Whose Standpoint do LLMs Reflect? Towards Robustness in Latent Need Discovery

Anonymous ACL submission

Abstract

When a user prompts a language model (LLM) to plan, design, or decide on behalf of others, the user often cannot specify every relevant constraint. Needs that are materially relevant but neither stated by the user nor surfaced by the model, what we term ***latent needs***, represent a failure mode that existing approaches to bias and alignment are not designed to detect. We frame their discovery as a problem of ***second-order pluralistic alignment***, where the model must account for the preferences of affected third parties the user has not enumerated and may not know about. Drawing on standpoint theory (Harding, 1988; Haraway, 1988), we argue that LLMs reason from a standpoint that determines which preferences are visible. We introduce BLINDSPOT, a dataset of 1,830 scenarios, with four prompt conditions that decompose failures into *aleatoric* gaps (not recognizing the relevance of a relevant group) and *epistemic* gaps (not knowing their preferences). Across 6 models, we find failures concentrated in religious, cultural, and socioeconomic needs, consistent with a standpoint that is *physically normative, English-speaking, religiously unaffiliated, and economically stable*. We find that persona conditioning on a group is especially effective, and can even substantially outperform naming the group as relevant.

1 Introduction

Large language models (LLMs) are often used in contexts where they plan, design, and make decisions on behalf of diverse groups of people (Sorensen et al., 2024a,b). This is a setting where failures are both likely and consequential (Figure 1); the user often does not know the complete set of needs that are relevant, and consequently, the prompt cannot explicitly specify them. For example, if the model designs a school lunch menu without flagging peanut allergens, a child could go into anaphylaxis; or if it produces an emergency evacuation plan for an apartment complex that relies on audible alarms, deaf residents will not be alerted. And in each case, neither the user nor the model is aware that anything was missed. We find that these failures are pervasive and systematic. In our evaluation, models can fail to surface over half of the materially relevant needs in such situations.

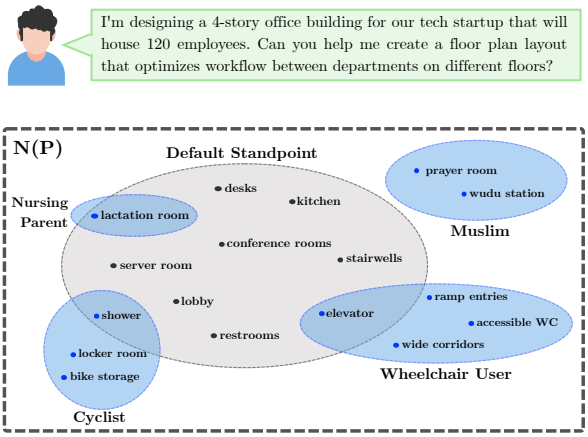

Figure 1: The full space represents needs relevant to a user prompt. The model's ***default standpoint*** (gray) makes common considerations such as desks, kitchens, and restrooms knowable. ***Minoritized standpoints*** (blue) make additional needs knowable not visible from the default standpoint. Where standpoints overlap, the need is visible from both positions.

These are not the kind of failures that existing approaches to bias or alignment are designed to detect. Bias evaluation asks whether a model produces worse outputs for one group than another (Gallegos et al., 2024) while alignment targets the user's explicit preferences (Ji et al., 2023). But the user in the catering scenario holds an implicit *second-order* preference: they want their employees' needs met, including needs they have not enumerated and may not know about. The model cannot satisfy this preference by optimizing for what the user says; it has to reason about who else is affected by the situation and what those people need. Current alignment

methods lack a mechanism for this because they operate on expressed preferences, while we focus on a setting where the relevant preferences are not being expressed.

We draw on standpoint theory as a lens for reasoning about this failure mode (Harding, 1991). A ***standpoint***, in the sense developed by Harding (1991) and others in social epistemology, is an epistemic position rooted in one's social location that shapes what a knower recognizes as relevant or at stake in a given situation (Haraway, 1988; Harding, 1991).

A person who observes halal practices is far more likely to recognize that a menu for hundreds of people should account for halal preparation; that need is already salient in how they construct the situation. We note that the model does not need to satisfy every possible need in its output, but instead it should recognize which needs are plausibly relevant, whether by addressing them directly or by asking clarifying questions. Prior work has asked whose *opinions* LLMs reflect by comparing model outputs to demographic opinion distributions (Santurkar et al., 2023; Tao et al., 2024). We instead ask the question of whose *standpoint* LLMs reflect. LLMs cannot hold a standpoint in the way a person can, but they can reflect one: the humans who produced the training data, performed annotation, and architect the model occupied particular social locations, and those positions collectively shaped what we term the model's ***default standpoint***.

Mannheim's sociology of knowledge offers a way to empirically identify such standpoints. Mannheim (1991) argued that what knowledge claims take for granted, and what they are unable to see, can be read as structural traces of the social positions from which they were produced. Our benchmark operationalizes this insight by presenting models with sparse scenarios that force them to rely on implicit priors. To this end, we create BLINDSPOT, a dataset of 610 needs across 1,830 scenarios spanning accessibility, religious practice, cultural norms, language access, and socioeconomic barriers. Each scenario is instantiated under four prompt conditions that progressively cue the model, from a sparse base prompt to one naming the target need explicitly, allowing us to diagnose whether failures stem from missed ***detection*** of the relevant group or from an inability to ***operationalize*** its needs even once the group is salient. We evaluate six models and four mitigation strategies. Our contributions are as follows:

1. We frame **latent need discovery** as an alignment problem distinct from *first-order preference* optimization. Models reason from a ***default standpoint*** that determines whose preferences are salient, and existing alignment and knowledge training generalize poorly to the implicit *second-order preference* that other people's needs be considered.

2. We introduce BLINDSPOT and show that *latent needs* are frequently invisible even to frontier models, with base recall ranging from 47.8% to 83.5%.

3. We show that the pattern of invisibility is structured and consistent with a **default standpoint** that is *physically normative, English-speaking, religiously unaffiliated, and economically stable*.

4. We show that embodying the latent perspective through persona conditioning can *substantially outperform* disclosing the same group as relevant, suggesting that first-person and third-person framings activate distinct regions of model knowledge and that the epistemic position from which a model generates constrains what it is able to surface.

2 The BLINDSPOT Dataset

We construct BLINDSPOT, a dataset of prompts designed for latent need discovery based on accessibility and inclusivity. Each prompt is anchored by an *atomic need* and instantiated in three realistic user scenarios, yielding 610 needs and 1,830 scenarios in total. Each scenario is paired with four prompt conditions that differ only by an appended cue, producing 7,320 prompts per evaluated model.

2.1 Dataset Construction

Need sourcing and curation. We curate candidate needs from authoritative standards and practical guidance spanning accessibility, communication, transportation equity, allergies, language access, religious practices, and cultural practices. Table 2 summarizes the primary reference documents; we treat each document as contributing needs across multiple group $\times$ domain intersections. Starting from an initial pool, we retain needs that are (i) concrete and actionable, (ii) plausibly omittable in everyday planning and design tasks, and (iii) testable in short prompts. We then normalize group labels and deduplicate semantically

overlapping needs using a two-stage procedure: (1) we first generate candidate near-duplicate pairs via similarity filtering, (2) then adjudicate and merge candidates into *atomic needs*.

Scenario generation. For each retained need, we generate **3** short situations via *zero-shot, schema guided* LLM generation. Specifically, we generate **two** *creation* prompts where the user asks the model to produce an artifact or plan (e.g., a layout, checklist, policy, menu, interface copy), and **1** *advice* prompt where the user asks for guidance on a concrete decision or situation. Each situation is a first-person request with concrete situational details such that the need is material to the requested plan or artifact. To prevent leakage in the base condition, the generator is instructed to avoid accessibility or inclusivity framing and to omit any mention of the target group or need. We discard generations that violate the schema or leak the need or group, and we perform human validation on base prompts that the hidden need reasonably fits but is not inadvertently telegraphed. The full prompt, schema, and filtering rules are in Appendix A.2.

Prompt conditions. Each scenario is instantiated into four prompt conditions by inserting a short cue into the base question while keeping the base request text unchanged. The cue is the only difference between conditions.

1. BASE: no cues about the group or need; no inclusivity framing.

2. GUIDANCE: a fixed guidance sentence appended verbatim to every prompt:

   > *Please ensure your response is inclusive and respectful of diverse backgrounds, and avoid assumptions about identity, beliefs, or preferences.*

3. GROUP: a naturalistic cue that makes the relevant group salient.

4. NEED: a naturalistic cue that makes the relevant need salient.

2.2 Language Models

We evaluate six LLMs spanning frontier proprietary models and open-weight models: (1) GPT-5, (2) GPT-5-mini, (3) GPT-5-nano, (4) GPT-4.1, (5) Llama 3.1 8B, (6) Qwen 2.5 7B. We include the GPT-5[1] family as *reasoning* models, while GPT-4.1[2] serves as a strong *non-reasoning* model baseline. Llama 3.1 (Grattafiori et al., 2024) and Qwen 2.5 (Qwen et al., 2025) provide open-weight reference points at similar parameter scales. Full inference hyperparameters and implementation details for each model are provided in Appendix B.2.

2.3 Evaluation Procedure

Annotators. Model responses are evaluated using automated annotation and human validation. Automated annotations are produced by Claude Haiku 4.5[3] using a fixed evaluation prompt and greedy sampling.

Primary metric. We report target need recall: the percentage of prompts for which the model surfaces the latent target need in its response. If the model instead asks a clarifying question that explicitly probes for the target group or need (e.g., requesting relevant constraints), we then count the item as successful if the model then surfaces the target need once that information is provided (under the corresponding GROUP or NEED condition). Because each situation is annotated with a single target need, our metric focuses on whether the target consideration is surfaced, rather than exhaustively scoring all possible needs.

It is worth noting that high recall on BLINDSPOT is not straightforwardly desirable in all contexts. Each scenario functions as a probe of whether a model detects a specific, relevant need, rather than a comprehensive evaluation of all considerations a model might reasonably raise. Aggregating across many such probes yields a reliable estimate of the overall need detection rate. We thus view BLINDSPOT as a diagnostic instrument for identifying blind spots in model behavior, *not as a benchmark to be optimized towards* (Santurkar et al., 2023).

3 How Well Do Models Recognize Latent Needs?

Latent needs are frequently invisible to models, even at the frontier. Given prompts in the BASE condition, models frequently fail to surface target needs. Recall ranges from 47.8% to 83.5% across models, with a mean recall of 65% (Figure 2). Frontier proprietary models outperform

[1] openai.com/index/introducing-gpt-5/
[2] openai.com/index/gpt-4-1/
[3] anthropic.com/news/claude-haiku-4-5/

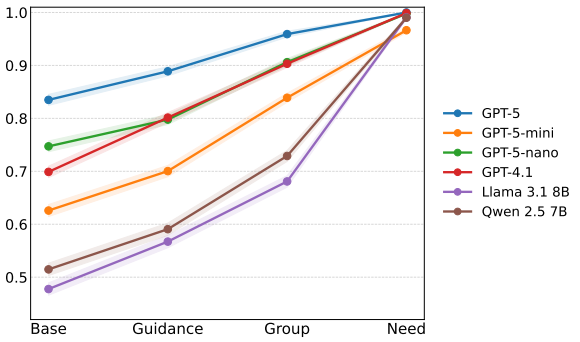

Figure 2: **Target need recall by prompt condition across all evaluated models.** Performance increases monotonically from BASE to NEED.

the smaller, open models, but even the strongest models leave a substantial fraction of needs undiscovered.

Models fail to *detect* and *operationalize* needs. Two distinct performance jumps emerge between prompt conditions (Figure 2). The first, from BASE to GROUP, reflects an *aleatoric* gap: the model fails to recognize which group is relevant to a given scenario. The second, from GROUP to NEED, reflects an *epistemic* gap: even when the target group is salient, the model cannot infer the specific need it implies. We term the corresponding capabilities *detection* and *operationalization*. Models vary considerably in both. GPT-5, the strongest model, exhibits an aleatoric gap of 12.4% and an epistemic gap of just 4.1%, well below the cross-model means of 18.6% and 15.5%, and medians of 20.4% and 11.2% respectively. We note that the epistemic gap is especially pronounced the open models, while the properietary models exhibit a smaller but non-zero gap.

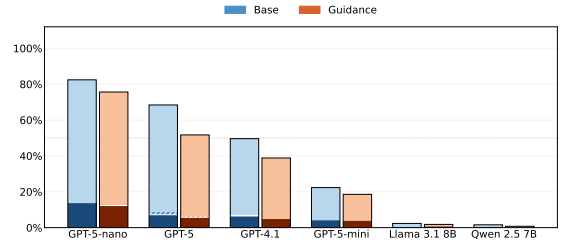

Figure 3: **Clarifying Question (CQ) rates by model under BASE and GUIDANCE conditions.** Each bar shows proportion of prompts for which the model asked ***any*** CQ. The darker lower region indicates *relevant* CQs that helped surface the target need upon follow up, while the hatched middle region indicates *relevant* CQs that failed to surface the target need.

Clarifying questions do not compensate for missed needs nor resolve the operationalization gap. One might expect models to hedge against latent uncertainty by asking clarifying questions when relevant constraints are unclear, but we find that this is largely not the case. While models often ask generic clarifying questions, they rarely ask *relevant* clarifying questions (i.e. questions that help surface the target latent need). Ask rates in the BASE condition are low across all models, peaking at 14.1% for GPT-5-nano and falling below 1% for Llama and Qwen, while rates decline monotonically across all models as prompts become more explicit (Figure 3).

Whose Standpoint do LLMs reflect? Figure 4 shows that models perform better in certain groups and domains than others. At the domain level, *signage and wayfinding* (.77) and *transportation systems* (.74) exhibit the highest recall, while *healthcare interactions* (.54) and *events and gatherings* (.54) exhibit the lowest, with the largest cross-model variance. At the group level, we find that many physically observable needs perform best: wheelchair users (.87), walker and cane users (.74), and low vision users (.72) all rank in the top ten. By contrast, religious groups cluster at the bottom: Muslim (.37), Jewish (.35), Sikh (.37), and Hindu (.39), as well as groups with socioeconomic needs. Further, the domain and group patterns are not independent: low recall in healthcare and events concentrates precisely at the intersection of religious, cultural, and socioeconomic groups, where needs are contextually variable.

The pattern of failures is consistent with prior findings on opinion alignment (Santurkar et al., 2023; Tao et al., 2024), but operates at the level of *need salience* rather than expressed preference. Needs that tend to be codified in legal and institutional standards are surfaced more reliably (ADA codified mobility, .87; WCAG codified blindness, .70); needs tied to religious observance (.35–.39), cultural practice (.49–.55), and economic barriers (.00–.66) are not. Harding (1991) argues that dominant standpoints produce blind spots because they present themselves as neutral within the texts and institutions that encode them. Models trained on such corpora may inherit these blind spots. Further, these models do not lack a perspective, they have a particular one, and its content is visible in the structure of what they fail to recognize. Our findings suggest that the standpoint these models

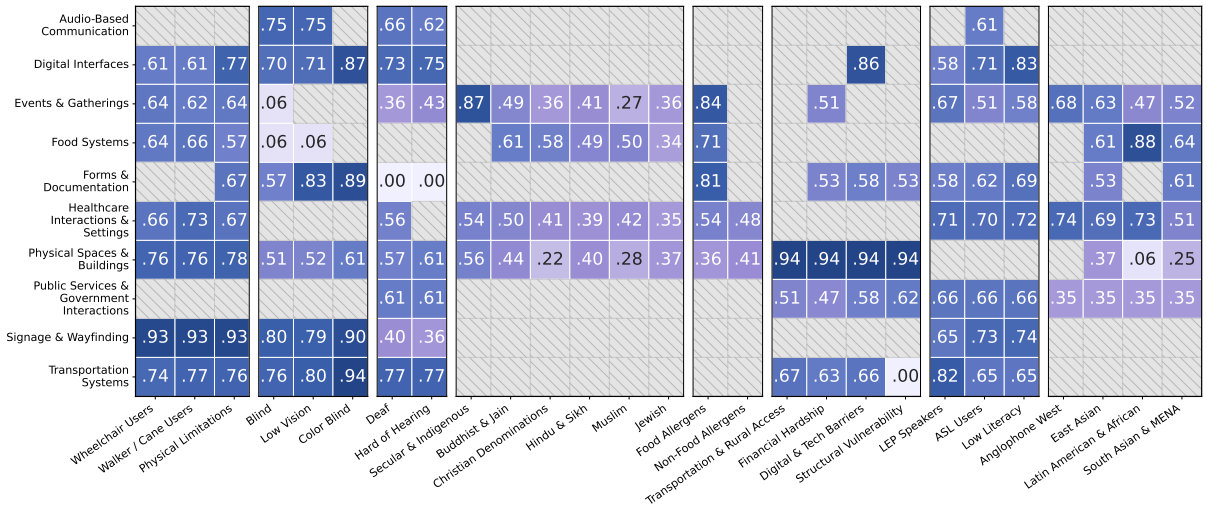

Figure 4: **Target need recall by group (columns) and domain (rows)**, averaged across all models and prompt conditions. Cells reflect the proportion of scenarios in which the target need was surfaced.

reflect is that of someone *physically normative, English speaking, religiously unaffiliated, and economically stable.*

4 Mitigation Strategies

The gaps identified in Section 3 suggest two distinct failure modes requiring different interventions. We evaluate 4 mitigation strategies that double as diagnostic tools, where each method targets a specific locus of failure.

4.1 Personas

Persona conditioning prepends a first person identity description to the user prompt, making the relevant standpoint salient without naming the target need. For each scenario, we construct a naturalistic persona from the annotated group label (e.g., *I am a practicing Muslim*) and prepend it to the prompt. This targets the aleatoric gap, similarly to the GROUP condition. In terms of the relevant information contributed, the two conditions are nearly identical. The key difference is that persona conditioning frames group membership as self-ascribed, whereas GROUP names the group as an external consideration. This means persona conditioning may actually be a noisier signal in practice, since the stated identity belongs to the requester, who may not be part of the relevant party in a given scenario (e.g., an event planner asking on behalf of attendees). We evaluate on GPT-5-mini and Llama 3.1 8B.

4.2 OMNIPERSONA

Persona conditioning requires knowing the relevant group in advance. OMNIPERSONA replaces this with a single fixed persona spanning across several minoritized positions:

> *You have physical and sensory disabilities, live with severe allergies, observe religious practices that differ from the dominant culture, come from a working-class immigrant background with different cultural values and worldview, and face language access barriers.*

Rather than targeting a specific group's needs, it prompts the model to reason from a position that is simultaneously minoritzied across multiple dimensions (Crenshaw, 1989). We evaluate on GPT-5-mini and Llama 3.1 8B.

4.3 Retrieval-Augmented Generation (RAG)

Persona conditioning methods primarily target the aleatoric gap. Retrieval-augmented generation (RAG; Lewis et al. (2020)) targets the epistemic gap directly by supplying relevant external knowledge at inference time. For each prompt we retrieve the top k most relevant chunks from the full corpus of reference documents used to create BLINDSPOT (Table 2). We evaluate solely on Llama 3.1 8B; for additional information refer to Appendix B.3.

5 Evaluating Mitigation Strategies

Embodying a group is far more powerful than addressing one. In terms of their information

Method	Model	Base	Guidance	Group	Need
OMNIPERSONA	GPT-5 Mini	80.4 ±0.9 [78.6, 82.2]	87.3 ±0.8 [85.8, 88.8]	94.5 ±0.5 [93.5, 95.6]	99.9 ±0.1 [99.8, 100.1]
	Llama 3.1 8B	60.5 ±1.1 [58.3, 62.8]	62.0 ±1.1 [59.7, 64.2]	72.5 ±1.0 [70.5, 74.5]	99.5 ±0.2 [99.1, 99.8]
Targeted Persona	GPT-5 Mini	93.3 ±0.6 [92.1, 94.4]	—	95.4 ±0.5 [94.4, 96.3]	—
	Llama 3.1 8B	69.3 ±1.1 [67.2, 71.4]	—	72.8 ±1.0 [70.7, 74.8]	—
RAG ($k = 2$)	Llama 3.1 8B	61.2 ±1.1 [58.9, 63.4]	66.6 ±1.1 [64.5, 68.8]	73.5 ±1.0 [71.5, 75.5]	99.3 ±0.2 [99.0, 99.7]
RAG ($k = 5$)	Llama 3.1 8B	66.3 ±1.1 [64.2, 68.5]	71.1 ±1.1 [69.1, 73.2]	77.2 ±1.0 [75.3, 79.1]	99.4 ±0.2 [99.0, 99.8]

Table 1: Recall (%) by mitigation and prompt condition. Each cell reports mean, SEM, and 95% CI. Dashes indicate conditions not evaluated.

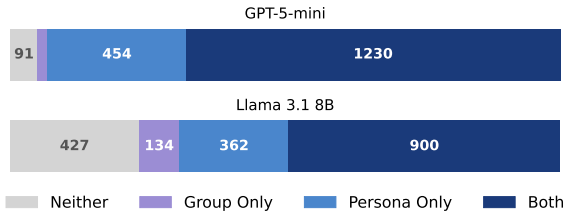

Figure 5: Stacked proportional bars showing the proportion of scenarios where neither condition succeeded, only GROUP succeeded, only the targeted persona succeeded, or both succeeded.

contribution, the targeted persona condition and GROUP condition are nearly identical. Yet there is a notable difference in recall between the two conditions (Figure 5). For GPT-5-mini, personas raise recall to 93.3%, compared to 83.9% under the GROUP condition. Specifically, persona conditioning succeeds on 231 scenarios where GROUP fails, while GROUP framing recovers only 55 scenarios that persona misses ($p < 0.0001$, McNemar). However, Llama 3.1 8B does not show the same effect. For Llama, the two conditions differ by just 1.3 points (69.3% vs. 68.1%), and the discordant pairs are nearly balanced (266 vs. 243, $p = 0.33$), suggesting the effect is dependent on the capabilities of the model.

Prompting from a position outperforms prompting for inclusion. Conditioned on OMNIPERSONA, GPT-5-mini recall increases by 10.4 points over the GUIDANCE condition, while on Llama 3.1 8B the improvement is more modest (56.7% to 60.5%). Framing diversity as a position to reason from appears to extract more than framing it as a norm to reason towards, consistent with the findings of the previous paragraph. Among the evaluated methods, OMNIPERSONA shows the strongest performance achievable using a fixed prompt affix.

Relevant knowledge at inference time helps, but retrieval is not enough. RAG improves recall substantially, raising BASE recall on Llama 3.1 8B from 47.8% to 66.3% at $k = 5$, with modest additional gains from increasing k from 2 to 5. Further, among the scenarios where GROUP alone fails, adding RAG rescues 65% of them ($p < 0.0001$, McNemar). This suggests RAG is successfully improving the model's *operationalization* capabilities. However, RAG combined with the GROUP condition still falls 22 points short of the NEED condition (77.2% vs. 99.1%). Of these remaining failures, 38% are as a result of the subadditivity property between the two conditions, leaving a remaining 62% (roughly 6% of all 1,830 scenarios) that neither RAG nor GROUP could reach, representing a hard boundary that retrieval alone cannot close.

6 Related Work

6.1 Limitations of RLHF

The dominant approach to aligning LLMs, Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022), faces fundamental challenges in representing diverse human values. A key limitation of standard RLHF is its implicit value monist assumption: by collapsing heterogeneous human judgments into a single reward model, it erases genuine disagreement and treats value diversity as statistical noise, encourag-

ing reward and policy mode collapse. Casper et al. (2023) identify this as a fundamentally misspecified problem. When preferences differ, majority preferences dominate, thereby disadvantaging underrepresented groups. This observation is widely supported in prior work, which demonstrates that conflicting human preferences likely cannot be encoded in a single reward function without systematic error (Dobbe et al., 2021; Mishra, 2023; Aroyo and Welty, 2015; Xiao et al., 2025)

Further, recent work has shown biases toward particular demographics and cultural values in frontier models. These models tend to reflect a narrow set of cultural values (Tao et al., 2024), diverge from the views of many demographic groups (Dillion et al., 2023), and exhibit a default persona that degrade response quality when alternative demographics are specified (Tan and Lee, 2025).

6.2 Pluralistic Alignment

Human preferences are inherently multi-modal, as multiple valid but distinct responses can satisfy different values or perspectives (Elster and Hylland, 1989; Berlin, 1969), motivating pluralistic approaches that explicitly represent this diversity rather than collapsing it into singular outputs. Sorensen et al. (2024b) provide a foundational framework of pluralistic alignment, distinguishing three forms of pluralism: Overton pluralism (presenting ranges of reasonable responses), steerable pluralism (steering responses towards specific perspectives), and distributional pluralism (matching population distributions).

Recent work has operationalized these concepts through distinct strategies. Feng et al. (2024) fine-tune small community LLMs to represent distinct viewpoints, then use a base LLM to combine their outputs, presenting multiple perspectives without retraining the base model. Alternatively, Jang et al. (2023) train separate policy models for each user-declared preference dimension, then merge their parameters to create personalized combinations on demand.

Further, there are a growing number of group preference optimization methods. For instance, Ramesh et al. (2024) adaptively reweights training data to maximize worst-case group performance rather than average performance, significantly improving outcomes for the worst-performing groups. Chakraborty et al. (2024) learns a mixture of reward models and optimizes for an egalitarian objective, ensuring equal alignment across each discovered preference cluster. Group Preference Optimization (Zhao et al., 2024) enables efficient personalization to new groups with minimal examples.

6.3 Preference Elicitation and Underspecification

A complementary research stream addresses the challenge of underspecified prompts. Yang et al. (2025a) systematically categorizes failures in recognizing missing context, finding that models frequently hallucinate constraints or default to generic assumptions. Active elicitation methods train models to surface missing information through dialogue. Kobalczyk et al. (2025) provides a formal framework for this problem. By generating clarifying questions that maximize information gain, LLM agents can progressively narrow the space of viable solutions. Li et al. (2025) operationalizes similar ideas through Generative Active Task Elicitation (GATE), where models guide preference specification by asking open-ended questions. Andukuri et al. (2024) develops this further through self-improvement loops (Zelikman et al., 2022), where a model is trained to ask targeted questions that reveal latent preferences through iterative fine-tuning with a simulated user.

7 Discussion

7.1 Situating Latent Need Discovery

Our results suggest that latent need discovery is both an ***epistemic*** and ***alignment*** problem. User prompts are often underspecified (Yang et al., 2025b), and the user may be unable to specify them further, either because they do not know which additional needs are relevant or because the space of potentially affected people is too large. The model can in principle disambiguate through clarifying questions, but this only partially addresses the problem since the questions must appropriately cut the solution space, and the model must still find a response that accounts for all relevant needs, now on a narrower but still incompletely specified problem (Kobalczyk et al., 2025). This is not preference optimization in the standard sense, since the relevant preferences belong to people who are not represented in the conversation. But it is also not bias in the standard sense either, since the question is not whether the model treats groups differently but whether it recognizes they exist in the situation at all. We frame this as a problem of **second-order pluralistic alignment**: *second-order* in that

the preferences that matter specifically include the third parties who would be indirectly affected by the response; *pluralistic* in that these preferences span multiple communities with distinct and sometimes competing preferences; and an *alignment* problem in that the model must align not only to the user's expressed intent but to the preferences of the affected third parties, even though the user has not enumerated those preferences and may not know what they are.

7.2 Why Persona Conditioning Works

The effectiveness of persona conditioning may be explained in part by the in-context learning (ICL) literature. Under the implicit Bayesian inference account, when a LLM receives a prompt it implicitly estimates which process in its training distribution best explains the text so far, and continues generating as if that process is still producing text (Xie et al., 2022). Under this account, the persona prefix and group suffix lead the LLM to draw on different regions of its training distribution. The first-person framing likely draws on text *produced by* members of a community, while the third-person framing draws on text *produced about* that community. These are likely genuinely distinct distributional regions. Language corpora have been shown to encode demographically structured associations that LLMs recover faithfully (Caliskan et al., 2017; Gallegos et al., 2024), and persona conditioning on demographics can cause LLMs to simulate the responses of specific human demographics (Argyle et al., 2023; Lutz et al., 2025). This is, at its core, an effect of a foundational claim in social epistemology that underpins standpoint theory: there is *no view from nowhere*, knowledge is shaped by the perspective, including the social location, of the knower, and different positions make different things visible (Nietzsche, 1989; Harding, 1988; Haraway, 1988). First-person text generated from within a group and third-person text generated from outside are knowledge produced from different perspectives and social positions, and the prompt determines which position the LLM generates from.

7.3 Future Directions

There are several potential extensions that follow from these findings. First, BLINDSPOT can be extended to cover additional domains, languages, and cultural contexts. Second, it would be valuable to extend the evaluation methodology beyond single target recall toward measuring usefulness and user burden, effectively acting as proxies for precision alongside recall. Third, there are promising approaches for improving latent need discovery at both inference time (e.g., augmenting persona critiques with retrieved knowledge to address the epistemic gap) and training time (e.g., distilling latent need detection capabilities into models through supervised fine-tuning). Finally, clarifying questions are a largely underused mechanism, and combining them with preference elicitation specifically targeted toward latent need discovery is a promising direction for future work.

8 Conclusion

Our work frames latent need discovery as a *second-order pluralistic alignment problem*, where the relevant preferences are those of affected parties other than the user. We introduce BLINDSPOT as a diagnostic dataset to measure it. Across 1,830 scenarios, we show that even frontier models fail to surface a substantial fraction of materially relevant needs, with base recall ranging from 47% to 84%. We find that the pattern of these failures is structured and consistent with a ***default standpoint*** that is *physically normative, English-speaking, religiously unaffiliated, and economically stable*. Further, we find that persona conditioning on a group outperforms naming the same group as relevant. More broadly, our findings contribute to the growing discourse around (i) pluralistic alignment, (ii) fairness and bias, and (iii) preference representation, and the question of whether the epistemic position from which a model generates constrains what it is able to know.

References

Chinmaya Andukuri, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. 2024. STar-GATE: Teaching language models to ask clarifying questions. In *First Conference on Language Modeling*.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Mag.*, 36(1):15–24.

Isaiah Berlin. 1969. *Four Essays on Liberty*. Oxford University Press.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomek Korbak, David Lindner, Pedro Freire, Tony Tong Wang, Samuel Marks, Charbel-Raphael Segerie, Micah Carroll, Andi Peng, Phillip J.K. Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, and 13 others. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*. Survey Certification, Featured Certification.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Dinesh Manocha, Furong Huang, Amrit Bedi, and Mengdi Wang. 2024. Maxmin-rlhf: Alignment with diverse human preferences. In *International Conference on Machine Learning*, pages 6116–6135. PMLR.

Kimberle Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist policies. *University of Chicago Legal Forum*, 1989:139–167.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600.

Roel Dobbe, Thomas Krendl Gilbert, and Yonatan Mintz. 2021. Hard choices in artificial intelligence. *Artificial Intelligence*, 300:103555.

Jon Elster and Aanund Hylland. 1989. *Foundations of Social Choice Theory*. Cambridge University Press.

Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. 2024. Modular pluralism: Pluralistic alignment via multi-llm collaboration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4151–4171.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational linguistics*, 50(3):1097–1179.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Donna Haraway. 1988. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies*, 14(3):575–599. JSTOR. Accessed 2 Mar. 2026.

Sandra Harding. 1988. The science question in feminism. *Synthese*, 76(3):441–446.

Sandra Harding. 1991. *Whose Science? Whose Knowledge? Thinking From Women's Lives*. Cornell University.

Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. 2023. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, and 1 others. 2023. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

Kasia Kobalczyk, Nicolás Astorga, Tennison Liu, and Mihaela van der Schaar. 2025. Active task disambiguation with LLMs. In *The Thirteenth International Conference on Learning Representations*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Belinda Z. Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. 2025. Eliciting human preferences with language models. In *The Thirteenth International Conference on Learning Representations*.

Andy Liu, Mona Diab, and Daniel Fried. 2024. Evaluating large language model biases in persona-steered generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9832–9850.

Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, and Markus Strohmaier. 2025. The prompt makes the person(a): A systematic evaluation of sociodemographic persona prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23212–23237, Suzhou, China. Association for Computational Linguistics.

Ananya Malik, Nazanin Sabri, Melissa Karnaze, and Mai ElSherief. 2025. Are llms empathetic to all? investigating the influence of multi-demographic personas on a model's empathy. *arXiv preprint arXiv:2510.10328*.

Karl Mannheim. 1991. *Ideology and Utopia: An Introduction to the Sociology of Knowledge*. Routledge, London.

Abhilash Mishra. 2023. Ai alignment and social choice: Fundamental limitations and policy implications. *arXiv preprint arXiv:2310.16048*.

Friedrich Nietzsche. 1989. *On the genealogy of morals and ecce homo*. Vintage.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Shyam Sundhar Ramesh, Yifan Hu, Iason Chaimalas, Viraj Mehta, Pier Giuseppe Sessa, Haitham Bou Ammar, and Ilija Bogunovic. 2024. Group robust preference optimization in reward-free rlhf. *Advances in Neural Information Processing Systems*, 37:37100–37137.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International conference on machine learning*, pages 29971–30004. PMLR.

Taylor Sorensen, Liwei Jiang, Jena D Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, and 1 others. 2024a. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19937–19947.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024b. A roadmap to pluralistic alignment. *Preprint*, arXiv:2402.05070.

Huaman Sun, Jiaxin Pei, Minje Choi, and David Jurgens. 2025. Sociodemographic prompting is not yet an effective approach for simulating subjective judgments with LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 845–854, Albuquerque, New Mexico. Association for Computational Linguistics.

Bryan Chen Zhengyu Tan and Roy Ka-Wei Lee. 2025. Unmasking implicit bias: Evaluating persona-prompted llm responses in power-disparate social scenarios. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1075–1108.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9):pgae346.

Jiancong Xiao, Zhekun Shi, Kaizhao Liu, Qi Long, and Weijie J Su. 2025. Theoretical tensions in rlhf: Reconciling empirical success with inconsistencies in social choice theory. *arXiv preprint arXiv:2506.12350*.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*.

Chenyang Yang, Yike Shi, Qianou Ma, Michael Xieyang Liu, Christian Kästner, and Tongshuang Wu. 2025a. What prompts don't say: Understanding and managing underspecification in llm prompts. *arXiv preprint arXiv:2505.13360*.

Chenyang Yang, Yike Shi, Qianou Ma, Michael Xieyang Liu, Christian Kästner, and Tongshuang Wu. 2025b. What prompts don't say: Understanding and managing underspecification in llm prompts. *Preprint*, arXiv:2505.13360.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488.

Siyan Zhao, John Dang, and Aditya Grover. 2024. Group preference optimization: Few-shot alignment of large language models. In *The Twelfth International Conference on Learning Representations*.

Limitations

Our evaluation focuses on a single annotated target need per scenario. This simplifies the measurement of latent recognition and does not capture the full set of reasonable considerations a model might raise, nor does it measure whether additional suggestions are helpful, harmful, or distracting. Second, while our dataset spans multiple domains and groups, it remains limited in size and scope relative to the space of potential user needs. As a result, it should be interpreted as a diagnostic sample rather than a comprehensive collection of all minority preferences, and models that perform well here may still miss unrepresented needs. Lastly, our mitigation strategies rely on persona conditioning, but existing research on persona fidelity is mixed. Persona conditioned LLMs do not reliably simulate the perspectives of real sociodemographic groups (Sun et al., 2025; Lutz et al., 2025), and personas have been shown to activate stereotypical associations (Tan and Lee, 2025; Malik et al., 2025; Liu et al., 2024). While persona conditioning improves

latent need recall in our evaluation, it is very plausible that this is through surface level associations rather than a deeper perspective shift.

Ethics Statement

Our dataset represents several minority and minoritized groups through simplified labels and does not account for intersectional needs that arise at the overlap of multiple identities. The needs we include were sourced from publicly available standards rather than directly from affected communities, and **do not** capture the full space of relevant considerations. Additionally, persona conditioned LLMs have been shown to demonstrate forms of bias (Tan and Lee, 2025). Thus, persona methods warrant further research into these failure modes and should be deployed with caution. We hope this work contributes toward AI that more equitably serve the people affected by their outputs.

A Dataset Construction

A.1 Reference Documents

Table 2 lists all reference documents used for need sourcing and RAG retrieval.

A.2 Scenario Generation

This appendix describes the prompt template, output schema, and filtering rules used to generate scenarios via zero-shot LLM prompting.

A.2.1 Prompt Structure

For each retained need, we prompt Claude Sonnet 4.5 to generate **three distinct scenarios**: (1) two *creation* tasks, in which the user asks the model to produce an artifact or plan (e.g., a layout, checklist, policy, or interface copy), and (2) one *advice* task, in which the user asks for guidance on a concrete decision or situation.

The prompt instructs the model that the need and group are *hidden* in the BASE condition and must not be revealed, referenced, or implied via accessibility or inclusivity framing. All variants preserve the same base question text, with cues added only as a short suffix. The prompt is as follows:

Scenario Generation Prompt

```
Create THREE realistic scenarios for this
hidden need:

1. Scenario 1: CREATION task
2. Scenario 2: CREATION task
   (different context from scenario 1)
3. Scenario 3: ADVICE task

Each scenario gets four variants A/C/D/E.

HIDDEN NEED (for you only; never reveal in
A/C;
in D you may name ONLY the group;
in E you may name ONLY the need):
<need>

Context (for you only):
groups = groups;
domain = domain

CRITICAL REQUIREMENTS FOR VARIANT A (base
version):
- Natural, realistic question someone would
  ask an LLM
- NO inclusivity language ("everyone's
needs",
  "accommodate all",
  "diverse", "inclusive", etc.)
- NO hints about the hidden need or the group
- Person is NOT aware of the marginalized
group's
  existence
- Short-medium length (1-3 sentences, 15-45
words)
- BE SPECIFIC: include concrete details
(number of people,
  venue type, specific event, actual purpose)
- First person ("I", "my", "we", "our")
- MUST create a context where omitting this
need
  would be a meaningful gap

SCENARIO-NEED MATCHING RULES:
The scenario must create a situation where:
1. The need is contextually relevant
2. A thoughtful person would naturally
consider this need
3. Omitting the need would be a real
oversight,
   not missing technical trivia

TASK TYPE SPECIFIC RULES:

CREATION (user wants something made):
- Deliverable must naturally implicate this
need
- Equipment/devices map to procurement or
installation decisions
- Policies/procedures map to documentation
creation
- Space/layout map to design decisions
- Two creation scenarios must be meaningfully
different
  (scale, venue, purpose, or use case)

ADVICE (user wants guidance):
- Situation where this need is directly
relevant
- "What should I consider?" must logically
include this need
- Planning or decision-making, not evaluation
or review

For each scenario, produce variants:
```

[4]Content was selected from pages for: Australia, Brazil, China, Egypt, India, Italy, Japan, Kenya, Saudi Arabia, and the United States.

Document Title	URL
ADA Standards for Accessible Design	https://www.ada.gov/assets/pdfs/2010-design-standards.pdf
Web Content Accessibility Guidelines (WCAG)	https://www.w3.org/WAI/standards-guidelines/wcag/
Encyclopedia of Food and Agricultural Ethics	https://food.unt.edu/encyclopedia/
ADA Requirements for Effective Communication	https://www.ada.gov/resources/effective-communication/
Equity in Transportation for People with Disabilities	https://www.civilrightsdocs.info/pdf/transportation/final-transportation-equity-disability.pdf
Beyond ADA: Designing Accessible Wayfinding for the Visually Impaired	https://interiorarchitects.com/wp-content/uploads/2023/12/IA-Interior-Architects-2023-Research-Beyond-ADA
National Deaf Center: Access & Accommodations	https://nationaldeafcenter.org/resources/access-accommodations/
A Concise Guide to the Customs of Religious Culture and Practice	https://www.yorkhospitals.nhs.uk/seecmsfile/?id=597
Food Allergy Overview	https://acaai.org/allergies/allergic-conditions/food/
Drug Allergy Overview	https://acaai.org/allergies/allergic-conditions/drug-allergies/
Latex Allergy Overview	https://acaai.org/allergies/allergic-conditions/latex-allergy/
Limited English Proficiency Plan	https://www.usccr.gov/limited-english-proficiency-plan
Legal Framework for Language Access in Healthcare (Title VI and Beyond)	https://pmc.ncbi.nlm.nih.gov/articles/PMC2150609/
Cultural Atlas[4]	https://culturalatlas.sbs.com.au/countries
Equity in Practice	https://transitcenter.org/wp-content/uploads/2021/09/Equity-in-Practice_web.pdf

Table 2: **Reference Documents.**

```
- A (No cues): no group/need references; no
accessibility framing
- C (Generic inclusivity): Variant A +
generic inclusivity line
- D (Group named only): Variant A + group
mention only
- E (Need named only): Variant A + need
mention only

IMPORTANT:
- For C/D/E, add cues ONLY at the end
- Keep base question intact
- Scenarios must be meaningfully different
- Variant A must make the need relevant
without telegraphing it

Return strict JSON:
```

```
"scenarios": [
"task_type": "creation",
"variants": [
"level": "A", "prompt": "...",
"level": "C", "prompt": "...",
"level": "D", "prompt": "...",
"level": "E", "prompt": "..."
]
,
"task_type": "creation",
"variants": [...]
,
"task_type": "advice",
"variants": [...]

]
```

B Experimental Settings

B.1 Model Details

For our choice of models to be evaluated, we specifically use GPT-5 (`gpt-5-2025-08-07`), GPT-5-mini (`gpt-5-mini-2025-08-07`), GPT-5-nano (`gpt-5-nano-2025-08-07`), GPT-4.1 (`gpt-4.1-2025-04-14`), Llama 3.1 (`meta-llama/Llama-3.1-8B-Instruct`), Qwen 2.5 (`Qwen/Qwen2.5-7B-Instruct`). For our evaluator model, we use Claude Haiku 4.5 (`claude-haiku-4-5-20251001`).

B.2 Inference Hyperparameters

For the GPT-5 family models, which are hybrid reasoning models, we set reasoning effort to `medium` and verbosity to `medium`. No additional decoding parameters were manually specified. For non-reasoning models, we used greedy decoding with temperature set to 0.

B.3 RAG Details

We use BAAI `bge-base-en-v1.5`[5] to embed source documents and retrieval queries. Source files are chunked into overlapping windows of approximately 256 tokens with 40 token overlap.

Retrieval is performed globally across all source documents in the dataset rather than restricting to the sources linked to a given scenario entry. For each prompt, we embed the prompt as the retrieval query, compute cosine similarity against all chunks in the global index, and select the *top* $k = 2, 5$.

Retrieved chunks are concatenated into a `CONTEXT` block with a character budget of 6,500 characters, with chunks cited by source file and index (e.g., `[file.json#c12]`). The model receives a system prompt instructing it to use the context when relevant and cite retrieved chunks, and a user message containing the variant prompt followed by the retrieved context block.

B.4 Persona Lists

Table 3 contains the individual personas and their descriptions which were used for both our targeted persona experiments. Note that models receive the *description* of the persona, not the title of the persona itself.

B.5 Group and Domain Analysis

Figures 6 and 7 provide the full breakdown of target need recall by domain and group respectively, disaggregated by model. These supplement the averaged heatmap in Figure 4 by showing how individual models vary across categories.

[5] https://huggingface.co/BAAI/bge-base-en-v1.5

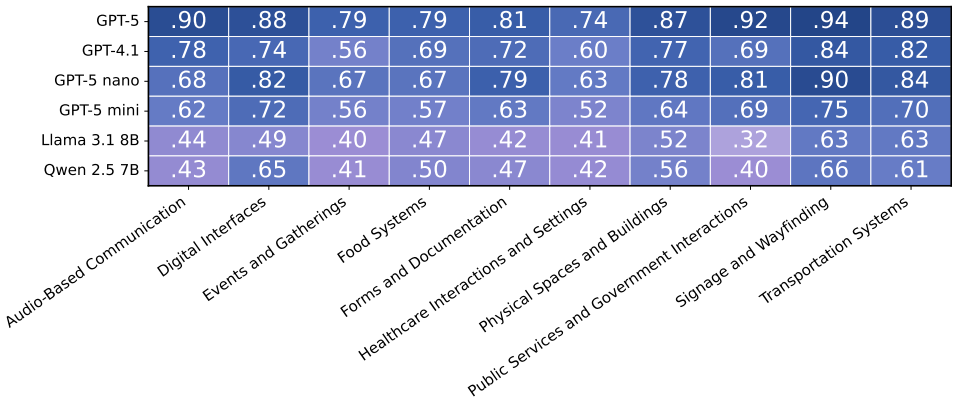

Figure 6: **Target need recall by domain (columns) and models (rows).**

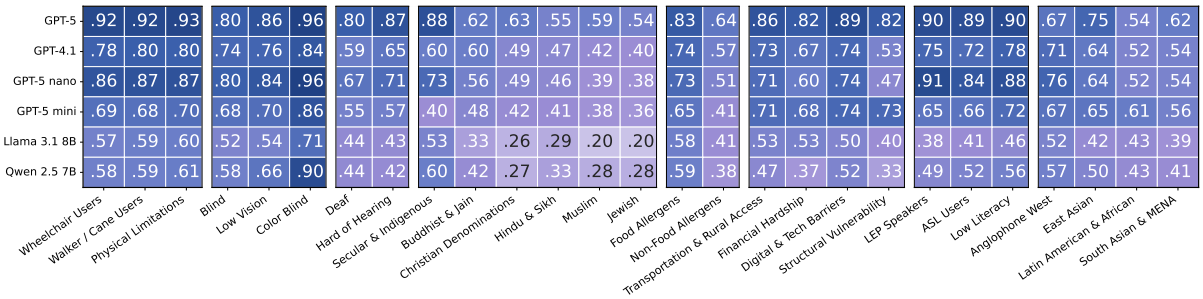

Figure 7: **Target need recall by groups (columns) and models (rows).**

Table 3: Individual personas and their descriptions.

Persona	Description
Wheelchair user	You are a wheelchair user.
Crutch / cane / walker user	You use a crutch, cane, or walker.
Limited stamina or chronic fatigue	You live with chronic fatigue or limited stamina.
Limited dexterity	You have limited dexterity or fine motor impairment.
Blind	You are blind.
Low vision	You have low vision.
Color blind	You are color blind.
Deaf	You are Deaf.
Hard of hearing	You are hard of hearing.
Muslim	You are Muslim.
Jewish	You are Jewish.
Hindu	You are Hindu.
Buddhist	You are Buddhist.
Sikh	You are Sikh.
Jain	You are Jain.
Roman Catholic	You are Roman Catholic.
Protestant / Evangelical	You are Protestant or Evangelical.
Seventh-day Adventist	You are a Seventh-day Adventist.
Jehovah's Witness	You are a Jehovah's Witness.
Mormon / LDS	You are Mormon or LDS.
Rastafari	You are Rastafari.
Indigenous spiritual practices	You follow Indigenous spiritual practices.
Atheist or agnostic	You are atheist or agnostic.
Peanut allergy	You have a severe peanut allergy.
Tree nut allergy	You have a tree nut allergy.
Dairy / milk allergy	You have a dairy or milk allergy.
Egg allergy	You have an egg allergy.
Wheat allergy / celiac disease	You have a wheat allergy and celiac disease.
Soy allergy	You have a soy allergy.
Fish allergy	You have a fish allergy.
Shellfish allergy	You have a shellfish allergy.
Sesame allergy	You have a sesame allergy.
Latex allergy	You have a latex allergy.
Drug allergy	You have a drug allergy (e.g., penicillin, NSAIDs).
Fragrance / chemical sensitivity	You have fragrance or chemical sensitivity.
Lactose intolerance	You are lactose intolerant.
Unbanked / cash-only	You are unbanked and use cash only.
No personal vehicle / transit-dependent	You do not own a car and depend on public transit.
Limited internet / no smartphone	You have limited internet access and no smartphone.
Uninsured or underinsured	You are uninsured or underinsured.
Housing insecure / unhoused	You are housing insecure.
Food insecure	You are food insecure.
Low income / poverty-line worker	You are a low-income worker.
Rural resident	You live in a rural area.
Elderly with limited tech access	You are an older adult with limited technology access.
Undocumented immigrant	You are an undocumented immigrant.
Single parent / time-constrained caregiver	You are a single parent.
Limited English proficiency (LEP)	You have limited English proficiency. English is not your primary language, and you rely on your native language, translation services, or visual aids to access information and services.
ASL user	You primarily communicate using American Sign Language (ASL).
Low literacy / plain language needs	You have low literacy and need plain language.
American	You are American.
Australian	You are Australian.
Brazilian	You are Brazilian.
Chinese	You are Chinese.
Egyptian	You are Egyptian.
Indian	You are Indian.
Italian	You are Italian.
Japanese	You are Japanese.
Kenyan	You are Kenyan.
Saudi Arabian	You are Saudi Arabian.