797.2024.2411170>
- [18] Xavier V. Caddle, Nurun Naher, Zachary P. Miller, Karla Badillo-Urquiza, and Pamela J. Wisniewski. 2022. Duty to Respond: The Challenges Social Service Providers Face When Charged with Keeping Youth Safe Online. *Proc. ACM Hum.-Comput. Interact.* 7, GROUP, Article 6 (Dec. 2022), 35 pages. <https://doi.org/10.1145/3567556>
- [19] Stevie Chancellor, Zhiyuan Lin, Erica L. Goodman, Stephanie Zerwas, and Munmun De Choudhury. 2016. Quantifying and Predicting Mental Illness Severity in Online Pro-Eating Disorder Communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, California, USA) (CSCW '16). Association for Computing Machinery, New York, NY, USA, 1171–1184. <https://doi.org/10.1145/2818048.2819973>
- [20] Charalampos Chelmis, Wenting Qi, and Wonhyung Lee. 2021. Challenges and Opportunities in Using Data Science for Homelessness Service Provision. In *Companion Proceedings of the Web Conference 2021 (WWW '21)*. Association for Computing Machinery, New York, NY, USA, 128–135. <https://doi.org/10.1145/3442442.3453454>
- [21] Hao-Fei Cheng, Logan Stapleton, Anna Kawakami, Venkatesh Sivaraman, Yanghuidi Cheng, Diana Qing, Adam Perer, Kenneth Holstein, Zhiwei Steven Wu, and Haiyi Zhu. 2022. How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 162, 22 pages. <https://doi.org/10.1145/3491102.3501831>
- [22] Hao-Fei Cheng, Logan Stapleton, Ruiqi Wang, Paige Bullock, Alexandra Chouldechova, Zhiwei Steven Wu, and Haiyi Zhu. 2021. Soliciting Stakeholders' Fairness Notions in Child Maltreatment Predictive Systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 390, 17 pages. <https://doi.org/10.1145/3411764.3445308>
- [23] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 134–148. <https://proceedings.mlr.press/v81/chouldechova18a.html>
- [24] Victoria Chui, Kelly McConvey, Erina Seh-Young Moon, Maya Ghai, and Shion Guha. 2025. Towards Sustainable Community-Designed AI Systems in the Public Sector. In *Proceedings of the 2025 ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS '25)*. Association for Computing Machinery, New York, NY, USA, 837–840. <https://doi.org/10.1145/3715335.3737683>
- [25] Paul Clarkson. 2008. Performance Measurement in Adult Social Care: Looking Backwards and Forwards. *The British Journal of Social Work* 40, 1 (June 2008), 170–187. <https://doi.org/10.1093/bjsw/bcn096>
- [26] Ned Cooper, Tiffanie Horne, Gillian R Hayes, Courtney Heldreth, Michal Lahav, Jess Holbrook, and Lauren Wilcox. 2022. A Systematic Review and Thematic Analysis of Community-Collaborative Approaches to Computing Research. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*

- (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 73, 18 pages. <https://doi.org/10.1145/3491102.3517716>
- [27] Victoria A Copeland. 2021. “It’s the Only System We’ve Got”: Exploring Emergency Response Decision-Making in Child Welfare. *Columbia Journal of Race and Law* 11, 3 (2021), 43–74.
- [28] Sasha Costanza-Chock. 2020. *Design Justice: Community-Led Practices to Build the Worlds We Need*. The MIT Press. <https://doi.org/10.7551/mitpress/12255.001.0001>
- [29] John W. Creswell and Dana L. Miller. 2000. Determining Validity in Qualitative Inquiry. *Theory Into Practice* 39, 3 (2000), 124–130.
- [30] Maria De-Arteaga, Riccardo Fogliato, and Alexandre Chouldechova. 2020. A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3313831.3376638>
- [31] Elizabeth Dwoskin, Jeff Stein, and Josh Dawsey. 2024. *Inside Elon Musk’s vision to remake government: ‘Delete, delete, delete’*. Retrieved Sept 2, 2025 from <https://www.washingtonpost.com/technology/2024/11/01/elon-musk-cuts-government-trump-election/>
- [32] Omar El-Gayar, Mohammad Al-Ramahi, Abdullah Wahbeh, Tareq Nasralah, and Ahmed Elnoshokaty. 2024. A comparative analysis of the interpretability of lda and llm for topic modeling: The case of healthcare apps. (2024).
- [33] Virginia Eubanks. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin’s Press.
- [34] S Ferguson. 2009. Clinical supervision in child welfare. In *Child welfare supervision*, C C R Potter C (Ed.). Oxford University Press, New York, NY, 296–329.
- [35] Anjalie Field, Amanda Coston, Nupoor Gandhi, Alexandra Chouldechova, Emily Putnam-Hornstein, David Steier, and Yulia Tsvetkov. 2023. Examining risks of racial biases in NLP tools for child protective services. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) (FAccT '23). Association for Computing Machinery, New York, NY, USA, 1479–1492. <https://doi.org/10.1145/3593013.3594094>
- [36] John D. Fluke, Tyler W. Corwin, Dana M. Hollinshead, and Erin J. Maher. 2016. Family preservation or child safety? Associations between child welfare workers’ experience, position, and perspectives. *Children and Youth Services Review* 69 (2016), 210–218. <https://doi.org/10.1016/j.childyouth.2016.08.012>
- [37] Sarah Fox, Jill Dimond, Lilly Irani, Tad Hirsch, Michael Muller, and Shaowen Bardzell. 2017. Social Justice and Design: Power and oppression in collaborative systems. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) (CSCW '17 Companion). Association for Computing Machinery, New York, NY, USA, 117–122. <https://doi.org/10.1145/3022198.3022201>
- [38] Jennifer M Geiger and Lisa Schelbe. 2021. Assessment in Child Welfare Practice. In *The Handbook on Child Welfare Practice*. Springer, 195–217.
- [39] Marissa Gerchick, Tobi Jegede, Tarak Shah, Ana Gutierrez, Sophie Beiers, Noam Shemtov, Kath Xu, Anjana Samant, and Aaron Horowitz. 2023. The Devil is in the Details: Interrogating Values Embedded in the Allegheny Family Screening Tool. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) (FAccT '23). Association for Computing Machinery, New York, NY, USA, 1292–1310. <https://doi.org/10.1145/3593013.3594081>
- [40] Github. 2021. *langdetect*. Retrieved Sept 2, 2025 from <https://github.com/Mimino666/langdetect>
- [41] Github. 2024. *Ollama: A lightweight, extensible framework for building and running language models*. Retrieved Sept 2, 2025 from <https://github.com/ollama/ollama>
- [42] Apoorva Gondimalla, Varshinee Sreekanth, Govind Joshi, Whitney Nelson, Eunsol Choi, Stephen C. Slota, Sherri R. Greenberg, Kenneth R. Fleischmann, and Min Kyung Lee. 2024. Aligning Data with the Goals of an Organization and Its Workers: Designing Data Labeling for Social Service Case Notes. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 765, 21 pages. <https://doi.org/10.1145/3613904.3642014>
- [43] Mitchell L. Gordon, Michelle S. Lam, Joont Sung Park, Kayur Patel, Jeff Hancock, Tatsumori Hashimoto, and Michael S. Bernstein. 2022. Jury Learning: Integrating Dissenting Voices into Machine Learning Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 115, 19 pages. <https://doi.org/10.1145/3491102.3502004>
- [44] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, and Anthony et al. Hartshorn. 2024. The Llama 3 Herd of Models. <https://doi.org/10.48550/arXiv.2407.21783>
- [45] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* (2022).
- [46] Brett A. Halperin, Gary Hsieh, Erin McElroy, James Pierce, and Daniela K. Rosner. 2023. Probing a Community-Based Conversational Storytelling Agent to Document Digital Stories of Housing Insecurity. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 304, 18 pages. <https://doi.org/10.1145/3544548.3581109>
- [47] MD Romael Haque, Devansh Saxena, Katy Weathering, Joseph Chudzik, and Shion Guha. 2024. Are We Asking the Right Questions?: Designing for Community Stakeholders’ Interactions with AI in Policing. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 301, 20 pages. <https://doi.org/10.1145/3613904.3642738>
- [48] Illugi Torfason Hjaltalin and Hallur Thor Sigurdarson. 2024. The strategic use of AI in the public sector: A public values analysis of national AI strategies. *Government Information Quarterly* 41, 1 (2024), 101914. <https://doi.org/10.1016/j.giq.2024.101914>
- [49] The White House. 2025. *ESTABLISHING AND IMPLEMENTING THE PRESIDENT’S “DEPARTMENT OF GOVERNMENT EFFICIENCY”*. Retrieved Sept 2, 2025 from <https://www.whitehouse.gov/presidential-actions/2025/01/establishing-and-implementing-the-presidents-department-of-government-efficiency/>
- [50] Saily Huusonen and Pertti Vakkari. 2015. Selective Clients’ Trajectories in Case Files: Filtering Out Information in the Recording Process in Child Protection. *The British Journal of Social Work* 45, 3 (April 2015), 792–808. <https://doi.org/10.1093/bjsw/bct160>
- [51] Angela Jin and Niloufar Salehi. 2024. (Beyond) Reasonable Doubt: Challenges that Public Defenders Face in Scrutinizing AI in Court. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 467, 19 pages. <https://doi.org/10.1145/3613904.3641902>
- [52] Eunkyun Jo, Young-Ho Kim, Sang-Houn Ok, and Daniel A. Epstein. 2025. Understanding Public Agencies’ Expectations and Realities of AI-Driven Chatbots for Public Health Monitoring. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (CHI '25). Association for Computing Machinery, New York, NY, USA, Article 951, 17 pages. <https://doi.org/10.1145/3706598.3713593>
- [53] Rebecca Ann Johnson and Simone Zhang. 2022. What is the Bureaucratic Counterfactual? Categorical versus Algorithmic Prioritization in U.S. Social Policy. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) (FAccT '22). Association for Computing Machinery, New York, NY, USA, 1671–1682. <https://doi.org/10.1145/3531146.3533223>
- [54] Anja Karadeglij. 2025. *Federal government taps Cohere to work on use of AI in public service*. Retrieved Sept 2, 2025 from <https://www.cbc.ca/news/politics/cohere-ai-public-service-1.7613424>
- [55] Naveena Karusala, Sohini Upadhyay, Rajesh Veeraraghavan, and Krzysztof Z. Gajos. 2024. Understanding Contestability on the Margins: Implications for the Design of Algorithmic Decision-making in Public Services. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 478, 16 pages. <https://doi.org/10.1145/3613904.3641898>
- [56] Anna Kawakami, Amanda Coston, Hodu Heidari, Kenneth Holstein, and Haiyi Zhu. 2024. Studying Up Public Sector AI: How Networks of Power Relations Shape Agency Decisions Around AI Design and Use. *arXiv:2405.12458* (May 2024). <https://doi.org/10.48550/arXiv.2405.12458> arXiv:2405.12458 [cs].
- [57] Anna Kawakami, Venkatesh Sivarajan, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. Improving Human-AI Partnerships in Child Welfare: Understanding Worker Practices, Challenges, and Desires for Algorithmic Decision Support. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (CHI '22). Association for Computing Machinery, New York, NY, USA, 1–18. <https://doi.org/10.1145/3491102.3517439>
- [58] Emily Kedell. 2015. The ethics of predictive risk modelling in the Aotearoa/New Zealand child welfare context: Child abuse prevention or neo-liberal tool? *Critical Social Policy* 35, 1 (2015), 69–88. <https://doi.org/10.1177/0261018314532224>
- [59] Seyun Kim, Bonnie Fan, Willa Yunqi Yang, Jessie Ramey, Sarah E. Fox, Haiyi Zhu, John Zimmerman, and Motahareh Eslami. 2024. Public Technologies Transforming Work of the Public and the Public Sector. In *Proceedings of the 3rd Annual Meeting of the Symposium on Human-Computer Interaction for Work* (Newcastle upon Tyne, United Kingdom) (CHIWOR '24). Association for Computing Machinery, New York, NY, USA, Article 20, 12 pages. <https://doi.org/10.1145/3663384.3663407>
- [60] Seyun Kim, Jonathan Ho, Yinan Li, Bonnie Fan, Willa Yunqi Yang, Jessie Ramey, Sarah E. Fox, Haiyi Zhu, John Zimmerman, and Motahareh Eslami. 2024. Integrating Equity in Public Sector Data-Driven Decision Making: Exploring the Desired Futures of Underserved Stakeholders. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 366 (Nov. 2024), 39 pages. <https://doi.org/10.1145/3686905>
- [61] Yasmine Kotturi, Julie Hui, TJ Johnson, Lutalo Sanifu, and Tawanna R. Dillahunt. 2024. Sustaining Community-Based Research in Computing: Lessons from Two Tech Capacity Building Initiatives for Local Businesses. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 178 (April 2024), 31 pages. <https://doi.org/10.1145/3641017>

- [62] Tzu-Sheng Kuo, Hong Shen, Jisoo Geum, Nev Jones, Jason I. Hong, Haiyi Zhu, and Kenneth Holstein. 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 (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–17. <https://doi.org/10.1145/3544548.3580882>
- [63] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, and Ariel D. Procaccia. 2019. WeBuildAI: Participatory Framework for Algorithmic Governance. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 181 (Nov. 2019), 35 pages. <https://doi.org/10.1145/3359283>
- [64] Tuukka Lehtiniemi. 2024. Contextual social valences for artificial intelligence: anticipation that matters in social work. *Information, Communication & Society* 27, 6 (2024), 1110–1125. <https://doi.org/10.1080/1369118X.2023.2234987>
- [65] Kelly LeRoux and Nathaniel S. Wright. 2010. Does Performance Measurement Improve Strategic Decision Making? Findings From a National Survey of Non-profit Social Service Agencies. *Nonprofit and Voluntary Sector Quarterly* 39, 4 (Aug. 2010), 571–587. <https://doi.org/10.1177/0899764009359942>
- [66] Karen Levy, Kyla E. Chaslaw, and Sarah Riley. 2021. Algorithms and Decision-Making in the Public Sector. *Annual Review of Law and Social Science* 17, 1 (Oct. 2021), 309–334. <https://doi.org/10.1146/annurev-lawsocsci-041221-023808>
- [67] Michael Lipsky. 2010. *Street-level bureaucracy: Dilemmas of the individual in public service*. Russell Sage Foundation.
- [68] Lydia T. Liu, Solon Barocas, Jon Kleinberg, and Karen Levy. 2024. On the actionability of outcome prediction. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'24/IAAI'24/EAII'24, Vol. 38)*. AAAI Press, 22240–22249. <https://doi.org/10.1609/aaai.v38i20.30229>
- [69] Lydia T. Liu, Inioluwa Deborah Raji, Angela Zhou, Luke Guerdan, Jessica Hullman, Daniel Malinsky, Bryan Wilder, Simone Zhang, Hammaad Adam, Amanda Coston, Ben Laufer, Ezinne Nwankwo, Michael Zanger-Tishler, Eli Ben-Michael, Solon Barocas, Avi Feller, Marissa Gerchick, Talia Gillis, Shion Guha, Daniel Ho, Lily Hu, Kosuke Imai, Sayash Kapoor, Joshua Loftus, Razieh Nabi, Arvind Narayanan, Ben Recht, Juan Carlos Perdomo, Matthew Salganik, Mark Sendak, Alexander Tolbert, Berk Ustun, Suresh Venkatasubramanian, Angelina Wang, and Ashia Wilson. 2025. Bridging Prediction and Intervention Problems in Social Systems. arXiv:2507.05216 [cs.LG]. <https://arxiv.org/abs/2507.05216>
- [70] Shuai Ma, Qiaoyi Chen, Xinru Wang, Chengbo Zheng, Zhenhui Peng, Ming Yin, and Xiaojuan Ma. 2025. Towards Human-AI Deliberation: Design and Evaluation of LLM-Empowered Deliberative AI for AI-Assisted Decision-Making. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–23. <https://doi.org/10.1145/3706598.3713423>
- [71] Narmene Marij, Mattia Thibault, and Janset Shawash. 2023. Blueprints of Tomorrow: Co-Designing Speculative Urban Futures. In *Proceedings of the 26th International Academic Mindtrek Conference (Tampere, Finland) (Mindtrek '23)*. Association for Computing Machinery, New York, NY, USA, 305–308. <https://doi.org/10.1145/3616961.3616980>
- [72] Kelly McConvey and Shion Guha. 2024. Designing for Fairness in Higher Education Early Warning Systems.. In *Canadian AI*.
- [73] Kelly McConvey and Shion Guha. 2024. "This is not a data problem": Algorithms and Power in Public Higher Education in Canada. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 16, 14 pages. <https://doi.org/10.1145/3613904.3642451>
- [74] Kelly McConvey, Shion Guha, and Anastasia Kuzminykh. 2023. A Human-Centered Review of Algorithms in Decision-Making in Higher Education. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (Hamburg, Germany) (CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 223, 15 pages. <https://doi.org/10.1145/3544548.3580658>
- [75] Leland McInnes, John Healy, and James Melville. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Retrieved August 31, 2024 from <https://arxiv.org/abs/1802.03426>
- [76] Rebecca Mead, Miranda Thurston, and Daniel Bloyce. 2022. From public issues to personal troubles: individualising social inequalities in health within local public health partnerships. *Critical Public Health* 32, 2 (2022), 168–180. <https://doi.org/10.1080/09581596.2020.1763916>
- [77] Darcey H. Merritt. 2020. How Do Families Experience and Interact with CPS? *The ANNALS of the American Academy of Political and Social Science* 692, 1 (2020), 203–226. <https://doi.org/10.1177/0002716220979520> PMID: 38075400. arXiv:<https://doi.org/10.1177/0002716220979520> PMID: 38075400.
- [78] Mario Molina and Filiz Garip. 2019. Machine Learning for Sociology. *Annual Review of Sociology* 45, 1 (2019), 27–45.
- [79] Erina Seh-Young Moon and Shion Guha. 2024. A Human-Centered Review of Algorithms in Homelessness Research. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 870, 15 pages. <https://doi.org/10.1145/3613904.3642392>
- [80] Erina Seh-Young Moon, Devansh Saxena, Dipto Das, and Shion Guha. 2025. The Datafication of Care in Public Homelessness Services. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 829, 16 pages. <https://doi.org/10.1145/3706598.3713232>
- [81] Erina Seh-Young Moon, Devansh Saxena, Tegan Maharaj, and Shion Guha. 2024. Beyond Predictive Algorithms in Child Welfare. In *Proceedings of the 50th Graphics Interface Conference (Halifax, NS, Canada) (GI '24)*. Association for Computing Machinery, New York, NY, USA, Article 37, 13 pages. <https://doi.org/10.1145/3670947.3670976>
- [82] Therese Moreau, Roberta Sinatra, and Vedran Sekara. 2024. Failing Our Youngest: On the Biases, Pitfalls, and Risks in a Decision Support Algorithm Used for Child Protection. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (Rio de Janeiro, Brazil) (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 290–300. <https://doi.org/10.1145/3630106.3658906>
- [83] Ramaravind Kommiya Mothilal, Shion Guha, and Syed Ishtiaque Ahmed. 2024. Towards a Non-Ideal Methodological Framework for Responsible ML. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 477, 17 pages. <https://doi.org/10.1145/3613904.3642501>
- [84] Ramaravind Kommiya Mothilal, Faisal M. Lalani, Syed Ishtiaque Ahmed, Shion Guha, and Sharifa Sultana. 2025. Talking About the Assumption in the Room. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 490, 16 pages. <https://doi.org/10.1145/3706598.3713958>
- [85] Whitney Nelson, Min Kyung Lee, Eunsol Choi, and Victor Wang. 2024. Designing LLM-Based Support for Homelessness Caseworkers. In *AAAI-2024 Workshop on Public Sector LLMs: Algorithmic and Sociotechnical Design*. <https://openreview.net/forum?id=XyLgJLwT5w>
- [86] Dong Nguyen, Maria Liakata, Simon DeDeo, Jacob Eisenstein, David Mimno, Rebekah Tromble, and Jane Winters. 2020. How We Do Things With Words: Analyzing Text as Social and Cultural Data. *Frontiers in Artificial Intelligence* 3 (2020), 62.
- [87] Lilian Ngweta, Kiran Kate, Jason Tsay, and Yara Rizk. 2025. Towards LLMs Robustness to Changes in Prompt Format Styles. arXiv:2504.06969 [cs.CL]. <https://arxiv.org/abs/2504.06969>
- [88] Trine Rask Nielsen, María Menéndez-Blanco, and Naja Holten Møller. 2023. Who Cares About Data? Ambivalence, Translation, and Attentiveness in Asylum Casework. *Computer Supported Cooperative Work (CSCW)* 32, 4 (Dec. 2023), 861–910. <https://doi.org/10.1007/s10606-023-09474-7>
- [89] Government of Canada. 2025. *Canada partners with Cohere to accelerate world-leading artificial intelligence*. Retrieved Sept 2, 2025 from <https://www.canada.ca/en/innovation-science-economic-development/news/2025/08/canada-partners-with-cohere-to-accelerate-world-leading-artificial-intelligence.html>
- [90] Government of Canada. 2025. *Ottawa drafting public registry of AI projects as tech spreads through government*. Retrieved Sept 2, 2025 from https://www.thestar.com/politics/federal/ottawa-drafting-public-registry-of-ai-projects-as-tech-spreads-through-government/article_f5bfdd9f-aecd-5b45-88b6-9a734943658c.html
- [91] Ihudu Ifinda Ogbonnaya-Ogburu, Angela D.R. Smith, Alexandra To, and Kentaro Toyama. 2020. Critical Race Theory for HCI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–16. <https://doi.org/10.1145/3313831.3376392>
- [92] Rock Yuren Pang, Hope Schroeder, Kynnedy Simone Smith, Solon Barocas, Ziang Xiao, Emily Tseng, and Danielle Bragg. 2025. Understanding the LLMification of CHI: Unpacking the Impact of LLMs at CHI through a Systematic Literature Review. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 456, 20 pages. <https://doi.org/10.1145/3706598.3713726>
- [93] Brian E. Perron, Barbara S. Hiltz, Erin M. Khang, and Sue Ann Savas. 2025. AI-Enhanced Social Work: Developing and Evaluating Retrieval-Augmented Generation (RAG) Support Systems. *Journal of Social Work Education* 61, 1 (Jan. 2025), 3–13. <https://doi.org/10.1080/10437797.2024.2411172>
- [94] Brian E. Perron, Hui Luan, Bryan G. Victor, Oliver Hiltz-Perron, and Joseph Ryan. 2025. Moving Beyond ChatGPT: Local Large Language Models (LLMs) and the Secure Analysis of Confidential Unstructured Text Data in Social Work Research. *Research on Social Work Practice* 35, 6 (2025), 695–710. <https://doi.org/10.1177/10497315241280686>
- [95] Brian E. Perron, Kelley A. Rivenburgh, Bryan G. Victor, Zia Qi, and Hui Luan. 2025. A Primer on Word Embeddings: AI Techniques for Text Analysis in Social Work. *Journal of the Society for Social Work and Research* 16, 2 (2025), 241–273. <https://doi.org/10.1086/735577>
- [96] Andrew Pithouse, Karen Broadhurst, Chris Hall, Sue Peckover, Dave Wastell, and Sue White. 2012. Trust, risk and the (mis)management of contingency

- and discretion through new information technologies in children's services. *Journal of Social Work* 12, 2 (March 2012), 158–178. <https://doi.org/10.1177/1468017310382151>
- [97] Joanna Redden. 2020. Predictive Analytics and Child Welfare: Toward Data Justice. *Canadian Journal of Communication* 45, 1 (2020), 101–111. <https://doi.org/10.22230/cjc.2020v45n1a3479>
- [98] Joanna Redden, Lina Dencik, and Harry Warne. 2020. Datafied child welfare services: unpacking politics, economics and power. *Policy Studies* 41, 5 (Sept. 2020), 507–526. <https://doi.org/10.1080/01442872.2020.1724928>
- [99] Tyler Reimmund, Lars Kunze, and Marina Denise Jirocka. 2024. Transitioning Towards a Proactive Practice: A Longitudinal Field Study on the Implementation of a ML System in Adult Social Care. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–19. <https://doi.org/10.1145/3613904.3642247>
- [100] Rashida Richardson. 2021. Defining and Demystifying Automated Decision Systems. *Maryland Law Review* 3811708 (2021). https://papers.ssrn.com/sol/papers.cfm?abstract_id=3811708
- [101] Shadan Sadeghian, Alarith Uhde, and Marc Hassenzahl. 2024. The Soul of Work: Evaluation of Job Meaningfulness and Accountability in Human-AI Collaboration. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1 (April 2024), 130:1–130:26. <https://doi.org/10.1145/3637407>
- [102] Samuel Salovaara and Katri Ylönen. 2022. Client information systems' support for case-based social work: experiences of Finnish social workers. *Nordic Social Work Research* 12, 3 (July 2022), 364–378. <https://doi.org/10.1080/2156857X.2021.1999847>
- [103] Anjana Samant, Aaron Horowitz, Sophie Beiers, and Kath Xu. 2021. *Family Surveillance by Algorithm: The Rapidly Spreading Tools Few Have Heard Of*. Retrieved May 2, 2023 from <https://www.aclu.org/news/womens-rights/family-surveillance-by-algorithm-the-rapidly-spreading-tools-few-have-heard-of>
- [104] Devansh Saxena, Karla Badillo-Urquiola, Pamela J. Wisniewski, and Shion Guha. 2020. A Human-Centered Review of Algorithms used within the US Child Welfare System. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–15.
- [105] Devansh Saxena, Karla Badillo-Urquiola, Pamela J. Wisniewski, and Shion Guha. 2021. A Framework of High-Stakes Algorithmic Decision-Making for the Public Sector Developed through a Case Study of Child-Welfare. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021).
- [106] Devansh Saxena and Shion Guha. 2020. Conducting Participatory Design to Improve Algorithms in Public Services: Lessons and Challenges. In *Companion Publication of the 2020 Conference on Computer Supported Cooperative Work and Social Computing (Virtual Event, USA) (CSCW '20 Companion)*. Association for Computing Machinery, New York, NY, USA, 383–388. <https://doi.org/10.1145/3406865.3418331>
- [107] Devansh Saxena and Shion Guha. 2023. Algorithmic Harms in Child Welfare: Uncertainties in Practice, Organization, and Street-level Decision-Making. *ACM J. Responsib. Comput.* (sep 2023). <https://doi.org/10.1145/3616473>
- [108] Devansh Saxena, Ji-Youn Jung, Jodi Forlizzi, Kenneth Holstein, and John Zimmerman. 2025. AI Mismatches: Identifying Potential Algorithmic Harms Before AI Development. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–23. <https://doi.org/10.1145/3706598.3714098>
- [109] Devansh Saxena, Zoe Kahn, Erina Seh-Young Moon, Lauren Marietta Chambers, Corey Jackson, Min Kyung Lee, Motahhare Eslami, Shion Guha, Sheena Erete, Lilly Irani, Deirdre Mulligan, and John Zimmerman. 2025. Emerging Practices in Participatory AI Design in Public Sector Innovation. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 777, 7 pages. <https://doi.org/10.1145/3706599.3706727>
- [110] Devansh Saxena, Seh Young Moon, Aryan Chaurasia, Yixin Guan, and Shion Guha. 2023. Rethinking "Risk" in Algorithmic Systems Through A Computational Narrative Analysis of Casenotes in Child Welfare (CHI '23). Association for Computing Machinery, New York, NY, USA, 19 pages. <https://doi.org/10.1145/3544548.3581308>
- [111] Devansh Saxena, Seh Young Moon, Dahlia Shehata, and Shion Guha. 2022. Unpacking Invisible Work Practices, Constraints, and Latent Power Relationships in Child Welfare through Casenote Analysis (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 120, 22 pages. <https://doi.org/10.1145/3491102.3517742>
- [112] Devansh Saxena, Charles Repaci, Melanie D Sage, and Shion Guha. 2022. How to Train a (Bad) Algorithmic Caseworker: A Quantitative Deconstruction of Risk Assessments in Child Welfare. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 1–7.
- [113] Homeland Security. 2024. *Department of Homeland Security Unveils Artificial Intelligence Roadmap, Announces Pilot Projects to Maximize Benefits of Technology, Advance Homeland Security Mission*. Retrieved November 22, 2025 from <https://www.dhs.gov/archive/news/2024/03/18/department-homeland-security-unveils-artificial-intelligence-roadmap-announces#:~:text=The%20pilot%20will%20specifically%20support,need%20for%20retraining%20over%20time>
- [114] Homeland Security. 2025. *DHS Playbook for Public Sector Generative Artificial Intelligence Deployment*. Retrieved Sept 2, 2025 from https://www.dhs.gov/sites/default/files/2025-01/25_0106_ocio_dhs-playbook-for-public-sector-generative-artificial-intelligence-deployment-508-signed.pdf
- [115] Ben Shneiderman. 2022. *Human-Centered AI*. Oxford University Press. <https://doi.org/10.1093/oso/9780192845290.001.0001>
- [116] Dilruba Showkat. 2025. *Towards Algorithmic Reform: Ethical Values-Informed Tool Design and Inclusive AI/ML Literacy*. Ph. D. Dissertation.
- [117] Dilruba Showkat, Angela D. R. Smith, Wang Lingqin, and Alexandra To. 2023. "Who is the right homeless client?": Values in Algorithmic Homelessness Service Provision and Machine Learning Research. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–21. <https://doi.org/10.1145/3544548.3581010>
- [118] Jonathan B. Singer, Johanna Creswell Báez, and Juan A. Rios. 2023. AI Creates the Message: Integrating AI Language Learning Models into Social Work Education and Practice. *Journal of Social Work Education* 59, 2 (April 2023), 294–302. <https://doi.org/10.1080/10437797.2023.2189878>
- [119] G Stevenson Smith. 1988. Performance Evaluation For Nonprofits. *Nonprofit world* 6, 1 (1988), 24.
- [120] Logan Stapleton, Min Hun Lee, Diana Qing, Marya Wright, Alexandra Chouldchova, Ken Holstein, Zhiwei Steven Wu, and Haiyi Zhu. 2022. Imagining new futures beyond predictive systems in child welfare: A qualitative study with impacted stakeholders. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*. ACM, 1162–1177. <https://doi.org/10.1145/3531146.3533177>
- [121] James P Stevens. 1984. Outliers and influential data points in regression analysis. *Psychological Bulletin* 95, 2 (1984), 334.
- [122] Dragan Stoll, Samuel Wehrli, and David Lätsch. 2025. Case reports unlocked: Harnessing large language models to advance research on child maltreatment. *Child Abuse & Neglect* 160 (Feb. 2025), 107202. <https://doi.org/10.1016/j.chab.2024.107202>
- [123] Ningting Tang, Jiayin Zhi, Tzu-Sheng Kuo, Calla Kainaroi, Jeremy J. Northup, Kenneth Holstein, Haiyi Zhu, Hoda Heidari, and Hong Shen. 2024. AI Failure Cards: Understanding and Supporting Grassroots Efforts to Mitigate AI Failures in Homeless Services. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (Rio de Janeiro, Brazil) (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 713–732. <https://doi.org/10.1145/3630106.3658935>
- [124] Alexandra To, Angela D. R. Smith, Dilruba Showkat, Adinawa Adjagbodjou, and Christina Harrington. 2023. Flourishing in the Everyday: Moving Beyond Damage-Centered Design in HCI for BIPOC Communities. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference (Pittsburgh, PA, USA) (DIS '23)*. Association for Computing Machinery, New York, NY, USA, 917–933. <https://doi.org/10.1145/3563657.3596057>
- [125] Pelle Tracey, Patricia Garcia, and Ricardo Punzalan. 2023. Recordkeeping, logistics, and translation: a study of homeless services systems as infrastructure. *Archival Science* (Feb. 2023), 1–27. <https://doi.org/10.1007/s10502-023-09410-0>
- [126] Amy Voids, Lynn Dombrowski, Gillian R. Hayes, and Melissa Mazmanian. 2014. Shared values/conflicting logics: working around e-government systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. Association for Computing Machinery, New York, NY, USA, 3583–3592. <https://doi.org/10.1145/2556288.2556971>
- [127] Angelina Wang, Sayash Kapoor, Solon Barocas, and Arvind Narayanan. 2024. Against Predictive Optimization: On the Legitimacy of Decision-making Algorithms That Optimize Predictive Accuracy. *ACM J. Responsib. Comput.* 1, 1 (March 2024), 9:1–9:45. <https://doi.org/10.1145/3636509>
- [128] Katharina Weitz, Ruben Schlagowski, Elisabeth André, Maris Männiste, and Ceenu George. 2024. Explaining It Your Way - Findings from a Co-Creative Design Workshop on Designing XAI Applications with AI End-Users from the Public Sector. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 745, 14 pages. <https://doi.org/10.1145/3613904.3642563>
- [129] Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Venkat Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. 2025. LiveBench: A Challenging, Contamination-Free LLM Benchmark. In *The Thirteenth International Conference on Learning Representations*.
- [130] Valerie Wirtshafer. 2025. *For AI to make government work better, reduce risk and increase transparency*. Retrieved Sept 2, 2025 from <https://www.brookings.edu/articles/for-ai-to-make-government-work-better-reduce-risk-and-increase-transparency/>
- [131] Katri Ylönen. 2023. The use of Electronic Information Systems in social work. A scoping review of the empirical articles published between 2000 and 2019. *European Journal of Social Work* 26, 3 (May 2023), 575–588. <https://doi.org/10.1080/13691457.2022.2064433>

A EXAMPLE TOPICS GROUPED FOR THEMATIC SIMILARITY

In Regular casenotes, we found that the topic model identified four topics related to legal matters. However, the topics differed in their mode of delivery, where one topic picked up on workers' descriptions of legal proceeding updates, while the other three topics picked up on legal updates written in the form of emails, texts, and phone calls. Similarly, in Service Plan Activities, we found four topics were related to substance abuse, wherein one focused on a person 'not' being under the influence, while other topics focused on a person receiving specific treatments and services for alcohol or substance abuse.

B EXEMPLAR SENTENCES FOR CASENOTE THEMES

This section provides exemplar sentences for the themes, shared and unique, across the different narrative types. All examples provided have been anonymized.

- **(Shared Theme) Family Relationship/ Visits & Parenting**

Objective: "Parent/caregiver has sufficient communication skills."

Activity: "[name] and [name] will be connected to family counselling/ parenting support program"

Regular Casenote: "I informed dad of placement. I advised [name] by text message that [name] was placed in foster care today and that I will be in touch with him about access visits next week. [name] responded saying thanks for the information and help."

- **(Shared Theme) Child Custody & Criminal/Legal**

Objective: "The child/youth does not engage in unlawful activity."

Activity: "[name] and [name] to attend mediation to work out a custody arrangement for [name]."

Regular Casenote: "I attended the home, [name] was home with her husband. [name] advised that [name] is sleeping, it seems she is tired, she was at the day care for the all day... I served [name] with court documents. I asked her if she would like me to review the papers with her but she said that she is fine. We talked about court, [name] is ready for mediation and full custody to continue to care for [name]."

- **(Shared Theme) Medical/Mental Health**

Objective: "Child/youth receives preventative medical, dental and/or vision care."

Activity: "[name] will take [name] for regular check up with the family doctor"

Regular Casenote: "Voicemail from walk in doctor - scratch may be old, it's not tender - no bruising and no bruising elsewhere - he is not in any distress - pupils are good - nothing in his ears - no marks of abuse - just wanted to follow up - looks comfortable, alert and bright - no worries at this time. I called Dr. [name] and he stated what he said in the voicemail. He said he had no worries or concerns and was not sure why he was seeing [name]."

- **(Shared Theme) Anger Management Conflict & Safety**

Objective: "Child/youth will not be exposed to physical conflict/violence in the home."

Activity: "[name] and [name] will not expose [name] to partner/adult conflicts and issues are resolved amicably"

Regular Casenote: "[name] said she has always paid attention to her daughters, never let anyone carry her daughters when they were young, never left them alone with anyone."

- **(Shared Theme) Administration related tasks, resources, and scheduling**

Objective: "To connect [name] to appropriate community supports and services"

Activity: "The worker will make referrals to youth services to address [name]'s physical and mental health conditions"

Regular Casenote: "Call to victim witness assistance to obtain info. Spoke to a rep who advised [name]'s assigned worker is [name]. I advised her [name] only speaks [language] so won't be able to speak without an interpreter. She said she will update the file, send [name] an email to contact [name] with an interpreter."

- **(Unique Theme) Attempts to Contact**

Regular Casenote: "Unsuccessful scheduled phone call with [name]: she did not call. [name] and I were scheduled to have a phone call to get an update about how things are going with her and her access, and to discuss her plan and how that is progressing."

- **(Unique Theme) Support Network**

Objective: "For the family to have a large support network surrounding them that does not include the society."

Activity: "For there to be a suitable network of supports to access as needed to help [name] and the family during this challenging time."

- **(Unique Theme) Child Development**

Objective: "Child/youth's physical and cognitive skills are age appropriate."

Objective: "Child/youth demonstrates adequate social skills."

C ANALYSIS OF SERVICE PLAN OBJECTIVES AND ACTIVITIES

When we manually examined our 38-topic model solution for Service Plan Objectives, we found over half of the Objectives in our dataset (74%, 1242/1677) were identically worded goals such as "Family resolves conflict regarding cultural differences," "Family engages with a strong support systems" or "Parent/caregiver physical or mental health does not affect parenting, family functioning and/or resources." The remaining Objectives included specific goals intended to be carried out by specific people or detailing specific actions, such as "[name] ensures [name]'s medical needs are met" or "To continue to try to locate biological mother," but even these objectives fell under consistent themes listed in Table 3. These findings suggest that high-level goals (i.e., Objectives) in child welfare at the agency are relatively standardized across the agency. On the other hand, manual inspections of our 82-topic model solution for Activities revealed how caseworkers provide customized supports to families based on individual circumstances so that a family may successfully meet their Service Plan's Objectives. There were more Activities in our dataset compared to Objectives because there are multiple paths for a family to achieve a Service Plan's Objectives. Moreover, Activities provided concrete, actionable steps to achieve a case's Objective. Each Activity included information on 1) the

person(s) the activity should be carried out by E.g., a child, parent, or caseworker. And 2) required action to complete the activity. This could be a single, multiple/ongoing behavioral actions or an absence of an action, such as a child regularly attending school or a father not drinking alcohol in the presence of children.

D AMBIGUITIES IN CASE PROGRESS

The following bullet points provide a few examples of cases in which case progress toward Service Plan Activity completion was ambiguous.

- When an activity is to, “[Name] to attend sobriety treatment programs”, it was unclear if a caseworker attempting to refer and schedule the client to sobriety treatment programs would be considered Activity-relevant.
- In one case, a child was not enrolled in school. An Activity for the family included, “Conduct academic assessments for school reading, writing literacy for [name].” The caseworker assisted the family in enrolling the child in the school so

that they could take the academic assessments. In this case, it was unclear if school registration facilitation would be considered Activity-relevant.

- A family’s Activity stated that, “[name] to demonstrate parenting interactions and activities that match with the developmental age of her children.” In this case, a bio-parent had access to visits with a child who was not living with the parent. Regular casenotes showed that the bio-parent would cancel scheduled access visits with the child, which occurs commonly across all child welfare cases when parents face unexpected scheduling conflicts. The first few cancellations appeared harmless, but it became increasingly clear that the bio-parent was cancelling many pre-scheduled access visits and not meeting the Service Plan’s Activity. In this case, our co-authors at the agency agreed that the cancellations can indicate Activity-relevant signals but one would only know this retrospectively after several cancellations had been made.