Child & Family Social Work

WILEY

CHILD & FAMILY SOCIAL WORK

ORIGINAL ARTICLE OPEN ACCESS

Making Accurate Judgements in Child Welfare: Comparing ChatGPT With Qualified Social Workers

David Wilkins | Verity Benett

CASCADE, Cardiff University, Cardiff, UK

Correspondence: David Wilkins (wilkinsd3@cardiff.ac.uk)

Received: 19 July 2024 | **Revised:** 22 January 2025 | **Accepted:** 11 April 2025

Keywords: artificial intelligence | bias | child protection | decision-making | judgement | machine learning | social work

ABSTRACT

This study compares the judgemental accuracy of child and family social workers ($n = 581$) with ChatGPT, a generative AI model. Using 12 anonymized referrals, participants were asked predictive questions to evaluate accuracy through Brier scores. ChatGPT outperformed the average social worker on 11 of the 12 referrals, though the difference was not statistically significant. These findings highlight the potential *and* the limitations for AI to support decision-making in social work while emphasising the need to address ethical concerns and AI's inadequacies for understanding complex human needs and social contexts. The study contributes to ongoing discussions on integrating AI into social work, advocating for a balanced approach that enhances effectiveness while preserving the profession's essential human elements.

1 | Introduction

Making judgements and decisions is a crucial aspect of the child and family social work role (Taylor and White 2001, 2006). Practitioners must balance ethical and legal considerations, weigh evidence and incorporate the views of children, parents and families (Forrester 2024). In child protection, they regularly assess and predict future parenting behaviour based on past and present risk factors (Juhasz 2020), often operating under uncertainty (Taylor and White 2006) and taking decisions 'laden with risk' (Budd 2005). As one social worker noted, 'you cannae see in the future ... that would make the job a lot easier, wouldn't it? (laughs)' (Bleasby 2023).

Social work decisions implicitly or explicitly claim truths about the world as it is and as it may become (Clardy 2022; Chen et al. 2014). For instance, closing a referral assumes no significant risk to the child, although a family group conference may promote reunification (Wood et al. 2024) and uphold the family's right to participation (Holland et al. 2005). These decisions link means to ends, even when the causal relationships are unclear (Conklin 2006) and information is incomplete (Flanagan 2020). Such judgements profoundly impact children and families, underscoring why decision-making in social work has been the focus of sustained debate and scrutiny.

In her seminal book *Social Diagnosis* (1917), Mary Richmond highlighted the potential for errors in practitioners' judgements due to limitations in assessing case-specific information, emphasising the role of skilled supervisors in identifying patterns and encouraging alternative perspectives. The inquiry into Maria Colwell's death in 1974 criticised her social worker's judgement, though systemic factors also played a role (Minty 1994; Munro 2019). Similarly, Lord Laming's report on Victoria Climbié's death noted that supervision often '[amounted] to no more than a rubber-stamping of ... decisions' (Laming 2003, 210). Munro (2011a) later observed that 'management and inspection processes [have hindered] professionals' ability to exercise their ... judgement' in child protection. More recently, the Independent Review of Children's Social Care (MacAlister 2022) criticised decision-making as 'too inconsistent' and 'underpinned by a lack of knowledge' (11, 37). Both reviews emphasised the ecological complexity of social work decision-making (Baumann et al. 2014).

BOX 1

A brief explanation of some key terms: AI, machine learning and large language models.

AI refers to any form of intelligence that is not animal or human. Machine learning is a specific kind of AI based on statistical algorithms capable of learning and performing various tasks without explicit instruction. This is achieved by generalising from existing data (El Naqa and Murphy 2015). A machine learning algorithm in the field of social work might be assigned the task of identifying children who are at risk of significant harm by examining local authority records. A large language model (LLM) is a form of machine learning, made using vast quantities of data such as the entire world wide web, to create a generalised ability to understand and generate language. The autocomplete function on most mobile phones is a relatively simple language model. OpenAI's ChatGPT is a form of generative AI using machine learning to identify word sequence patterns in its training data, enabling it to respond to natural language prompts from the user by predicting the next word, sentence or paragraph.

In response, various reforms have been proposed, including more detailed policies (Laming 2003), greater professional discretion (Munro 2011b) and specialist child protection teams (MacAlister 2022). Others advocate for AI and machine learning to augment professional judgements, prompting significant debate (Reed and Karpilow 2002; Vaithianathan et al. 2023; Hodgson et al. 2023).

When considering the use of AI and machine learning in social work, ethical implications must take precedence (Hall et al. 2024). Concerns include the depersonalization of services and the potential for algorithms to reinforce systemic bias and discrimination (Gillingham 2016; Glaberson 2019). For instance, a predictive algorithm tested in the New York criminal justice system discriminated against Black defendants, recommending lower pre-trial release rates (Arnold et al. 2021). However, human judgement is not immune to bias; for example, Black children are over-represented in care compared to White children in England (Edney et al. 2023). Some argue predictive models could help reduce such bias in child welfare decision-making (Vaithianathan et al. 2023). Additionally, preferences for algorithmic versus human judgements are not universal. Laypeople in one study preferred algorithmic decisions, especially when perceived as more accurate (Logg et al. 2019). To address these tensions, ethical frameworks for AI in social work must prioritise dignity, empowerment, social justice and equity (Leslie et al. 2021).

Empirical evidence further highlights AI's potential in improving judgemental accuracy. Machine learning algorithms have outperformed traditional risk assessments, as demonstrated by Pan et al. (2017), whose model improved the identification of high-risk pregnancies by 36%, enabling more targeted support. Victor et al. (2021) found similar success using AI to predict domestic abuse risks in child welfare investigations, though they cautioned against applying these models to individual cases. In English social work, Clayton et al. (2020) developed machine learning models with natural language processing (NLP) to predict outcomes such as re-referral, child protection plans and care placements. These models achieved an average area under the curve (AUC) of 0.75, indicating moderate accuracy. However, much of this success stemmed from avoiding false positives, a result easier to achieve with rare outcomes.

Although AUC thresholds offer benchmarks, they can be arbitrary, and we lack direct comparisons between algorithmic and social worker accuracy. In related research (authors' own), social workers achieved an AUC of 0.68 on similar tasks, slightly below Clayton et al.'s models. Given the importance of judgemental accuracy to decision-making quality (Hood et al. 2022), AI may offer some potential to support social work. This study explores this by directly comparing the accuracy of social worker judgements with those made by perhaps the current most popular and widely known example of such technology, namely, ChatGPT.

2 | Method

To compare the judgemental accuracy of social workers and AI, ChatGPT (v3.5) was asked to analyse a series of 12 referrals and answer questions about the likelihood of subsequent actions, events and outcomes (Figure 1). Its responses were compared with those previously given by 581 social workers in relation to the same referrals and questions (authors own). The research question for the study was:

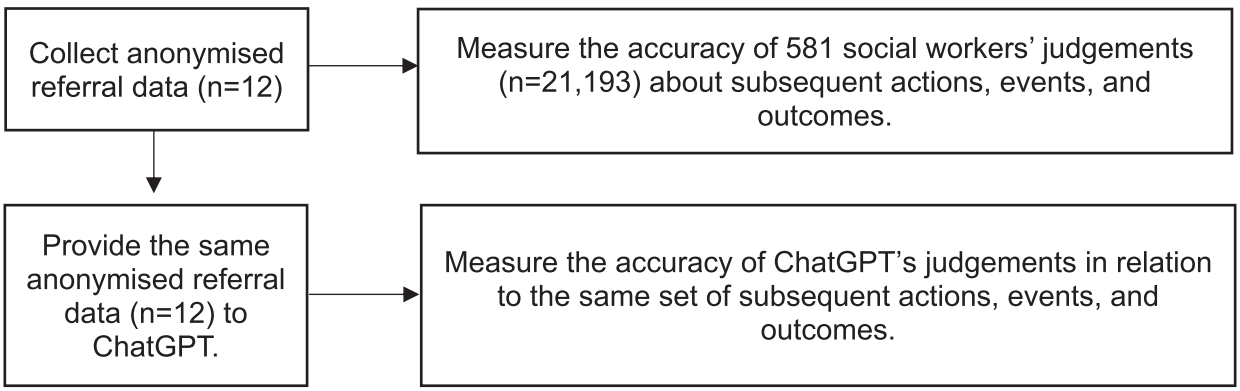

FIGURE 1 | An overview of the study design.

> Given the same information and in response to the same questions, can ChatGPT make more accurate forecasts than social workers?

ChatGPT is an NLP tool, underpinned by AI technology. From the user perspective, ChatGPT is available as a website (https://chat.openai.com) and can engage in text-based conversations (and more recently voice conversations) about a range of topics. It works by analysing user-provided prompts and responding with a series of words that it predicts will provide the best answer based on its training data (Briganti 2024). For this study, ChatGPT was given the following brief (Schoenegger et al. 2024):

> In this chat, you are a child protection social worker in England. As an experienced practitioner, you can evaluate different forms of evidence and make careful judgements about the nature of risk and harm to children and the likelihood of different future actions, events and outcomes. In a moment, I am going to ask you to read a series of anonymised referrals about children and answer a series of questions about what might happen next. For each question, you will provide me with a number between 0 and 100 that is your best prediction of the outcome.

No other instructions were provided for ChatGPT or for the social workers who took part in the earlier comparative surveys. For example, they were not advised that the probabilities for mutually exclusive outcomes should sum to 100%. The rationale for this is that such instructions could have impacted on the outcome, as part of the 'test' of making good forecasts is understanding when events may be independent, related and/or mutually exclusive.

ChatGPT was then provided with the anonymised referral information and a series of questions about the likelihood of subsequent actions, events and outcomes for each one (Tables 1 and 2). For every referral, the same two questions were asked each time:

TABLE 1 | An example of one of the referrals used in this study to prompt ChatGPT.

'I am concerned about Aadesh (aged 11) and his siblings due to domestic abuse at home. Aadesh's uncle is drug taking and selling within the family home, this is having a grave effect upon his mental health. There is domestic violence between mother and her brother-in-law. He is known to use cannabis, heroin, and crack-cocaine. He was previously admitted to a mental health unit under section. Diagnosed with schizoaffective disorder and a personality disorder. The children must be witnessing domestic violence, and I think they would be scared of their uncle. I've spoken to the mother, and she said she is scared of him, and they cannot ask him to leave the home in case he reacts aggressively. I think he may also be drug dealing within the home. He is currently under a community mental health treatment order'
(Referral received from a community mental health nurse)

1. In response to this referral, how likely are the following actions—no further action, a social work or other form of assessment; emergency removal into care; something else *and*
2. Within the next 12 weeks, how likely is it that the child will become the subject of—no plan; a child in need plan; a child protection plan; a looked after child plan.

In addition, several bespoke questions were also posed for each referral. The 12 anonymised referrals were collected from one local authority in England in 2020. These referrals were selected from a 6-month period within that year to ensure temporal consistency. The local authority was asked to generate an anonymised list of all referrals received during this timeframe, stratified by the outcome of the referral. This stratification included categories such as no further action, assessment, child in need plan, child protection plan and other outcomes. From these stratified sub-lists, referrals were selected to ensure a representative sample that captured the diversity of outcomes typically encountered in child welfare practice. Specifically, the final selection included (a) three referrals that resulted in no further action and (b) nine referrals that led to an assessment. Among the nine assessed cases, further stratification was applied to include (i) three that led to a child in need plan, (ii) three that resulted in a child protection plan and (iii) three with other outcomes. This process was designed to ensure that the sample adequately reflected the range of scenarios and outcomes that practitioners routinely face. After selection, the lead author reviewed the corresponding case files to generate a series of bespoke questions for each referral, aimed at evaluating judgemental accuracy in forecasting subsequent actions, events and outcomes.

In a separate study (authors' own), a sample of social workers ($n = 581$) read these anonymised referrals and collectively provided 21 193 judgements in relation to a series of questions, providing a response for each on a scale from 0 (*the specified action, event or outcome will definitely not happen*) to 100 (*the specified action, event or outcome will definitely happen*). These participants were recruited through social media announcements and email invitations distributed to a limited number of local authorities in England. All participants self-reported that they were qualified social workers actively working with children and families in England at the time of the study. The sample predominantly consisted of female respondents (85.2%), aligning with the gender distribution of the broader social work workforce. Age demographics revealed that around one-third of the participants were aged 25–34 (34.9%), with a similar proportion aged 35–44 (30.7%). Most respondents identified their ethnicity as White English, Welsh, Scottish, Northern Irish or British (86.1%). Regarding professional experience, nearly one-quarter had been qualified between 1 and 3 years (23.6%), whereas approximately one-third had over 10 years of practice experience (35.1%). In terms of practice focus, just over a quarter worked within child in need or child protection teams (26.6%), with smaller proportions in referral and assessment teams (11.4%) and Looked After Children services (12.7%). As the study employed a non-probability sampling method, participation was contingent upon individuals' availability and interest, meaning the sample is not representative of the entire social work population.

To measure the accuracy of responses to the questions, a series of multi-category Brier scores were calculated for each one and means per referral. Brier scores are a statistical measure used

TABLE 2 | An overview of the 12 referrals and associated questions used in this study to prompt ChatGPT.

Child's pseudonym	Summary of the referral	Associated questions
Salma	From the police, with concerns about possible child sexual exploitation	Within the next 12 weeks, will the authority convene a strategy meeting? Within the next 6 months, how likely is it that there will be a further referral about Salma from any source? Within the next 6 months, how likely is it that Salma will come into care?
Unborn Clarke	From a hospital, with concerns about a high-risk pregnancy and the mother's mental health	Within the next 12 weeks, will the mother consent to a social work assessment? Within the next 12 weeks, will the mother attend at least one antenatal appointment? Within the next 12 weeks, will the authority convene a strategy meeting?
Taryn	From the mother, with concerns about domestic abuse and her son's behaviour	Within the next 12 weeks, will the authority convene a strategy meeting? Within the next 8 weeks, will Taryn's school attendance improve? Within the next 12 weeks, will the social worker be able to visit the family at home?
Stephanie	From the Court Advisory Service, with concerns about physical abuse	Within the next 12 weeks, will the children's school report to the social worker any concerns about the children? Within the next 12 weeks, will the father consent to further checks being made with professionals about the children's welfare? Within the next 12 weeks, will the social worker be able to visit the father at home?
Aadesh	From a community mental health team, with concerns about domestic abuse, mental health problems and substance misuse	Within the next 12 weeks, will the local authority convene a strategy meeting? Within the next 6 months, will there be another referral about Aadesh? Within the next 6 months, will Aadesh come into care?
Emelia	From the Court Advisory Service, with concerns about domestic abuse and parenting capacity	Within the next 12 weeks, will the social worker be able to visit the father at home? Within the next 6 months, how likely is it that the child will come into care? Within the next 6 months, how likely is it that there will be a further referral about this child from any source?
Omar	From another local authority, with concerns about maternal learning disability and domestic abuse	Within the next 12 weeks, will the father be arrested? Within the next 12 weeks, will the mother and children return to live with the father? Within the next 12 weeks, will the social worker be able to contact the father and talk to him about the referral?
Malalai	From a hospital emergency department, with concerns about physical abuse, mental health problems and self-harming behaviours	Within the next 12 weeks, will the social worker be able to visit the mother at home? Within the next 6 months, how likely is it that there will be a further referral about this child from any source? Within the next 12 weeks, will Malalai talk to the social worker about her mental health problems?
Poppy	From the police, with concerns about domestic abuse and parental alcohol misuse	Within the next 12 weeks, will the father be arrested? Within the next 12 weeks, will the father attend an appointment at an alcohol support service? Within the next 12 weeks, will the mother attend an appointment at an alcohol support service?

(Continues)

TABLE 2 | (Continued)

Child's pseudonym	Summary of the referral	Associated questions
William	From the police, with concerns about domestic abuse	Within the next 12 weeks, will the social worker be able to contact mother to discuss the referral? Within the next 12 weeks, will the authority convene a strategy meeting? Within the next 12 weeks, will the social worker be able to meet the mother and stepfather at home
Unborn Wooten	From another local authority, with concerns about a young person in care being pregnant	Within the next 12 weeks, will the authority convene a strategy meeting? Within the next 12 weeks, will the mother agree to attend a parenting programme? Within the next 12 weeks, will the social worker be able to meet mother at home?
Ava	From school, with concerns about neglect	Within the next 6 months, how likely is it that there will be a further referral about this child from any source? Within the next 6 months, how likely is it that the child will come into care? Within the next 12 weeks, will the social worker be able to meet mother and father at home?

to assess the accuracy of probabilistic judgements or forecasts (Brier 1950). They are calculated as the mean squared difference between anticipated probabilities and actual outcomes for a set of events or cases. Mathematically, a Brier score is expressed as follows, where BS is the Brier score, N is the number of events considered, t indexes the events from 1 to n (first event, second event, etc.), R is the number of possible outcomes for each event, i indexes the possible outcomes for each event, f is the forecast probability and o is the outcome (0 or 1).

$$BS = \frac{1}{N}\sum_{t=1}^{N}\sum_{i=1}^{R}\left(f_{ti} - o_{ti}\right)^2$$

Using this formula, Brier scores range from 0 to 2, whereas 'a hedged fifty-fifty call, or random guessing in the aggregate, will produce a Brier score of 0.50' (Tetlock and Gardner 2016, 64). A worked example is shown in Table 3. Having previously calculated Brier scores to measure the accuracy of social workers' judgements in relation to the same 12 referrals, the aim of this study was to record ChatGPT's responses to the same questions and calculate equivalent Brier scores (using Excel for Mac, v16.86).

2.1 | Ethics

The study was given ethical approval by (The School of Social Sciences, Cardiff University) in May 2023. The main ethical consideration was to protect the anonymity of the referrals and ensure that no data were retained by ChatGPT. As outlined above, the referrals were anonymised at the point of collection. To prevent the data being retained by ChatGPT, it was necessary to set up a new account and opt out of having any data retained. Once the responses of ChatGPT had been recorded, this account was deleted. According to the relevant terms and conditions, this means that all the data were permanently deleted within 4 weeks. Thus, none of the anonymised referral information or any of the associated questions were stored, used or processed by ChatGPT for any reason other than as part of the study.

3 | Findings

For all 12 referrals combined, ChatGPT achieved a mean Brier score of 0.42, compared with 0.49 for the social work sample ($n = 581$). ChatGPT was more accurate than the mean of the social work sample in relation to 11 out of the 12 referrals (Table 4). The most accurate individual social worker in the sample achieved an overall Brier score of 0.22, whereas the least accurate achieved a Brier score of 1.55 (authors' own). Overall, ChatGPT was *more* accurate than the average social worker, *far more* accurate than the individual practitioner with the least accurate Brier score and *far less* accurate than the individual practitioner with the most accurate Brier score.

A two-sample t-test was performed to compare the overall mean Brier scores (Table 5). Although the Brier score achieved by ChatGPT was lower than the mean average for the social work sample, this difference was not significant (M = 0.4854, SD = 0.09355) and ChatGPT (M = 0.4174, SD = 0.11292); $t(\mathrm{df}) = 1.913$, $p = 0.065$.

4 | Discussion

The aim of this small-scale study was to compare the accuracy of judgements made by social workers and ChatGPT in relation to the same anonymised referrals and associated questions about the likelihood of subsequent actions, events and outcomes. The judgements made by ChatGPT were on average more accurate than those made by social workers, albeit the most accurate social worker in the sample was 53% *more* accurate than this

TABLE 3 | A worked example of how to calculate Brier scores.

In response to this referral, how likely is each of the following outcomes	
Options	**Respondent's forecasts**
No further action	12%
Social work or other form of assessment	60%
Emergency removal into care	2%
Something else	26%

Note: Assume the outcome that did happen was **social work or other form of assessment**: $BS = (0.6-1)^2 + (0.4-0)^2 = 0.16 + 0.16 = 0.32$.

(authors' own). In any case, the difference was non-significant ($p < 0.05$). Given the limitations and exploratory nature of the study, these findings should be treated with caution, albeit this is an area of rapid technological development and one that many believe will increasingly challenge current ways of working (Tambe and Rice 2018; Hodgson et al. 2022; Singer et al. 2023). Reflecting on these findings, there are three questions that come to mind: (1) How come ChatGPT outperformed the average judgemental accuracy of the social work sample and many of the individual practitioners? (2) To what extent does judgemental accuracy matter in child and family social work? And (3) is it ethical to think that ChatGPT (or other forms of AI) might be used to help support and improve the accuracy of social work judgements?

TABLE 4 | A comparison of Brier scores achieved by ChatGPT and a sample of social workers in relation to 12 anonymised referrals.

Referral	ChatGPT Brier score	Social work sample mean average Brier score	Lowest (most accurate) social work Brier score	Highest (least accurate) social work Brier score
Salma	0.66	0.48	0.01	2.00
Unborn Clarke	0.60	0.63	0.04	1.55
Taryn	0.35	0.46	0.16	0.91
Steph	0.54	0.60	0.17	1.02
Aadesh	0.31	0.38	0.05	0.85
Emelia	0.39	0.48	0.09	1.03
Omar	0.50	0.53	0.20	1.14
Malalai	0.23	0.48	0.20	0.89
Poppy	0.31	0.37	0.06	0.85
William	0.40	0.59	0.20	1.59
Unborn Wooten	0.34	0.34	0.40	0.98
Ava	0.45	0.52	0.32	0.83
Overall	0.42	0.49	0.22	1.55

TABLE 5 | The results of a two-sample *t*-test comparing the Brier scores achieved by social workers and ChatGPT.

	Levene's test for equality of variances		*t*-test for equality of means							
					Sig.				95% confidence interval of the difference	
	F	Sig.	t	df	One-sided *p*	Two-sided *p*	Mean diff.	Std. error diff.	Lower	Upper
Equal variances assumed	0.305	0.585	1.913	32	0.032	**0.065**	0.06805	0.03556	−0.00440	0.14049
Equal variances not assumed			1.913	30.929	0.032	**0.065**	0.06805	0.03556	−0.00449	0.14059

The first of these questions may be the easiest. After all, making accurate judgements is difficult, especially about the future (Doyle et al. 2012). People are prone to systematic errors in their thinking commonly known as biases (Kahneman and Tversky 1996) and non-systematic errors less commonly known as noise (Kahneman et al. 2021). ChatGPT may have an in-built advantage, insofar as while it may also be prone to systematic errors (based on the biased nature of its training data), it may be less prone to non-systematic errors (Maclure 2021). ChatGPT is also immune to many other challenges that people face, such as tiredness, boredom and a lack of motivation. One can imagine that asking a social worker to read and answer questions about 12 referrals in a row might engender any or all these things—yet not for ChatGPT, which appeared as keen to answer questions about the 12th referral as it was about the first. No doubt social workers in real practice are susceptible to a similar kind of decision-making fatigue (Pignatiello et al. 2020), as are study participants when completing surveys (Backor et al. 2007).

In addition, in a range of studies, Tetlock and Gardner (2016), Tetlock et al. (2017) and Friedman et al. (2018) have found that the judgements of non-specialists (such as ChatGPT in relation to social work) can be more accurate than those of specialists. How come? Because in some cases, there is such a thing as knowing (and caring) too much. Perhaps when reading these referrals, instead of answering questions about the likelihood of different actions, events and outcomes, many social workers responded to a subtly different question, thinking about what *they* would do if faced with the same referral. The theory of motivated reasoning suggests that depending on the situation and the context, people may be less concerned with judgemental accuracy than they are with providing the socially acceptable, expected or personally preferred outcome while still constructing seemingly reasonable justifications (Kunda 1990). In practice, this may manifest in phenomena such as 'optimism bias', wanting to think the best of people, especially parents in relation to their own children (Kettle and Jackson 2017) or perhaps more contemporaneously as 'pessimism bias', being encouraged to *think the unthinkable* about possible child maltreatment out of fear that the worst *could* happen (Burton and Revell 2018).

However, it is also important to note that ChatGPT did not outperform the average social worker by a huge amount, and such tools self-evidently do not provide 'easy' solutions to the more complex practice challenges of social work. ChatGPT generates responses by predicting the most likely sequence of words based on its training data, which consist of a vast amount of text sourced from the internet. Although the parameters provided in this study directed its responses toward numerical outputs, the underlying principle remains unchanged: ChatGPT does not engage in mathematical calculations (perhaps most social workers do not either) or deep understanding (which social workers can and do). Rather, it creates text responses by identifying patterns in the data it has been exposed to, much like a parrot mimics phrases without comprehending their meaning.

Importantly, the training data for ChatGPT likely do not include formal social care records, as these are highly sensitive and protected. This raises questions about its relevance to social work, as it lacks direct exposure to the context and nuances inherent in social care scenarios. Intuitively, models trained on domain-specific data, such as formal social care records, might yield more accurate and contextually relevant predictions. For example, using NLP techniques, a supervised machine learning algorithm could be trained on a curated dataset of real-life scenarios labelled with outcomes. Such an approach might provide more precise predictions, including Brier scores that reflect performance metrics tailored to social work contexts. However, this would require access to high-quality, representative datasets and raises ethical considerations around data privacy, transparency and the appropriate use of sensitive information.

The second question to consider is whether and to what extent judgemental accuracy matters in social work. According to a recent literature review by Hood et al. (2022), there are five standards commonly applied when studying social work judgement and decision-making: (i) consistency, the extent to which different professionals make similar judgements and decisions about the same child or family, *or* the extent to which the same professional makes similar judgements and decisions about similar children and families at different times; (ii) outcomes, the extent to which judgements and decisions are associated with positive benefits; (iii) practice standards, the extent to which judgements and decisions are aligned with the values and ethics of social work; (iv) equity, the extent to which people from different socio-economic and demographic groups are treated fairly; and (v) accuracy, the extent to which judgements correspond with other sources of evidence, including subsequent events. In their review, Hood et al. identified several examples of how judgemental accuracy has been measured in social work. For example, Forrester (2007) analysed re-referral rates in three local authorities, and because they were quite low concluded that 'the identification of risk of serious abuse demonstrates a relatively high level of accuracy' (296). This may be so, although it may also reflect the relative rarity of serious abuse. By contrast, Farmer (2014) looked at what happened to children returning home after a period in care. Finding that 59% experienced subsequent maltreatment and 65% came back into care within 5 years, Farmer suggested that the judgements made by social workers before the children went home about the risk of future harm were less accurate than they might have been.

More generally, the notion of judgemental accuracy has been applied in child protection by Munro (1999) via the concept of error types. From this perspective, a *true* (or accurate) judgement is one in which we say 'X' and 'X' is true or 'not-X' and 'not-X' is true. A *false* (or inaccurate) judgement is one in which we say 'not-X' and 'X' is true, or 'X' and 'not-X' is true. In which cases, we have made an error—either false positive or false negative (Table 6).

Although this may seem relatively straightforward, the practice of making social work judgements is much more complicated than this may imply. It is important not to make the mistake of naïve empiricism—believing that we can make judgements based only on observable data, without the need for theory, context or moral reasoning (Bealer and Strawson 1992; Taylor and White 2006). When a social worker says 'this child is at risk of domestic abuse', this may represent their belief that there is a relatively high chance in the near future of the child's father assaulting the child's mother. Asked to justify this belief, the social worker may point to evidence from the recent past, including

TABLE 6 | Error types when judging whether a child is 'safe' or 'not safe'.

	Actual—child is safe	Actual—child is at risk of significant harm
Judgement—child is safe	*True negative*	*False negative*
Judgement—child is at risk of significant harm	*False positive*	*True positive*

reports of domestic abuse made by the mother or child, as well as information from other sources such as the police and hospital emergency departments. However, social work judgements are not mere descriptions of the world. They are also a method for constructing social reality (Taylor and White 2006) and of 'taking action' (Austin 1975; Searle 1983). When a social worker says, 'this child is at risk of domestic abuse', to a greater or lesser extent they create and change the nature of the thing described, influencing the beliefs and behaviour of others, including family members and professionals. This is very different from making mundane forecasts about the weather, in which your belief that it is going to be sunny has no bearing on whether it rains. Adding further to this complexity, when a social worker says, 'this child is at risk of domestic abuse', they are also making a moral judgement (Taylor and White 2001)—that *this kind* of behaviour is *wrong*, beyond the limits of social and moral acceptability, while also implying that something needs to be done about it. Although it is true that 'good' social work judgements must accurately represent the world, as it is now and as it may be in future, it is imperative to recognise that 'facts themselves are not simple and stable but can be used to produce accounts which are rhetorically potent' (Taylor and White 2001, 45). How much more potent such judgements might be if they are made based on an AI-informed analysis, rather than made by 'mere' human beings.

Finally, is it ethical to think that ChatGPT (or other forms of AI) might be used in practice to help support and improve the accuracy of social work judgements? For some, this may seem like putting the moral cart before the empirical horse. After all, no single study—especially one as small scale as this—can 'prove' that AI judgements are more accurate than those made by social workers, not least because the referral information was limited in scope and because the questions are not all directly comparable to those social workers are concerned about in their day-to-day practice. That said, it would be a mistake to think we need only address ethical questions once the data are in. Even if these results were replicated in a larger, representative and more rigorously designed study, could or should ChatGPT (or something like it) ever be used to augment the judgements made by social workers? As noted already and especially in child protection, judgement and decision-making is fraught with complexity. Even if ChatGPT *were* more accurate than social workers on average and in relation to certain questions, we will always need wise and humane social workers to think through the meaning and possible implications of such judgements (Forrester 2024).

Currently, ChatGPT and similar large language models (LLMs) do not adequately address these ethical concerns. Although ChatGPT can generate coherent text, it lacks transparency in its decision-making processes and does not allow for human contestability of its outputs. However, this does not mean that LLMs have no utility in social care. Bespoke LLMs, tailored to the specific needs of social work, could provide significant value in areas such as summarising and simplifying large volumes of text or translating documents into different language styles to suit various audiences.

Ultimately, good social work is about the fair and reasonable balancing of competing rights and responsibilities, and the resolution of ethical dilemmas that resist easy resolution. Social work will never be an exact science—not least because, what is good for one person may not be good for someone else, and because all of us have the right, within the limitations of a classically liberal society, to define for ourselves how we want to live and if we are parents how we want to raise our children (Forrester 2024).

5 | Limitations

This study, while providing some insights into the comparative judgemental accuracy of ChatGPT and child and family social workers, has several limitations that must be acknowledged. First, the comparative sample of social workers may not be representative of the wider population. Participants were recruited based on their motivation to take part, and this approach probably fails to represent the diversity of practice settings and experiences found across the profession. This limits the generalisability of the findings. Second, the scope of the study is inherently limited due to the small sample size of 12 referrals, all of which were drawn from a single local authority in England. Although the selection process ensured some element of representation across referral outcomes, the findings may not fully capture the variability and complexity of cases encountered across different local authorities, regions or broader national contexts. Additionally, the use of referrals from a single authority may introduce biases related to local practice patterns or demographic factors. These limitations restrict the generalisability of the results and highlight the need for caution when interpreting the findings.

Moreover, the forecasting questions were generated by the lead author based on personal experience as a social worker and may not constitute an objectively 'good test' of social workers' abilities. Indeed, asking the right questions is probably the most important aspect of forecasting research and is certainly one of the most important aspects of good practice. Third, ChatGPT was only asked to make a series of one-off predictions in relation to each referral. A better test of its capabilities would have been to ask it to make multiple predictions on the same questions, generating a mean average for each one as in a recent study by Schoenegger et al. (2024). Fourth, although ChatGPT did achieve a more accurate Brier score than the average social worker, this difference was not statistically significant. Finally,

the experimental design does not account for the interactive and iterative nature of social work decision-making in the real world, in which judgements are often revised because of new information and via consultation with colleagues and other professionals.

In acknowledging these limitations, this study highlights the need for further research that employs a more diverse and representative sample of social workers and utilises a broader array of case referrals and other information. Additionally, future studies should aim to develop forecasting questions that more accurately mirror the real-world challenges experienced by social workers, thereby enhancing the ecological validity of the findings.

6 | Conclusion

These findings and subsequent discussion underscore the need for a balanced and critical approach to the use of AI in social work. Although AI tools like ChatGPT may show some promise in assisting with 'mundane' social work tasks, they can provide no 'magic' or 'off-the-shelf' solution for the most complex things that social workers do, such as making judgements and decisions. There is a pressing need for further research into the responsible development and implementation of AI in social care. This includes both assistive technologies and predictive analytics. Specifically in relation to this study, future research should include a greater number and range of referrals, drawn from a diverse range of geographical areas. This would allow for greater confidence in the findings and enable exploration of any potential differences between different 'types' of referral.

More generally, evaluations of existing proprietary software and tools are needed to assess their impact on social work process efficiency and efficacy, their technological robustness and their ability to meet ethical requirements. Research should also focus on identifying priorities for AI tool development in social care, alongside the careful co-creation of bespoke tools tailored for social workers and people who use services. These tools must be rigorously evaluated in practice to ensure they deliver value while upholding ethical and practical standards.

Furthermore, transparency and explainability remain critical challenges. Future research should explore how algorithms can be made more understandable and accessible to key stakeholders, enabling informed decision-making and maintaining accountability. AI could also play a pivotal role in identifying service needs and risk factors within specific populations, providing valuable insights for targeted interventions.

Regulatory frameworks will be essential to govern the ethical deployment of AI in social care. These should align with international standards while addressing the specific use cases for AI in social work and covering key considerations for different types of AI. Establishing quality standards for AI tools, along with an independent evaluation committee, could help ensure the reliability and ethical compliance of these technologies. Practical ethical guidelines must also be developed, focusing on real-world applications.

Finally, training programmes will play a vital role in preparing social workers to use AI responsibly. These should include both general training to raise awareness of AI and its implications, as well as tool-specific training that equips practitioners to use these technologies effectively and ethically. Regular updates to these programmes will be critical to keeping pace with ongoing advancements in AI.

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

Arnold, D., W. Dobbie, and P. Hull. 2021. "Measuring Racial Discrimination in Algorithms." *AEA Papers and Proceedings* 111: 49–54.

Austin, J. L. 1975. *How to Do Things With Words*. Harvard University Press.

Backor, K., S. Golde, and N. Nie. 2007. "Estimating Survey Fatigue in Time Use Study." International Association for Time Use Research Conference, Washington, DC; Citeseer.

Baumann, D. J., J. D. Fluke, L. Dalgleish, and H. Kern. 2014. "The Decision Making Ecology." In *From Evidence to Outcomes in Child Welfare: An International Reader*, 24–40. Oxford Academic.

Bealer, G., and P. F. Strawson. 1992. "The Incoherence of Empiricism." *Proceedings of the Aristotelian Society, Supplementary Volumes* 66: 99–143.

Bleasby, C. 2023. "'Nobody Wants to Remove a Baby … That's the Crux of It': Social Workers' Experiences of Undertaking Pre-Birth Assessments." University of Dundee.

Brier, G. W. 1950. "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review* 78, no. 1: 1–3. https://doi.org/10.1175/1520-0493(1950)078%3C0001:VOFEIT%3E2.0.CO;2.

Briganti, G. 2024. "How ChatGPT Works: A Mini Review." *European Archives of Oto-Rhino-Laryngology* 281: 1565–1569.

Budd, K. S. 2005. "Assessing Parenting Capacity in a Child Welfare Context." *Children and Youth Services Review* 27: 429–444.

Burton, V., and L. Revell. 2018. "Professional Curiosity in Child Protection: Thinking the Unthinkable in a Neo-Liberal World." *British Journal of Social Work* 48: 1508–1523.

Chen, Y., I. A. Kash, M. Ruberry, and V. Shnayder. 2014. "Eliciting Predictions and Recommendations for Decision Making." *ACM Transactions on Economics and Computation (TEAC)* 2: 1–27.

Clardy, A. 2022. "What Can We Know About the Future? Epistemology and the Credibility of Claims About the World Ahead." *Foresight* 24: 1–18.

Clayton, V., M. Sanders, E. Schoenwald, L. Surkis, and D. Gibbons 2020. "Machine Learning in Children's Services." What Work's for Children's Social Care, London.

Conklin, J. 2006. "Wicked Problems & Social Complexity." CogNexus Institute San Francisco, CA.

Doyle, C. C., W. Mieder, and F. R. Shapiro. 2012. *The Dictionary of Modern Proverbs*. Yale University Press.

Edney, C., B. Alrouh, and M. Abouelenin. 2023. "Ethnicity of Children in Care and Supervision Proceedings in England." Briefing paper. Nuffield Justice Observatory, London.

El Naqa, I., and M. J. Murphy. 2015. "What Is Machine Learning?" In *Machine Learning in Radiation Oncology*, edited by I. El Naqa, R. Li, and M. Murphy, Springer, Cham. https://doi.org/10.1007/978-3-319-18305-3_1.

Farmer, E. 2014. "Improving Reunification Practice: Pathways Home, Progress and Outcomes for Children Returning From Care to Their Parents." *British Journal of Social Work* 44: 348–366.

Flanagan, N. 2020. "The Information Behaviour of Social Workers: Needs, Seeking, Acquiring and Using Information in Practice." *British Journal of Social Work* 50: 1588–1610.

Forrester, D. 2007. "Patterns of Re-Referral to Social Services: A Study of 400 Closed Cases." *Child & Family Social Work* 12: 11–21.

Forrester, D. 2024. *The Enlightened Social Worker: An Introduction to Rights-Focused Practice*. Policy Press.

Friedman, J., J. Baker, B. Mellers, P. E. Tetlock, and R. Zeckhauser. 2018. "The Value of Precision in Probability Assessment: Evidence From a Large-Scale Geopolitical Forecasting Tournament." *International Studies Quarterly* 62, no. 2: 410–422.

Gillingham, P. 2016. "Predictive Risk Modelling to Prevent Child Maltreatment and Other Adverse Outcomes for Service Users: Inside the 'Black Box' of Machine Learning." *British Journal of Social Work* 46: 1044–1058.

Glaberson, S. K. 2019. "Coding Over the Cracks: Predictive Analytics and Child Protection." *Fordham Urban Law Journal* 46: 307.

Hall, S. F., M. Sage, C. F. Scott, and K. Joseph. 2024. "A Systematic Review of Sophisticated Predictive and Prescriptive Analytics in Child Welfare: Accuracy, Equity, and Bias." *Child and Adolescent Social Work Journal* 41: 831–847. https://doi.org/10.1007/s10560-023-00931-2.

Hodgson, D., S. Goldingay, J. Boddy, S. Nipperess, and L. Watts. 2022. "Problematising Artificial Intelligence in Social Work Education: Challenges, Issues and Possibilities." *British Journal of Social Work* 52: 1878–1895.

Hodgson, D., L. Watts, and S. Gair. 2023. *Artificial Intelligence and Implications for the Australian Social Work Journal*. Taylor & Francis.

Holland, S., J. Scourfield, S. O'neill, and A. Pithouse. 2005. "Democratising the Family and the State? The Case of Family Group Conferences in Child Welfare." *Journal of Social Policy* 34: 59–77.

Hood, R., S. Abbott, B. Coughlan, et al. 2022. "Improving the Quality of Decision Making and Risk Assessment in children's Social Care: A Rapid Evidence Review."

Juhasz, I. B. 2020. "Child Welfare and Future Assessments–An Analysis of Discretionary Decision-Making in Newborn Removals in Norway." *Children and Youth Services Review* 116: 105137.

Kahneman, D., O. Sibony, and C. R. Sunstein. 2021. *Noise: A Flaw in Human Judgment*. Hachette UK.

Kahneman, D., and A. Tversky. 1996. "On the Reality of Cognitive Illusions."

Kettle, M., and S. Jackson. 2017. "Revisiting the Rule of Optimism." *British Journal of Social Work* 47: 1624–1640.

Kunda, Z. 1990. "The Case for Motivated Reasoning." *Psychological Bulletin* 108: 480–498.

Laming, H. 2003. "The Victoria Climbie Inquiry."

Leslie, D., C. Burr, M. Aitken, J. Cowls, M. Katell, and M. Briggs. 2021. "Artificial Intelligence, Human Rights, Democracy, and the Rule of Law: A Primer." https://doi.org/10.2139/ssrn.3817999.

Logg, J. M., J. A. Minson, and D. A. Moore. 2019. "Algorithm Appreciation: People Prefer Algorithmic to Human Judgment." *Organizational Behavior and Human Decision Processes* 151: 90–103.

MacAlister, J. 2022. "The Independent Review of Children's Social Care (in England): Final Report Executive Summary."

Maclure, J. 2021. "AI, Explainability and Public Reason: The Argument From the Limitations of the Human Mind." *Minds and Machines* 31: 421–438.

Minty, B. 1994. *Maria Colwell: The Legacy*. Taylor & Francis.

Munro, E. 1999. "Common Errors of Reasoning in Child Protection Work." *Child Abuse & Neglect* 23: 745–758.

Munro, E. 2011a. *The Munro Review of Child Protection*. Stationery Office.

Munro, E. 2011b. *The Munro Review of Child Protection: Final Report, a Child-Centred System*. Stationery Office.

Munro, E. 2019. "Decision-Making Under Uncertainty in Child Protection: Creating a Just and Learning Culture." *Child & Family Social Work* 24: 123–130.

Pan, I., L. B. Nolan, R. R. Brown, et al. 2017. "Machine Learning for Social Services: A Study of Prenatal Case Management in Illinois." *American Journal of Public Health* 107: 938–944.

Pignatiello, G. A., R. J. Martin, and R. L. Hickman Jr. 2020. "Decision Fatigue: A Conceptual Analysis." *Journal of Health Psychology* 25: 123–135.

Reed, D. F., and K. Karpilow. 2002. "Understanding the Child Welfare System in California: A Primer for Service Providers and Policymakers." California Center for Research on Women & Families.

Schoenegger, P., I. Tuminauskaite, P. S. Park, and P. E. Tetlock. 2024. "Wisdom of the Silicon Crowd: LLM Ensemble Prediction Capabilities Match Human Crowd Accuracy". arXiv preprint, arXiv:2402.19379.

Searle, J. R. 1983. *Intentionality: An Essay in the Philosophy of Mind*. Cambridge University Press.

Singer, J. B., J. C. Báez, and J. A. Rios. 2023. "AI Creates the Message: Integrating AI Language Learning Models Into Social Work Education and Practice." *Journal of Social Work Education* 59: 294–302.

Tambe, M., and E. Rice. 2018. *Artificial Intelligence and Social Work*. Cambridge University Press.

Taylor, C., and S. White. 2001. "Knowledge, Truth and Reflexivity: The Problem of Judgement in Social Work." *Journal of Social Work* 1: 37–59.

Taylor, C., and S. White. 2006. "Knowledge and Reasoning in Social Work: Educating for Humane Judgement." *British Journal of Social Work* 36: 937–954.

Tetlock, P. E., and D. Gardner. 2016. *Superforecasting: The Art and Science of Prediction*. Random House.

Tetlock, P. E., B. Mellers, and J. Scoblic. 2017. "Bringing Probability Judgements Into Polic Debates via Forecasting Tournaments." *Science* 355: 481–483.

Vaithianathan, R., S. Cuccaro-Alamin, and E. Putnam-Hornstein. 2023. "Improving Child Welfare Practice Through Predictive Risk Modeling: Lessons From the Field." In *Strengthening Child Safety and Well-Being Through Integrated Data Solutions*. Springer.

Victor, B. G., B. E. Perron, R. L. Sokol, L. Fedina, and J. P. Ryan. 2021. "Automated Identification of Domestic Violence in Written Child Welfare Records: Leveraging Text Mining and Machine Learning to Enhance Social Work Research and Evaluation." *Journal of the Society for Social Work and Research* 12: 631–655.

Wood, S., J. Scourfield, M. Meindl, et al. 2024. "Family Group Conference Provision in UK Local Authorities and Associations With Children Looked After Rates." *British Journal of Social Work* 54, no. 5: 2045–2066. https://doi.org/10.1093/bjsw/bcae019.