Street-Level AI: Are Large Language Models Ready for Real-World Judgments?

Gaurab Pokharel[1], **Shafkat Farabi**[1], **Patrick J. Fowler**[2] **Sanmay Das**[1]

[1]Virginia Tech
[2]Washington University in St. Louis
gaurab@vt.edu, mfarabi@vt.edu, pjfowler@wustl.edu, sanmay@vt.edu

Abstract

A surge of recent work explores the ethical and societal implications of large-scale AI models that make "moral" judgments. Much of this literature focuses either on alignment with human judgments through various thought experiments or on the group fairness implications of AI judgments. However, the most immediate and likely use of AI is to help or fully replace the so-called street-level bureaucrats, the individuals deciding to allocate scarce social resources or approve benefits. There is a rich history underlying how principles of local justice determine how society decides on prioritization mechanisms in such domains. In this paper, we examine how well LLM judgments align with human judgments, as well as with socially and politically determined vulnerability scoring systems currently used in the domain of homelessness resource allocation. Crucially, we use real data on those needing services (maintaining strict confidentiality by only using local large models) to perform our analyses. We find that LLM prioritizations are extremely inconsistent in several ways: internally on different runs, between different LLMs, and between LLMs and the vulnerability scoring systems. At the same time, LLMs demonstrate qualitative consistency with lay human judgments in pairwise testing. Findings call into question the readiness of current generation AI systems for naive integration in high-stakes societal decision-making.

1 Introduction

Large language models (LLMs) have captured public attention and have been broadly touted for their ability to reduce and streamline human work. A question of particular interest of late has been how the moral judgments of LLMs compare with those of humans. This has been engaged in the context of philosophical moral dilemmas (Kim et al. 2025; Jha et al. 2024; Rathje 2024; Dillion et al. 2025), organ transplantation (Murray et al. 2025; Hasjim et al. 2024), and fair division (Hosseini and Khanna 2025), among other domains. With the exception of organ transplantation, most research attends to general conceptions of ethical judgment or definitions of fairness from the fair division community. The central questions of these lines of research have been in trying to understand which general theories of ethics or definitions of fairness LLMs appear to satisfy or whether they replicate certain aspects of human behavior.

However, LLMs are likely to see rapid uptake in domains that involve the allocation of scarce societal resources. The stakes will be more immediate and higher than in toy examples or constructed moral dilemmas, given the scale enabled by AI and the potential number of people affected by such decision-making in high-stakes domains like homelessness services, post-disaster medical triage, and organ transplantation. By analogy to "vibe coding," those working in these domains may feel pressure to use AI to make "vibe prioritization" decisions.

Such social and health service settings share key features uniquely relevant for adapting AI applications. First, current delivery reflects *local justice* principles and prioritization practices that emerged to fit the specific contexts over time (Elster 1992). Scarcity requires the adoption of agreed upon rules of who should get what, when, and how. For example, homelessness services – the focus of the present study – currently use a "vulnerability first" prioritization, where households deemed to be at greatest risk receive highest priority for resource allocation (Kube, Das, and Fowler 2023). Medical triage, by contrast, often prioritizes based on maximum expected improvement from receiving the resource. Prioritization frameworks are not inherent and rather emerge from complex social and political processes at play in many countries' social safety net contexts. For homelessness, *implementation* involves the use of point systems based on questionnaire responses administered to at-risk households upon contact with the system (often in so-called "coordinated entry" settings). Point systems are what we refer to as *bureaucratic* scoring systems frequently observed in other domains, like public housing and prioritization for liver transplant (Johnson and Zhang 2022; Been et al. 2018; Cholongitas, Germani, and Burroughs 2010).

Second, actual allocation decisions occur in the province of *street-level bureaucrats* – civil servants and caseworkers allowed to use their experience to exercise discretion in decision making (Lipsky 1980). Bureaucratic scoring systems often inform, but do not determine, decisions. Bureaucrats themselves bring considerable domain knowledge and expertise from interacting directly with those who ultimately face the outcomes of decisions. Given the embedding within local contexts, street-level judgments likely differ from those of lay humans. Importantly, it is these frontline workers who are likely to be replaced or supplemented by LLMs or "vibe

prioritization."

It is crucial to understand how LLMs would make decisions in such domains and to be as close to reality as possible. Prior work on organ transplantation appears to come closest to doing so. Murray et al. (2025) examine the abilities of LLMs to assess medical compatibility in kidney transplantation, as well as the group fairness implications of LLM prioritization. Likewise, Hasjim et al. (2024) investigate the abilities of LLMs to predict medical benefits and risks of liver transplantation. Yet, neither paper compares LLM prioritization with an established human system of prioritization.

This is the gap that we seek to fill. By systematically querying LLMs, we show that they behave in ways that are qualitatively similar to *lay* humans, as evidenced by a different set of pairwise comparison experiments. Moreover, using a novel dataset of households experiencing homelessness in a major metro area, we compare LLM prioritization with two different standard bureaucratic prioritization methods and assess how well each method predicts the actual allocation decisions of homeless service providers. In contrast with their similarities to lay humans in the simulated task, we find that LLMs are (1) internally inconsistent in the rankings they generate across different runs; (2) inconsistent with the rankings generated by bureaucratic ranking systems; and (3) worse at predicting the allocation decisions of caseworkers (although the scoring systems themselves are not great at this). Taken together, the results seriously question the replacement of street-level bureaucrats with AI in high-stakes societal allocation domains and emphasize the importance of understanding the behavior of front-line workers interacting with the uncertainties inherent in homeless service provision.

2 Background

In just the past few years, public interest in LLMs has surged remarkably, and researchers have explored their potential to make moral judgments and guide fair resource allocation. In particular, abstract allocation tasks – where agents distribute indivisible goods and, optionally, monetary compensation among individuals with heterogeneous valuations – have assessed LLM behavior against foundational fairness axioms such as equitability and envy-freeness. Hosseini and Khanna (2025) show that although LLMs can sometimes satisfy single fairness criteria, their choices often diverge from human distributional preferences and are sensitive to prompt phrasing and persona assignments. Complementing these findings, Kim et al. (2025) document substantial variance in moral reasoning when models adopt different sociodemographic personas, with politically charged framings exacerbating bias and polarization. At the same time, LLMs hold the promise of articulating trade-offs and justifying their decisions in rich natural language rationales – fueling an emerging concern: do these "toy" successes hold up under the pressures of real-world stakes?

Such stylized experiments, whether dividing cookies, dollars, or chores, inevitably abstract away the contextual nuances and ethical complexities of high-stakes domains. Yet, the public increasingly perceives LLM outputs as possessing moral expertise comparable to or exceeding that of professional ethicists (Dillion et al. 2025). Misplaced confidence underscores a key limitation: high-stakes domains demand more than metaphorical toy examples and call for careful evaluation in concrete, socially consequential contexts.

To move beyond the question of "which pie slice feels fairest" and toward "who receives life-altering resources first," we must ground our inquiry in a domain defined by: first, the existence of standardized triage tools; second, access to rich real-world data; and third, clear human benchmarks against which to judge model performance. Homelessness resource allocation is one such domain. Communities across North America establish protocols – commonly referred to as coordinated entry systems – to assess household vulnerability and to prioritize scarce homelessness services for the neediest.

Operationalizing this vulnerability-forward framework, the Vulnerability Index-Service Prioritization Decision Assistance Tool (VI-SPDAT) suite — the core VI-SPDAT (OrgCode Consulting Inc. and Community Solutions 2015b) and its derivatives, the VI-F-SPDAT for families (OrgCode Consulting Inc. and Community Solutions 2015a) and the TAY-VI-SPDAT for transition-aged youth (OrgCode Consulting Inc. et al. 2015) - consists of brief, self-reported intake questionnaires covering domains, such as housing history, health status, and social support. Responses are aggregated through a predefined rubric into a composite acuity score, allowing minimally trained assessors to triage clients rapidly. To date, these tools have been deployed in over one thousand communities in North America (OrgCode Consulting Inc. and Community Solutions 2015b). Despite widespread use, the VI-SPDAT faces criticism for introducing biases against certain subpopulations and for oversimplifying complex needs (Brown et al. 2018; Cronley 2022; Shinn and Richard 2022).

Bureaucrats in Social Services: The responsibility for helping navigate this intricate triage process is typically entrusted to *street-level bureaucrats*, who wield significant discretion in policy interpretation and implementation (Lipsky 1980). Bureaucrats decipher policies and allocate homeless resources under tight budgets, standardization pressures, and their own professional judgment. Pokharel, Das, and Fowler (2024) show that, at a time when intervention assignments were not formulaic, beyond rule-based assignments, caseworkers exercise discretion that sometimes favors less vulnerable households but delivers larger marginal gains when more intensive services are provided, highlighting how experiential knowledge complements standardized scores.

Efforts to introduce caseworkers to algorithmic decision support tools are not new and have met resistance in the past. In child welfare, bureaucrats reject recommendations deemed contextually inappropriate despite institutional conformity pressures (Kawakami et al. 2022). Furthermore, front-line workers and unhoused individuals alike voice concerns about biases in AI-driven homelessness services, underscoring tensions between algorithmic objectives and individual needs (Kuo et al. 2023).

Illustrating these dynamics in homelessness services,

Kube et al. (2022) enlisted 458 Mechanical Turk participants to simulate street-level allocation decisions by completing ten pairwise comparisons of homeless household profiles drawn from St. Louis's Homeless Management Information System (HMIS) – randomizing subjects to view either only demographic and service-request data or that same data augmented with Bayesian Additive Regression Trees derived low/medium/high risk scores of reentering homelessness within two years. Choices clustered into two distinct prioritization styles – "vulnerability-oriented" (favoring households with higher reentry risk) and "outcome-oriented" (favoring those with lower risk) – and prior exposure to algorithmic predictions both nudged undecided participants toward outcome-oriented allocations and amplified each individual's intrinsic decision-making style. These results highlight how even relatively simple risk scores can subtly but systematically reshape frontline judgments, underscoring the complexity of introducing algorithmic support into real-world resource-allocation contexts.

At the same time, LLMs have been enthusiastically deployed in other social service domains, often without much scrutiny (Maitra et al. 2025; Taylor 2024; Bender and Hanna 2025). Empirical studies expose their brittleness and fairness vulnerabilities: small input tweaks can drastically alter clinical risk recommendations (Acharya et al. 2024), shift demographic outcomes under Borda-based rankings (Murray et al. 2025), or overlook subtle but critical factors in AI-driven committee decisions (Hasjim et al. 2024). Theoretical work also warns about interpretability challenges and the potential for rapid, catastrophic errors in the absence of human oversight (Conitzer 2024).

The documented vulnerabilities, alongside the critical and nuanced nature of the problem, beg the question: Should LLMs be considered in the context of homelessness resource allocation at all? Our analysis specifically examines the alignment of LLM judgments with both human caseworker assessments and established vulnerability scoring systems in homelessness resource allocation. Leveraging real-world intake data, we investigate LLM internal consistency, agreement across model variants, and compatibility with existing social and political prioritization mechanisms, insights that are crucial for assessing the practical viability of incorporating LLMs into high-stakes societal decision-making contexts.

3 Methods

To assess how well LLMs align with both human judgments and established vulnerability scoring systems in homelessness resource allocation, we design two complementary experimental tasks:

1 *Pairwise Comparisons*: We replicate an experiment by Kube et al. (2022) who used 10 pairs of household data drawn from St. Louis homelessness services records (including features on demographics, income, disability, service requests, and reentry to homelessness risk scores). The main purpose is to investigate whether the LLM being tested is more "outcome-oriented" or "vulnerability-oriented" in deciding which household to allocate a more intensive intervention and how the orientations align with lay humans. We limit our experiments to the same 10 pairs of household data Kube et al. (2022) used in order to accurately compare the LLM allignments with humans.

2 *Ranking Task*: In this experiment, we use data on homeless service vulnerability assessments collected from 2021 through 2024 in St. Louis and elicit pairwise prioritization comparisons from several LLMs to create a complete ranked list of households for prioritization. We then compare the LLM rankings with those obtained from standard rankings actually used in the real-world prioritization scheme.

In both tasks, we measure internal consistency (variance across runs), inter-model agreement, and alignment with established vulnerability scores using Spearman's ρ (Spearman 1904), a non-parametric statistic that measures the strength and direction of association between two variables by comparing the ranked order of the values. We now explain the two tasks in detail.

3.1 Pairwise Comparisons

Kube et al. (2022) drew on St. Louis's Homeless Management Information System (HMIS) to create ten hypothetical pairs of tabular data of households. These data tables are already public and published on the web, thus we are also able to use cloud-based proprietary models for evaluation. The data contained information on demographics (number of adults and children, monthly income, disability status), service requests, and a precomputed risk score (low/medium/high probability of reentering homelessness within two years). In that experiment, human subjects mimicked street-level bureaucrats who were told that two types of assistance were available: 1) **Transitional Housing (TH)**, which is a longer-term and more intensive support program, reserved for households with greater needs; and, 2) **Emergency Shelter (ES)** – a less intensive option that provides a place to stay overnight. The lay humans were then presented with features of household pairs and asked to prioritize one for TH, with the understanding that the other would receive ES.

We run a similar set of experiments with LLMs. We test four information conditions:

- **No Prediction:** The model sees only the household data table (no reentry to homelessness risk estimates).
- **Only Prediction:** The model sees only the outcome prediction (no data table).
- **Prediction First:** The model first generates its own outcome prediction from the data tables provided and *then* makes the allocation decision.
- **Shared Prediction:** The model sees both the household data table *and* a pre-computed "outcome prediction" (the probability – labeled low/medium/high –that the household will reenter homelessness services within two years, conditional on receiving each intervention).

These four setups allow us to identify how access to risk estimates affects the LLM's choices. Then, for each of the ten household pairs, we issue ten independent prompts per

```
<begin_of_text>
<start_header_id>system<end_header_id>
Communities across the US provide homeless services that respond to household requests for assistance in securing
stable housing. Households call a hotline to request help and provide basic demographic details such as household
size, monthly income, disability support status, and prior residence. Service providers must decide what services
to offer based on household need and available resources.
Two key services are:
1. Emergency Shelter - provides an immediate response to homelessness.
2. Transitional Housing - provides long-term housing along with case management that may include treatment for
disabilities or health conditions.
Emergency Shelter requires fewer resources, while Transitional Housing is more resource-intensive and scarcer due
to budget constraints. Households unable to access Transitional Housing typically receive Emergency Shelter or
remain in it until other options become available.
Researchers have developed models using the hotline demographic information to predict whether a household will
need services again within 2 years. All households are assumed to have a lower or equal likelihood of needing
future services if given Transitional Housing compared to Emergency Shelter.
During the following activities, you will see the same information provided to service providers and make
decisions on how to allocate homeless services.
<eot_id>
Household Data:
<insert data for comparison>
<start_header_id>user<end_header_id>
Decide which household to prioritize for transitional housing. Remember, you must choose different households for
Transitional Housing and Emergency Shelter! Your output should strictly be in the following format (with nothing
else outside the template):
Emergency Shelter <Household-#>. Transitional Housing: <Household-#>.
Replace # with your chosen household number.
<prompt>
<start_header_id>assistant<end_header_id>
```

Figure 1: The 'base prompt' used in both the pairwise comparison and ranking tasks. Depending on the experiment type, we replace the blue placeholder text with either the tabular data (for the pairwise comparison) or the questionnaire responses (for the ranking task).

condition for every LLM that we test. Each response is labeled *outcome-oriented* (favoring the lower-risk household) or *vulnerability-oriented* (favoring the higher-need household). We then compute an *outcome score* for each pair:

$$\text{Outcome Score} = \left\lceil \frac{\text{\# of outcome decisions}}{\text{total decisions}} \times 100 \right\rceil \tag{1}$$

An outcome score of 0 means that all ten runs were vulnerability-oriented; a score of 100 means that all were outcome-oriented. The exact prompt and details on the LLMs used are provided in Section 3.3.

3.2 Ranking Task

The second set of experiments evaluates how LLMs perform in the core bureaucratic task of coordinated entry – that is, producing a complete ordering of a given set of households by relative vulnerability. In practice, such rankings drive the allocation of scarce resources according to socially and politically defined priorities. We compare the LLM-generated rankings with those produced by established assessment instruments: the VI-SPDAT for single adults, families, and youth (hereafter denoted VI_{SA}, VI_{F}, and VI_{Y} respectively) (OrgCode Consulting Inc. and Community Solutions 2015b,a; OrgCode Consulting Inc. et al. 2015), and the Risk/Medical Frailty Score (RMFS), another tool used in St. Louis homeless services (Fefer 2022). Our dataset contains, for each household, the VI-SPDAT acuity score, the RMFS frailty score, and the full set of raw questionnaire responses. These questions aim to assess vulnerability. Examples include *How long has it been since you lived in permanent, stable housing?*, and *Where do you sleep most frequently?* There are in total 35 questions in VI_{SA}, 41 questions in VI_{Y}, 54 questions in VI_{F}, and 25 in RMFS. The full set of questions used to collect the data is included in Appendix B and C. Table 3 documents the number of available assessments. We refer to these scores collectively as bureaucratic rankings.

It is widely acknowledged that asking an LLM to rank directly yields unreliable global orderings. We follow the recommendation of Qin et al. (2024) and instead use a methodology that constructs the ranking from pairwise comparisons using Rank Centrality (Negahban, Oh, and Shah 2017). We start by comparing each pair of households with the LLM and recording if i is preferred to j. For each ordered pair (i, j), we compute the raw "win fraction" and treat that as the weight on the edge between i and j. To convert raw weights into transition probabilities for a random walk, we *normalize* each household's outgoing edges so that they sum to one by simply dividing each edge-weight by the total

of that node's outgoing weights. This ensures each node's edges form a valid probability distribution.

We then interpret the normalized edge weights as transition probabilities of a Markov chain and compute the stationary distribution of this Markov chain to recover a global score for each household. The rank centrality algorithm is closely connected to the *Bradley–Terry–Luce (BTL)* model (Bradley and Terry 1952; Luce et al. 1959), which assumes that each household i has a positive "strength" θ_i, and that in any single comparison the probability i beats j is $\frac{\theta_i}{\theta_i+\theta_j}$. Under this model, the stationary distribution of the normalized comparison graph converges to the true normalized strengths, and it can be shown that only on the order of $\theta(N \log N)$ independent comparisons are required to recover accurate rankings with high probability (Negahban, Oh, and Shah 2017). Since exhaustively comparing all $\frac{1}{2}N(N-1)$ pairs is infeasible, we instead include each (unordered) pair independently with probability 0.4, yielding about $0.2\,N(N-1)$ comparisons—comfortably above the $\theta(N \log N)$ threshold—while keeping compute and LLM calls tractable (see Table 3). Importantly, throughout this entire process, we never send data to the cloud. All computations and LLM inferences are performed locally, ensuring strict data privacy and confidentiality. Our decision to avoid cloud-based LLM calls constrains us to utilize only models locally available, leading us to employ different sets of models for the pairwise comparison task and the ranking task.

Each time we sample an edge, the LLM decides which household "wins", that is, the model judges it more vulnerable. Thus, the sampled edge accrues a higher transition probability in the rank-centrality graph. Consequently, when we compute the stationary distribution of this Markov chain, those households deemed most vulnerable by the LLM naturally receive the highest global scores in the final aggregated ranking.

3.3 LLMs and Prompting

The details of the LLMs that we use (and their aliases) are detailed in Table 1.

Pairwise Comparison	
Full Model Name	**Alias**
DeepSeek-R1-Distill-Llama-70B (DeepSeek-AI 2025)	DS-R1-L70B
Llama-3.3-70B-Instruct (Touvron et al. 2023)	LL-3.3-70B
Gemini-2.0-flash-thinking-exp (Google 2025a)	G2-FTX
Gemini 2.0 flash (Cloud 2025)	G2-FLS
Gemma-3-27b-it (Google 2025b)	GM3-27B
GPT-4 (OpenAI 2023b)	GPT-4
GPT-3.5-turbo (OpenAI 2023a)	GPT-3.5T
GPT-4.0125-preview (OpenAI 2024)	GPT-4-P125
Ranking Task	
Llama-3-8B-Instruct (AI@Meta 2024)	LL-3-8B
deepseek-llm-7b-chat (DeepSeek-AI 2024)	DS-7B

Table 1: The LLMs used in our experiments and their shorthand aliases. Models are grouped by task.

Pairwise Comparisons: For the pairwise comparison task, because the data is publicly available in Kube et al. (2022) structured as *blocks* numbered one through ten, we evaluated both open-source and proprietary LLMs. The open source models included DS-R1-L70B and LL-3.3-70B, while the proprietary models comprised G2-FTX, G2-FLS, GM3-27, GPT-4, GPT-3.5T, and GPT 4-P125. We take the data from Kube et al. (2022) and convert them into a json-like string and use a base prompt (see Figure 1) to replace the households data for each pair, such that we are working with the same prompt for every pairwise comparison.

Ranking Task: As mentioned above, we sample approximately 40% of the undirected edges for the rank centrality algorithm. In order to derive edge weights, we need the outcomes of pairwise comparisons. We reuse the same base prompt from the pairwise experiments, but replace the data now with the responses from each household's raw questionnaire. Given the proprietary nature of our questionnaire data and to ensure that information never leaves our secure environment, we only use open-source LLMs that can run locally. Considering the computational requirements for the large number of prompts necessary, we chose two 7-billion-parameter models – LL-3-8B and DS-7B – that come from different development ecosystems (the former rooted in Western open-source communities, the latter from Chinese academic collaborations). This allows us to explore whether varying design philosophies or cultural contexts affect vulnerability judgments. For each LLM, we execute the full rank-centrality pipeline twice, producing two independent ranked lists per model to quantify the consistency of their rankings.

All local models were run on a cluster compute node containing two 24GB VRAM Nvidia A30 GPUs using vLLM's api-server's OpenAI compatible endpoints. Both models were deployed using publicly available pre-trained weights from the Hugging Face Hub, and no additional fine-tuning was performed. We opted not to fine-tune the models because our core research question is: Which types of households do current state-of-the-art LLMs, without any form of fine-tuning, consider more vulnerable / needing a more intensive intervention? Evaluating "vanilla" models tells us how well they transfer to a novel, socially important task without bespoke adaptation, which mirrors many real-world settings where practitioners lack labeled data, compute budgets, or permission to retrain proprietary APIs. This is the sense in which we can think of this "vibe prioritization" as an analogy to "vibe coding," where humans ask AI models to produce code with minimal intervention or oversight. We note that we access the proprietary models via API tokens, but follow the same experimental pipeline as our local models.

For most models, we use regular expressions to parse the LLM outputs. DS-7B, however, produces overly verbose responses, so we route its output through a secondary LLM (G2-FLS) to distill the household selections. On the rare occasions when the models failed to yield a coherent choice, we re-prompt until the LLM unambiguously identifies which household should receive the more intensive intervention.

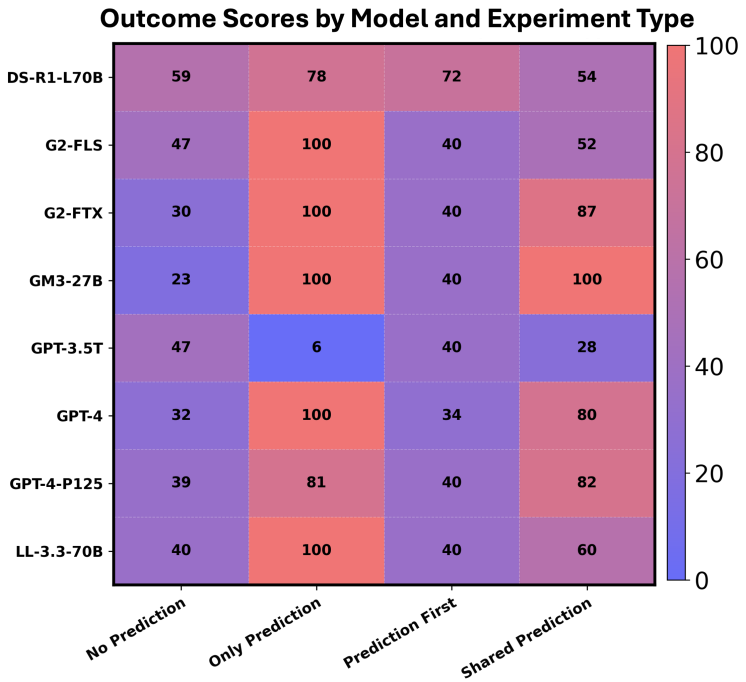


Figure 2: Outcome-oriented scores from our pairwise comparison task, broken down by model and information configuration. In the heatmap, red denotes higher scores (more outcome-oriented) and blue denotes lower scores (more vulnerability-oriented). These results reveal substantial variability in LLM behavior across different models—and show that, overall, LLMs perform similarly to non-expert humans when making prioritization judgments.

4 Results

4.1 Pairwise Comparisons

We repeatedly prompted each LLM in the pairwise comparison experiments until it produced a coherent, categorizable response.[1] Our main goal here is to compare LLM prioritization behavior with that observed in lay humans on carefully constructed data. This is similar in spirit to much of the ongoing work on LLMs making moral judgments in different domains (Kim et al. 2025; Jha et al. 2024; Dillion et al. 2025; Nazi and Peng 2024). We calculate the outcome scores of the decisions rendered by different LLMs under different configurations and average them over household pairs. Figure 2 presents the outcome scores of the results, where each column corresponds to the different experiment types and each row represents the LLMs tested for the task. Recall that scores range from 0 (completely vulnerability oriented, colored blue) to 100 (completely outcome oriented, colored red).

No Prediction. When no outcome information is provided (only tabular data), LLMs are, on average, inconsistent in terms of whether they present as vulnerability- or outcome-oriented in decision-making. This is not due to inconsistency across runs within a specific pairwise comparison, but instead due to making different decisions across pairs (see Figure 3). This is similar to the behavior Kube et al observe in humans, indicating that LLMs may have similar difficulty in assessing vulnerability or outcomes from the raw data.

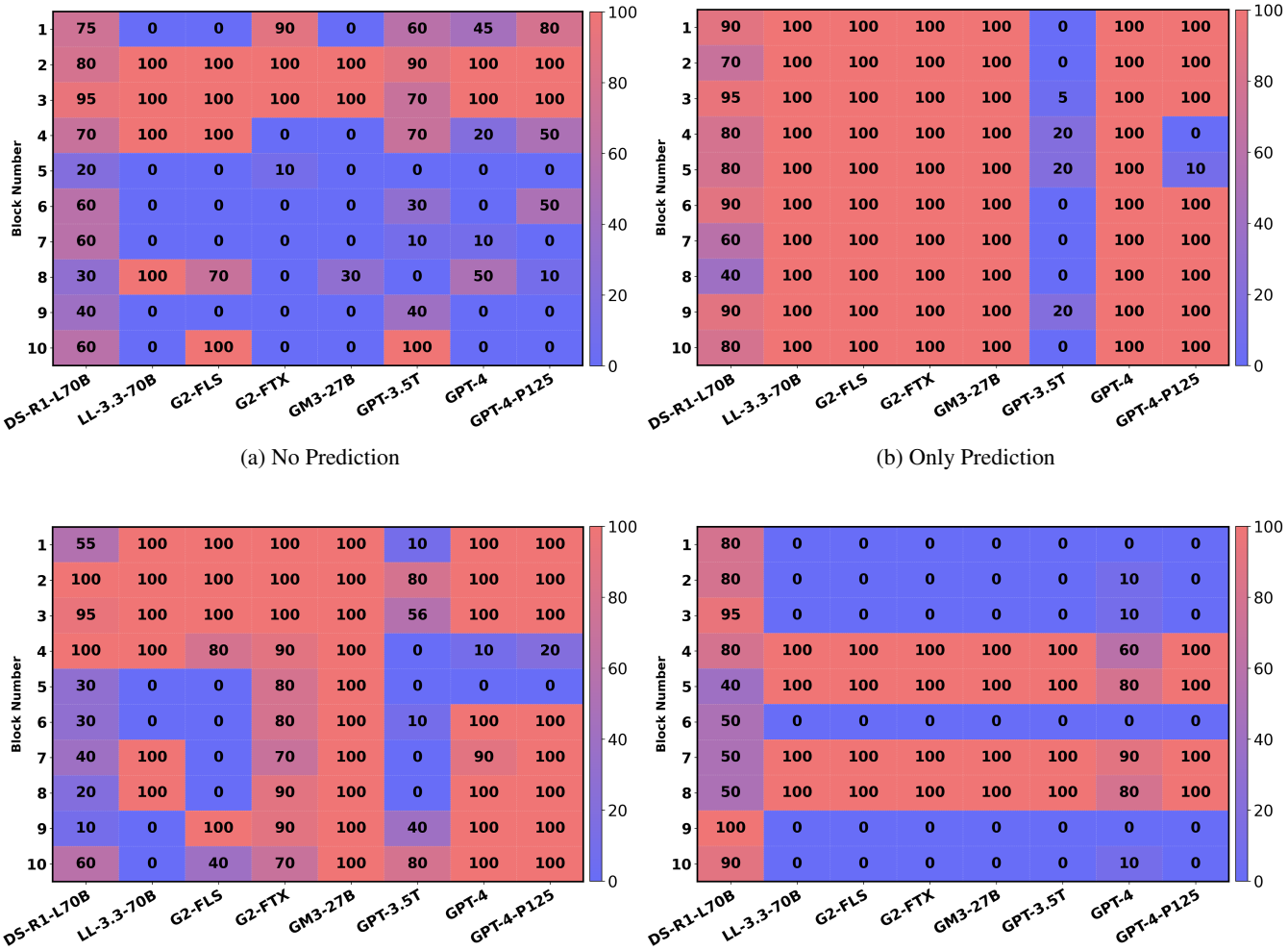
Only Prediction. When *only* outcome information is provided, similar to humans, each LLM is highly consistent in manifesting either a vulnerability or an outcome orientation. Notably, GPT-3.5T is the only LLM that is vulnerability-oriented, with an outcome score of 6.

Prediction First. Kube et al. provide evidence that prior exposure to quantitatively predicting outcomes can alter the orientation of some (about one third) of humans from vulnerability to outcome. For the LLMs we tested, the outcome orientation is not substantively different from the No Prediction condition, indicating that there is no such "framing" effect here.

Shared Prediction. This experiment demonstrates substantial heterogeneity between different LLMs. In human behavior, when comparing No Prediction to Shared Prediction, Kube et al. note that the addition of outcome information allows each human to manifest their "true" type (equivalent, in our setting, to whether or not they appear vulnerability- or outcome-oriented in the Prediction Only task). In fact, we see a similar movement of the LLMs. Each moves towards their Prediction Only score, with the exception of DS-R1-L70B. GS-FLS is somewhat limited in its movement.

Together, the results indicate substantial heterogeneity in

[1] DS-R1-L70B often diverged into tangents, and GPT-4 initially refused on grounds of unqualification; nonetheless, all models ultimately yielded categorizable outputs after repeated prompting.



(a) No Prediction

(b) Only Prediction

(c) Shared Prediction

(d) Prediction First

Figure 3: Pairwise comparison results under four conditions: no predictions (Fig. 3a); only outcome predictions (Fig. 3b); shared prediction (Fig. 3c); and prediction first (Fig. 3d). Each block corresponds to a household. Scores in each cell are calculated using Equation 1 from outcome decisions averaged across 10 different runs. These results indicate that LLM decision inconsistencies do not occur within the block level, but instead across blocks.

LLM behavior across different models. They also show that LLMs in general are not dissimilar from lay humans (with no specific subject matter expertise) in making judgments on the problem of prioritization.

4.2 Rankings

Before presenting results for the ranking task, we address how outcome score ties affect score-based rankings, since multiple households may receive the same score. To evaluate the impact of random tie breaking on the overall order, we generate ten tie-broken variants for each assessment tool by permuting tied entries uniformly at random. In all cases, Spearman's ρ between each pair of variants remained above 0.98, indicating that arbitrary tie resolutions do not significantly alter the global rank structure. To provide a complete statistical picture, 95% confidence intervals for this and all following analyses are in Table 8, Appendix A[2].

We now turn to examining LLM judgments in comparison with the judgment of experts on a task involving real-world data.

LLM Internal Consistency After generating the rankings induced by pairwise comparisons performed by the LLMs, as described in the Methods section, we check for consistency between two different executions of the rank-centrality pipeline. We calculate Spearman's rank correlation between different rankings produced by DS-7B and LL-3-8B and present them in Table 4.

Rankings produced by LL-3-8B demonstrate weak positive correlations among themselves. DS-7B rankings show moderate positive correlation on the VI_{SA} (single adults)

[2]Full paper on arXiv (Pokharel et al. 2025)

LLaMA Ranking 1 on VI_{SA}

Feature	Coefficient
Current living arrangement: Shelter/Outdoors	-0.3796
Current living arrangement: Car	0.0216
Current living arrangement: No answer	0.0517
Attempts of self-harm? No answer	0.1025
Number of past incarcerations > 10	0.1296

LLaMA Ranking 2 on VI_{SA}

Feature	Coefficient
Current living arrangement: No answer	-0.0632
Current living arrangement: Job corp	-0.0290
Current living arrangement: Car	-0.0283
Current living arrangement: Hotel	-0.0236
Current living arrangement: Couch-surfing/Shelter	0.5033

LLaMA Ranking 1 on VI_{F}

Feature	Coefficient
Current living arrangement: No answer	-0.0469
Current living arrangement: Couch-surfing	-0.0228
Current living arrangement: Place-to-place	-0.0135
Current living arrangement: Car	-0.0081
Current living arrangement: Couch-hopping	0.5707

LLaMA Ranking 2 on VI_{F}

Feature	Coefficient
Current living arrangement: No answer	-0.0145
Expected additional children to join after receiving housing: No answer	0.0619
Parents in family > 3	0.0757
Nights spent incarcerated > 3	0.1110
Current living arrangement: Car	0.1511

LLaMA Ranking 1 on VI_{Y}

Feature	Coefficient
Reselling prescribed painkiller? No answer	-0.0558
Number of ER visits > 10	0.0929
Avoids seeking medical help? No answer	0.0969
Current living arrangement: Boyfriend's house	0.0981
Physical health impacted homelessness: No answer	0.1353

LLaMA Ranking 2 on VI_{Y}

Feature	Coefficient
Did abusive relationship cause homelessness? No answer	-0.1244
Marijuana use before age of 12? No answer	-0.0588
Chronic health issues? No answer	-0.0505
Time since permanent housing? No answer	0.0547
Homelessness caused by running away? No answer	0.1200

Table 2: Top 5 most influential question-answer pairs for two independent LLaMA rankings. The low overlap and inconsistent polarity of the few shared features (highlighted in blue) demonstrate the model's instability. Questions are shortened for brevity (see Appendix B (Pokharel et al. 2025) for full versions.)

Assessment Type	Dataset Name	# Samples	Min Corr (SPDAT)	Min Corr (RMFS)
VISPDAT	VI_{SA}	325	0.98350	0.99360
VI-FSPDAT	VI_{F}	698	0.98419	0.99216
TAY-VISPDAT	VI_{Y}	561	0.98692	0.98853

Table 3: Stability of baseline rankings induced by bureaucratic scores. The table lists the number of households for each dataset. The last two columns report the minimum Spearman's ρ from 10 independent random tie-breaking trials on the SPDAT and RMFS scores. These high values ($\rho > 0.98$) confirm that these baseline rankings are stable for comparisons with other rankings.

Assessment Data	Between LLaMA Rankings	Between DeepSeek Rankings
VI_{SA}	0.24692	0.47951
VI_{F}	0.12805	0.27571
VI_{Y}	0.19748	0.02719

Table 4: Spearman's ρ between two rankings produced in independent runs by the same model. The generally low correlations indicate that these LLMs produce significantly different rankings when run on the exact same data.

dataset, weak correlation in the VI_{F} (families) dataset, and little to no correlation on the VI_{Y} (youth) dataset. The low degree of correlations presented in Table 4 indicates clearly that LLMs are inconsistent in their vulnerability determinations from questionnaire data commonly used by service providers.

Comparing LLMs with Bureaucratic Scores In the US, homelessness services must implement a prioritization system that considers household vulnerability in allocating scarce housing resources, according to federal policies. The St. Louis community has used two different scoring systems, the VI-SPDAT (which has three different forms for three different populations) and the RMFS. How do LLM rankings compare with these well-established systems?[3]

Table 6 shows that the rankings by LLMs have near-zero correlation with the VI-SPDAT and RMFS rankings, while LL-3-8B rankings show negative correlation with VI-SPDAT and RMFS. This demonstrates that LLM ranking judgments diverge substantially from those of the vetted bureaucratic ranking systems adopted in coordinated entry systems.

What Features Drive LLM Decisions? To probe deeper into the LLM decision-making criteria, we examine which questions receive higher focus from the models. Each question q in the HMIS questionnaire is accompanied by an answer $\hat{a}$ chosen from a set A_q (which could include multiple options written into a catch-all category as well) by the applicant household. For each question answer pair $(q, \hat{a} \in A_q)$, we generate binary categorical features in the form $(q_{\hat{a}}) = 1$ and $(q_a) = 0$ for all $a \in A_q$ and $a \neq \hat{a}$. Next, we train an ordinal logistic regressor with all thresholds (Rennie and Srebro 2005) with L2 regularization (coefficient 1) to map these categorical features generated from the questionnaire to the ranks produced by LL-3-8B and DS-7B. We analyze the normalized feature coefficients assigned to each categorical question to better understand which question-answer combinations impact the LLM decisions the most. The LLMs seem to focus on different question-answer pairs in different runs on the same data. As

[3]Note that there are many issues researchers and practitioners have identified with such bureaucratic systems, including various kinds of bias, inconsistencies, etc. However, the decision to use these scoring systems at least arises from complex social and political processes that reflect the collective decision of society on how to allocate public benefits and harms.

DeepSeek Ranking 1 on VI_{SA}

Feature	Coefficient
Current living arrangement: Motel/Hotel	-0.1498
Current living arrangement: Church	-0.1057
Current living arrangement: Family's house	-0.0144
Prescription medication use? No answer	0.1475
Current living arrangement: Couch-surfing/Shelter	0.3555

DeepSeek Ranking 2 on VI_{SA}

Feature	Coefficient
Number of past incarcerations $>$ 10	-0.2378
Current living arrangement: No answer	-0.0400
Current living arrangement: Motel/Hotel	-0.0292
Current living arrangement: Job corp	-0.0189
Current living arrangement: Couch-surfing/Shelter	0.3229

DeepSeek Ranking 1 on VI_{F}

Feature	Coefficient
Homeless during previous year? No answer	-0.0003
Nights spent incarcerated $>$ 3	0.0001
Parents in family $>$ 3	0.0002
Current living arrangement: No answer	0.0002
Expected additional children to join after receiving housing: No answer	0.0559

DeepSeek Ranking 2 on VI_{F}

Feature	Coefficient
Current living arrangement: Relative's house	-0.2136
Current living arrangement: Couch-hopping	-0.0263
Homeless during previous year? No answer	-0.0125
Current living arrangement: Couch-surfing	0.0107
Current living arrangement: No answer	0.0220

DeepSeek Ranking 1 on VI_{Y}

Feature	Coefficient
Prescription medication use? No answer	-0.0472
Alcohol/drug use related eviction? No answer	-0.0472
History of head injury? No answer	0.0500
Marijuana use before age of 12? No answer	0.0552
Used a crisis service? No answer	0.1150

DeepSeek Ranking 2 on VI_{Y}

Feature	Coefficient
# Ambulance rides to hospital $>$ 10	0.0371
Alcohol/drug use impact housing? No answer	0.0505
Chronic health issues? No answer	0.0579
# of interactions with police: 6	0.0921
Current living arrangement: Boyfriend's house	0.0927

Table 5: Top 5 most influential question-answer pairs for two independent DeepSeek rankings. Similar to Table 2, the low overlap and inconsistent polarity of the few shared features (highlighted in blue) demonstrate the model's instability. Questions are shortened for brevity (see Appendix B (Pokharel et al. 2025) for full versions.)

Assesment Subpopulation	Ranking Criterion	LLaMA Ranking 1	LLaMA Ranking 2	DeepSeek Ranking 1	DeepSeek Ranking 2
VI_{SA}	VI-SPDAT	-0.18009	-0.18640	0.12056	0.07167
	RMFS	-0.00952	0.12469	0.16421	0.14639
VI_{F}	VI-SPDAT	-0.06960	-0.19930	0.21722	0.17079
	RMFS	0.08557	-0.01491	0.12232	0.08018
VI_{Y}	VI-SPDAT	-0.15286	-0.16309	0.06188	0.04425
	RMFS	-0.01904	-0.02552	0.06974	0.02503

Table 6: Correlation between LLM and bureaucratic rankings. We report the Spearman's ρ between each of the four LLM-generated rankings and the two baseline bureaucratic rankings. The correlations are close to zero, showing that LLM rankings unreliable fail to capture the vulnerability principles embedded in existing systems.

illustration, the top five most important features for LL-3-8B and DS-7B are reported in Tables 2 and 5, respectively. Even when features appear in both rankings, they have inconsistency in the polarity. This shows a struggle to identify consistent criteria for judging household vulnerability. Interestingly, refusals to answer questionnaire items and unrecorded answers influence both LL-3-8B and DS-7B decisions on VI_{F} and VI_{Y} subpopulations. Household current living arrangements seem to be a recurring factor impacting the decisions. These may be intuitively aligned with what human perception, however, note that there is considerable inconsistency in which living conditions impact the LLMs most and whether the impact is favorable or adverse to recommending more intensive intervention.

Comparing LLMs and Bureaucratic Scores with Human Decisions Although caseworkers are supposed to use bureaucratic scores in determining assignments, street-level bureaucrats inherently can exercise judgment and discretion in deciding who receives services (Lipsky 1980; Pokharel, Das, and Fowler 2024; Kawakami et al. 2022). Therefore, it is plausible that LLM judgments match up better with human judgments than bureaucratic scores. In this case, since we only have data on whether or not households received services, we can test how predictive the different rankings are of receipt of intensive services (Rapid Rehousing, Transitional Housing, or Permanent Supportive Housing) by computing the area under the ROC curve for predicting receipt of one of these services for each of the rankings.

Assessment Data	# Positive Samples	VI-SPDAT Ranking	RMFS Ranking	LLaMA Ranking 1	LLaMA Ranking 1	DeepSeek-R1 Ranking 1	DeepSeek-R1 Ranking 2
VI_{SA}	68	0.68511	0.6042	0.61667	0.62434	0.51957	0.51247
VI_{F}	116	0.53022	0.5253	0.51518	0.5374	0.51271	0.54361
VI_{Y}	103	0.50674	0.56923	0.51015	0.53258	0.54498	0.50854

Table 7: Predictive validity of rankings for service allocation. The table reports the ROC AUC for each ranking's ability to predict the receipt of an intensive housing intervention, with the number of positive samples (recipients) noted. All rankings are weak predictors, and LLM-generated rankings offer no improvement over existing bureaucratic tools in forecasting real-world decisions on prioritization.

Table 7 presents the revealing results. The bureaucratic scores are only predictive for single adults; LLaMA is also somewhat predictive in this case, but less so than the VI-SPDAT scores. Meanwhile, for families and youth, neither the bureaucratic scores nor the LLM rankings predict actual caseworker decisions. There are many factors determining if a household receives services, especially the availability of appropriate housing at the time of assessment. We should not expect very high AUC values, but the *differences* are illuminating. In short, LLM rankings show virtually no correlation with well-known scoring systems and are less predictive of the actual decision-making of human experts.

5 Conclusion

In this work, we examine the viability of deploying off-the-shelf LLMs as de facto street-level bureaucrats in the context of homelessness resource allocation ("vibe prioritization"). Through two complementary experimental tasks – pairwise comparisons and global ranking reconstructed via Rank Centrality – we systematically evaluate (1) the internal consistency of LLM judgments, (2) their agreement with established bureaucratic scoring systems, and (3) their alignment with actual caseworker decisions.

Our pairwise comparison experiments (Section 4.1) demonstrate that, when presented with household profiles alone, LLMs exhibit variability reminiscent of non-expert human subjects: different LLMs are different in whether they naturally take vulnerability-oriented decisions or outcome-oriented ones, but relatively consistent within themselves *as long as they receive explicit risk information* (computed by an external conventional machine learning model) (Figures 2 and 3). They are, however, inconsistent in decision-making when not provided explicit risk information.

The ranking task (Section 4.2) reveals serious limitations. Neither LL-3-8B nor DS-7B produces stable vulnerability orderings: Spearman correlations between independent runs range from near zero to moderate (Table 4). This means that two executions of the same model can generate markedly different rankings for the same set of households, calling into question the robustness and reliability required for high-stakes decision-making. Furthermore, LLM-generated rankings show negligible, and even negative, correlation with bureaucratic scoring systems (Table 6), as well as being less predictive of actual homelessness service allocations (Table 7). Taken together, the results signal that LLMs, without domain-specific adaptation, fail to replicate either formalized points-based systems or the nuanced discretion of experienced caseworkers.

These findings raise important concerns for policymakers and practitioners considering integrating LLMs into high-stakes social service workflows. First, the pronounced inconsistency suggests that automated judgments may vary dramatically depending on incidental factors – model choice, prompt phrasing, or inference seed – undermining fairness and transparency. Second, the disconnect from established vulnerability metrics risks both inefficient resource deployment and erosion of community-driven prioritization principles. This misalignment indicates that LLMs may rely on superficial patterns rather than the policy-driven indicators of vulnerability, potentially exacerbating service gaps and deepening inequities. They are also not consistent with any of the prioritization principles that have been developed across many different domains of local justice.

Finally, the limited ability of models to predict real allocations indicates insufficient alignment with the tacit expertise of street-level bureaucrats. As they are, without careful consideration or thorough testing, off-the-shelf LLMs fail to capture the local justice principles and context-sensitive discretion that street-level bureaucrats embed into their decisions (Lipsky 1980; Elster 1992), emphasizing that LLM-driven approaches as of now cannot supplant embodied professional judgment.

This study underscores the necessity of rigorous context-grounded evaluation before automating public resource allocation decisions. Future work must investigate whether practices like fine-tuning on localized caseworker data, integrating multi-modal client information, or embedding human-in-the-loop safeguards can enhance LLM reliability. Equally crucial is the need for ongoing engagement with service providers and stakeholders to ensure that algorithmic systems respect norms of justice and preserve the discretionary judgment essential to public service. Importantly, any AI augmentation in this domain must reflect the nuanced trade-offs and moral frameworks that have evolved through community engagement and political processes.

In conclusion, while LLMs offer promising avenues for augmenting decision support, the evidence from our study cautions against their wholesale replacement of street-level bureaucrats in homelessness services. Realizing the potential of AI in this domain requires carefully calibrated hybrid systems, where machine recommendations are transparent, consistently calibrated, and subject to human oversight, rather than unmediated reliance on present-day language models.

Acknowledgments

We are grateful for support from NSF Award 2533162. We also thank the various community partners who helped conceptualize the challenges facing the delivery of homeless services, as well as their ongoing efforts to support local families.

References

Acharya, A.; Shrestha, S.; Chen, A.; Conte, J.; Avramovic, S.; Sikdar, S.; Anastasopoulos, A.; and Das, S. 2024. Clinical risk prediction using language models: benefits and considerations. *Journal of the American Medical Informatics Association*, 31(9): 1856–1864.

AI@Meta. 2024. Llama 3 Model Card. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Been, V.; O'Regan, K.; Waldinger, D.; and Center, N. F. 2018. Allocation of the limited subsidies for affordable housing. *New York Times*.

Bender, E. M.; and Hanna, A. 2025. On the Very Real Dangers of the Artificial Intelligence Hype Machine. *Literary Hub*.

Bonett, D. G.; and Wright, T. A. 2000. Sample Size Requirements for Estimating Pearson, Kendall and Spearman Correlations. *Psychometrika*, 65(1): 23–28.

Bradley, R. A.; and Terry, M. E. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4): 324–345.

Brown, M.; Cummings, C.; Lyons, J.; Carrión, A.; and Watson, D. P. 2018. Reliability and validity of the Vulnerability Index-Service Prioritization Decision Assistance Tool (VI-SPDAT) in real-world implementation. *Journal of Social Distress and the Homeless*, 27(2): 110–117.

Cholongitas, E.; Germani, G.; and Burroughs, A. K. 2010. Prioritization for liver transplantation. *Nature Reviews Gastroenterology & Hepatology*, 7(12): 659–668.

Cloud, G. 2025. Gemini 2.0 Flash Model Documentation. https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash. Accessed: 2025-05-18.

Conitzer, V. 2024. Why should we ever automate moral decision making? *arXiv preprint arXiv:2407.07671*.

Cronley, C. 2022. Invisible intersectionality in measuring vulnerability among individuals experiencing homelessness–Critically appraising the VI-SPDAT. *Journal of Social Distress and Homelessness*, 31(1): 23–33.

DeepSeek-AI. 2024. DeepSeek LLM: Scaling Open-Source Language Models with Longtermism. *arXiv preprint arXiv:2401.02954*.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

DeLong, E. R.; DeLong, D. M.; and Clarke-Pearson, D. L. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3): 837–845.

Dillion, D.; Mondal, D.; Tandon, N.; and Gray, K. 2025. AI language model rivals expert ethicist in perceived moral expertise. *Scientific Reports*, 15(1): 4084.

Elster, J. 1992. *Local Justice: How Institutions Allocate Scarce Goods and Necessary Burdens*. Russell Sage Foundation.

Fefer, N. S. 2022. 2022 Consolidated Annual Performance and Evaluation Report.

Google. 2025a. gemini-2.0-flash-thinking-exp Experimental Model Documentation. https://ai.google.dev/gemini-api/docs/models. Accessed: 2025-05-18.

Google. 2025b. gemma-3-27b-it Model Card. https://huggingface.co/google/gemma-3-27b-it. Accessed: 2025-05-18.

Hasjim, B. J.; Azafar, G.; Lee, F.; Diwan, T. S.; Raju, S.; Gross, J. A.; Sidhu, A.; Ichii, H.; Krishnan, R. G.; Mamdani, M.; Sharma, D.; and Bhat, M. 2024. The AI Agent in the Room: Informing Objective Decision Making at the Transplant Selection Committee. *medRxiv*.

Hosseini, H.; and Khanna, S. 2025. Distributive Fairness in Large Language Models: Evaluating Alignment with Human Values. *arXiv*.

Jha, A.; Mann, P.; Tiwari, A.; Kadian, K.; and Sharma, A. 2024. Decoding Ethics: Proficiency of LLMs in Addressing Moral Dilemmas. In Illés, Z.; Verma, C.; Gonçalves, P. J. S.; and Singh, P. K., eds., *Proceedings of International Conference on Recent Innovations in Computing*, 593–605. Singapore: Springer Nature Singapore. ISBN 978-981-97-3442-9.

Johnson, R. A.; and Zhang, S. 2022. What is the bureaucratic counterfactual? Categorical versus algorithmic prioritization in US social policy. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1671–1682.

Kawakami, A.; Sivaraman, V.; Cheng, H.-F.; Stapleton, L.; Cheng, Y.; Qing, D.; Perer, A.; Wu, Z. S.; Zhu, H.; and Holstein, K. 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*.

Kim, J.; Kwon, J.; Vecchietti, L. F.; Oh, A.; and Cha, M. 2025. Exploring Persona-dependent LLM Alignment for the Moral Machine Experiment. *arXiv*.

Kube, A.; Das, S.; Fowler, P. J.; and Vorobeychik, Y. 2022. Just resource allocation? How algorithmic predictions and human notions of justice interact. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, 1184–1242.

Kube, A. R.; Das, S.; and Fowler, P. J. 2023. Fair and efficient allocation of scarce resources based on predicted outcomes: implications for homeless service delivery. *Journal of Artificial Intelligence Research*, 76: 1219–1245.

Kuo, T.-S.; Shen, H.; Geum, J.; Jones, N.; Hong, J. I.; Zhu, H.; and Holstein, K. 2023. Understanding Frontline Workers' and Unhoused Individuals' Perspectives on AI Used in Homeless Services. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. New York, NY, USA: Association for Computing Machinery.

Lipsky, M. 1980. *Street-level bureaucracy: Dilemmas of the individual in public services*. New York: Russell Sage Foundation.

Luce, R. D.; et al. 1959. *Individual choice behavior*, volume 4. Wiley New York.

Maitra, S.; Sleep, L.; Henman, P.; Fay, S.; and Conversation, T. 2025. AI is being used in social services—but we must make sure it doesn't traumatize clients. *Phys Org*.

Murray, H.; Kim, B. H.; Lee, I.; Byun, J.; Yogatama, D.; and Micha, E. 2025. Ethical AI on the Waitlist: Group Fairness Evaluation of LLM-Aided Organ Allocation. *arXiv*.

Nazi, Z. A.; and Peng, W. 2024. Large language models in healthcare and medical domain: A review. In *Informatics*, volume 11, 57. MDPI.

Negahban, S.; Oh, S.; and Shah, D. 2017. Rank Centrality: Ranking from Pairwise Comparisons. *Operations Research*, 65(1): 266–287.

OpenAI. 2023a. gpt-3.5-turbo Model Documentation. https://platform.openai.com/docs/models/gpt-3-5-turbo. Accessed: 2025-05-18.

OpenAI. 2023b. gpt-4 Model Documentation. https://platform.openai.com/docs/models/gpt-4. Accessed: 2025-05-18.

OpenAI. 2024. gpt-4-0125-preview Model Card. https://platform.openai.com/docs/models. Accessed: 2025-05-18.

OrgCode Consulting Inc.; and Community Solutions. 2015a. Family Service Prioritization Decision Assistance Tool (F-SPDAT): U.S. Version 2.0. https://everyonehome.org/wp-content/uploads/2016/02/F-SPDAT-2.0-Families.pdf. Accessed May 17, 2025.

OrgCode Consulting Inc.; and Community Solutions. 2015b. Vulnerability Index–Service Prioritization Decision Assistance Tool (VI-SPDAT): Prescreen Triage Tool for Single Adults. https://everyonehome.org/wp-content/uploads/2016/02/VI-SPDAT-2.0-Single-Adults.pdf. Accessed May 17, 2025.

OrgCode Consulting Inc.; Corporation for Supportive Housing; Community Solutions; and Eric Rice. 2015. Next Step Tool for Homeless Youth (TAY-VI-SPDAT): U.S. Version 1.0. https://letsendhomelessness.org/wp-content/uploads/2018/07/TAY-VI-SPDAT-v1-0-w.-Intro-Script.pdf. Accessed May 17, 2025.

Pokharel, G.; Das, S.; and Fowler, P. 2024. Discretionary Trees: Understanding Street-Level Bureaucracy via Machine Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20): 22303–22312.

Pokharel, G.; Farabi, S.; Fowler, P. J.; and Das, S. 2025. Street-Level AI: Are Large Language Models Ready for Real-World Judgments? (arXiv:2508.08193). ArXiv:2508.08193.

Qin, Z.; Jagerman, R.; Hui, K.; Zhuang, H.; Wu, J.; Yan, L.; Shen, J.; Liu, T.; Liu, J.; Metzler, D.; Wang, X.; and Bendersky, M. 2024. Large Language Models are Effective Text Rankers with Pairwise Ranking Prompting. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Findings of the Association for Computational Linguistics: NAACL 2024*, 1504–1518. Mexico City, Mexico: Association for Computational Linguistics.

Rathje, W. 2024. Learning When Not to Measure: Theorizing Ethical Alignment in LLMs. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 1190–1199.

Rennie, J. D.; and Srebro, N. 2005. Loss functions for preference levels: Regression with discrete ordered labels. In *Proceedings of the IJCAI multidisciplinary workshop on advances in preference handling*, volume 1, 1–6. Citeseer.

Shinn, M.; and Richard, M. K. 2022. Allocating homeless services after the withdrawal of the vulnerability index–service prioritization decision assistance tool.

Spearman, C. 1904. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1): 72.

Taylor, J. 2024. AI ban ordered after child protection worker used ChatGPT in Victorian court case. *The Guardian*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv*.