

Introduction & Motivation

LLMs are poised for rapid adoption in scarce resource-allocation settings which have two defining features:

- Agreed upon prioritization methods governed by principles of *local justice* (e.g. “vulnerability first” in homelessness services, marginal benefit in medical triage)
- Implementation by *street level bureaucrats* who can exercise significant discretion

Core Question: Can LLMs replace or assist street-level bureaucrats?

Why this matters:

- Prior work has relied on **thought experiments and stylized tasks** (e.g. trolley problems)
- Real world deployment demands testing on **actual data** from high stakes contexts

We examine this problem in the **homelessness services** setting, leveraging household records from St. Louis’ HMIS to compare LLM judgements with **both human caseworkers** and **established vulnerability metrics**.

Setting & Models

Street-Level Bureaucrats in Homelessness Services

- Allocate publicly funded interventions e.g., Emergency Shelter (ES), Transitional Housing (TH)) to at-risk households
- Decision making supported by **bureaucratic scoring tools**, e.g.:
 - ❖ **VI-SPDAT**^[1] (VI_S for single adults, VI_F for families, VI_Y for youth)
 - ❖ **RMFS Scores**^[2] (Risk/Medical Frailty Scores)

Experimental Design

Pairwise Comparison Task

- Based on **10 household pairs**^[3]
- Each LLM chooses which household should receive the **more intense intervention**
- Goal: identify if LLMs act more **vulnerability-oriented** (serve the neediest) or **outcome oriented** (maximize expected benefit), compared with lay-human decisions

Ranking Task

- Uses **2021-2024 St. Louis HMIS intake data**
- LLMs rank households by **relative vulnerability**, reconstructed from many pairwise comparisons via **Rank Centrality**^[4]
- Rankings are compared against **bureaucratic scoring tools** (VISPDAT/RMFS) to compare with established prioritization systems.

Pairwise Comparison Task	
Full Model Name	Alias
DeepSeek-R1-Distill-Llama-70B	DS-R1-L70B
Llama-3.3-70B-Instruct	LL-3.3-70B
Gemini-2.0-flash-thinking-exp	G2-FTX
Gemini-2.0-flash	G2-FLS
Gemma-3-27B-it	GM3-27B
GPT-4	GPT-4
GPT-3.5-Turbo	GPT-3.5T
GPT-4.0125-Preview	GPT-4-P125
Ranking Task	
Llama-3-8B-Instruct	LL-3-8B
DeepSeek-LLM-7B-Chat	DS-7B

Table 1: The list of LLMs that were tested in the two experiments and their aliases

Methodology

Pairwise Comparison Task

Data: Household pairs from St. Louis HMIS (demographic information + reentry risk estimates^[3]). The LLM is asked to assign one each to:

- Transitional Housing (TH):** more intensive
- Emergency Shelter (ES):** less intensive

Different degrees of information availability:

- No Prediction:** only demographic information
- Only Prediction:** only re-entry risk estimates
- Prediction First:** demographics given → LLM predicts outcome→ allocates.
- Shared Prediction:** demographic information along with risk estimates are given

	Household 1	Household 2
Number of Household Members	1	3
Number of Children	0	2
Head of Household Gender	Female	Female
Head of Household Age	45	35
Head of Household Disabling Condition	No	Yes
Head of Household Receives Substance Abuse Services	No	Yes
Monthly Income	\$400	\$1524
Number of Calls to Hotline	0	25
Prior Residence	Place not meant for habitation	Unknown
Predicted probability of needing future services within 2 years if given Transitional Housing	Low	High
Predicted probability of needing future services within 2 years if given Emergency Shelter	Medium	High

Each decision the LLM makes is classified as either **vulnerability oriented** or **outcome oriented**.

Ranking Task

Data: Household data from 2021-2024 St. Louis HMIS intake questionnaires (VISPDAT (VI_S, VI_F, VI_Y) and RMFS)

LLM Task: induce a ranking in the data-points **based on vulnerability**.

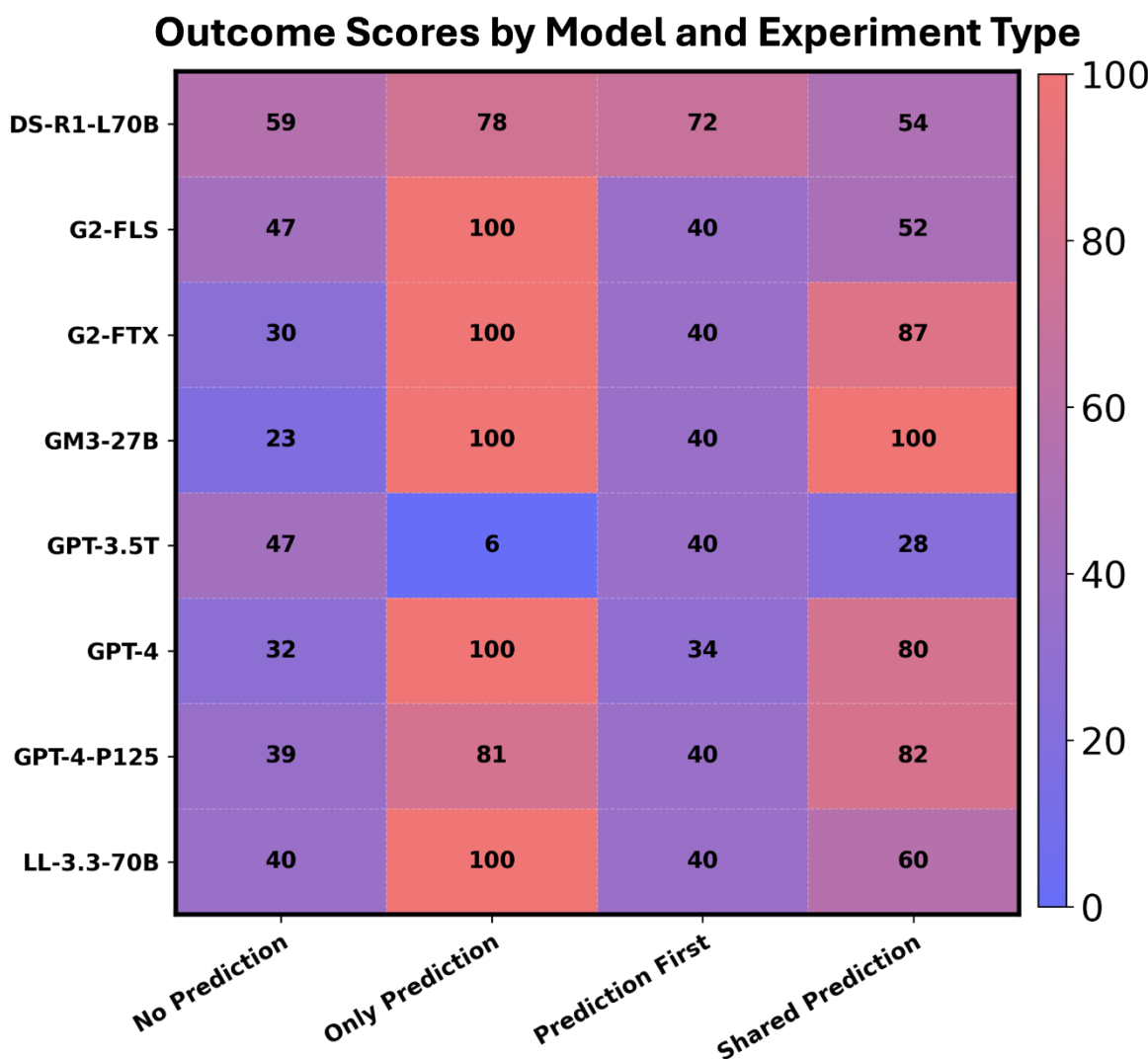
- LLMs are *notoriously unreliable rankers*.
- Infer rankings from pairwise comparisons using **Rank Centrality**^[4].
- Rankings are compared against allocations induced by:
 - Bureaucratic scoring tools** (VISPDAT, RMFS)
 - Actual human decisions** (receipt of a service)

Impractical to make $\binom{N}{2}$ comparisons to induce rankings. Sampled ~40% of pairs (well above the $\Omega(N \log N)$ threshold required).

NOTE: To ensure privacy of proprietary data, models (open-source) were used **locally** (Table 1). *Two independent runs* for each model were compared to evaluate internal consistency.

Results

Pairwise comparison Task



- Without outcome information, LLMs decision orientations are inconsistent

- Access to outcome predictions improves consistency

- Substantial heterogeneity in LLM behavior across models

- LLMs behave like non-expert humans on this task

Ranking Task

Assessment Data	Between LLaMA Rankings	Between DeepSeek Rankings
VI_S	0.24692	0.47951
VI_F	0.12805	0.27571
VI_Y	0.19748	0.02719

Table 2: Spearman’s ρ between rankings produced by two different runs of the same model

LLM judgement **diverges substantially** from bureaucratic scores.

Near 0 or negative(!) correlation

Assessment Subpopulation	Ranking Criterion	LLaMa Ranking 1	LLaMa Ranking 2	DeepSeek Ranking 1	DeepSeek Ranking 2
VI_S	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 3: Spearman’s ρ between LLM rankings and bureaucratic rankings

Assessment Data	# Positive Samples	VI-SPDAT Ranking	RMFS Ranking	LLaMa Ranking 1	LLaMa Ranking 2	DeepSeek Ranking 1	DeepSeek Ranking 2
VI_S	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	0.5692	0.51015	0.53258	0.54498	0.50854

Table 4: 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.

Do LLMs predict expert judgment?

- Bureaucrats exercise discretion with VI-SPDAT scores
- Could LLM judgements match up better with human decisions? Plausible
- Actual service allocation data used to test predictive capability (Table 4)
- Even RMFS & VI-SPDAT have low predictive power
- LLaMA rankings demonstrate some predictive capability for V_S (worse than VI-SPDAT, though)
- LLM rankings are no better than random guessing for V_Y and V_F

In short, LLM rankings show virtually no correlation with well-known scoring systems and are also not predictive of the actual decision-making of human experts.

Conclusion

- AI methods have promise for helping in the pipeline of social service provision (e.g. for measurement, prediction, and complex allocation)
- Policymakers considering the use of LLMs in social services must recognize that off-the-shelf current generation models:
 - Are inconsistent, and randomness affects their judgments
 - Are misaligned with established vulnerability metrics
 - Cannot substitute for the experience of caseworkers
- As always, adoption requires continuous calibration, transparent recommendations, and human oversight.

References:

- [1] 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.
- [2] Fefer, N. S. 2022. 2022 Consolidated Annual Performance and Evaluation Report.
- [3] 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.
- [4] Negahban, S.; Oh, S.; and Shah, D. 2017. Rank Centrality: Ranking from Pairwise Comparisons. Operations Research, 65(1): 266–287.