Is Your Chatbot a Tourist or a Townie? Quantifying Geographic and Localness Disparities in LLM Representations of Place

ZIHAN GAO, University of Wisconsin-Madison, USA
JACOB THEBAULT-SPIEKER, University of Wisconsin-Madison, USA

People are increasingly using Large Language Models (LLMs) for a sense of "localness," yet their ability to accurately and equitably represent local knowledge remains unexamined. To investigate this, we conducted a large-scale evaluation using a benchmark of over 12,000 question-answer pairs spanning structured census data, local news, and social media. Our results show that performance is strongly shaped by data modality: structured tasks expose deep limitations in numerical reasoning and calibration, while open-ended prompts reveal a clear performance hierarchy favoring informal user-generated content over professionally edited prose. Our primary finding is the existence of deep, context-dependent disparities that affect communities differently. We uncover a dual geographic bias: in formal news contexts, models exhibit a strong "urban advantage," leaving rural areas systematically underrepresented with lower semantic depth. Conversely, in social media data, models suffer an "urban penalty," struggling to navigate the conversational complexity and slang of high-density areas. This indicates that while rural locales face a "poverty of data," highly documented urban centers face a "poverty of precision." We also identify a domain bias: models are more adept at handling concrete, physical questions but consistently struggle to capture the nuanced relational and cognitive dimensions of a community. This work provides the first systematic audit of localness disparities in LLMs, revealing how they reflect and risk amplifying real-world inequities. Achieving equitable local representation requires moving beyond passive evaluation to active intervention. We call for a concerted effort from the CSCW community to build richer and more ethical datasets, design interfaces that prioritize user verification over blind trust, and architect AI systems for deeper and more just engagement with place.

CCS Concepts: • **Human-centered computing** → **Empirical studies in collaborative and social computing**; **HCI design and evaluation methods**; • **Information systems** → **Location based services**.

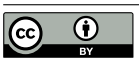
Additional Key Words and Phrases: Localness, Digital Placemaking, Large Language Models, Geographic Bias, Urban–Rural Disparities

ACM Reference Format:
Zihan Gao and Jacob Thebault-Spieker. 2026. Is Your Chatbot a Tourist or a Townie? Quantifying Geographic and Localness Disparities in LLM Representations of Place. *Proc. ACM Hum.-Comput. Interact.* 10, 2, Article CSCW022 (April 2026), 45 pages. https://doi.org/10.1145/3788058

1 Introduction

For decades, social computing research has explored how people use technology to understand and connect with the places around them [20, 36]. From neighborhood forums to location-based applications, these tools for digital placemaking have long mediated our sense of community and belonging [12, 14, 79]. Today, a powerful new tool is entering this space: LLMs. From planning a trip to moving to a new city, people are increasingly turning to LLMs for a sense of "localness"—asking

Authors' Contact Information: Zihan Gao, zihan.gao@wisc.edu, University of Wisconsin-Madison, Information School, Madison, Wisconsin, USA; Jacob Thebault-Spieker, jacob.thebaultspieker@wisc.edu, University of Wisconsin-Madison, Information School, Madison, Wisconsin, USA.



ACM 2573-0142/2026/4-ARTCSCW022
https://doi.org/10.1145/3788058

AI chatbots for guidance on everything from local culture and services [5, 22, 44] to personalized travel itineraries [32, 69]. This growing reliance on LLMs as the next generation of placemaking technology, however, rests on a largely unexamined assumption: that these models can accurately and equitably represent local life. This paper investigates this assumption, arguing that the CSCW community has a pressing need to systematically evaluate these tools before they become further embedded in civic life [13].

The challenge stems from a fundamental mismatch between the rich, multifaceted nature of "localness" and the inherent limitations of how LLMs acquire knowledge. Place is more than a point on a map; it is a complex tapestry of factual knowledge, sensory details, and the social-relational context that gives a community its character [21, 31, 56, 88]. Current LLMs often struggle to capture this nuance, demonstrably failing to generate responses that are consistently indistinguishable from those of a true local [21]. Trained on web-scale data that heavily overrepresents certain locales [39, 85, 86], the knowledge they represent is geographically biased and uneven, leading to factual errors and knowledge gaps for less-documented areas [67, 76]. This bias extends beyond mere information gaps to issues of representation. Recent studies confirm that LLMs can reproduce and amplify harmful real-world stereotypes, associating lower socioeconomic areas with negative attributes and reinforcing systemic biases [64, 71].

Given this mismatch, a systematic evaluation of LLMs' ability to represent localness is urgently needed; yet, to our knowledge, no such study exists. Existing benchmarks tend to target country-level factual recall [67] or focus on core geospatial facts rather than rich social context [65]. Furthermore, qualitative studies of place identity in social computing often focus on major, data-rich cities, reflecting a broader urban bias in the field [45], leaving critical questions about smaller or rural locales unanswered. While concepts like "place" and "localness" have been studied and theorized in HCI [40, 48, 50], sociology [28], environmental psychology [49], and geography [77, 88], it is only recently that scholars have developed a specific, operational taxonomy of "localness" that spans cognitive, physical, and relational dimensions [21]. This taxonomy provides the necessary foundation for a more holistic evaluation. Our work builds on this foundation to construct the first comprehensive framework for benchmarking "localness" in LLMs. Specifically, we ask the following research questions:

RQ1 How can the multidimensional nature of "localness" be operationalized into a robust framework for evaluating LLM performance?
RQ2 Using this framework, to what extent can current LLMs accurately represent different aspects of localness, and where do they succeed or fail?
RQ3 What geographical and social biases exist in LLMs' representation of localness, particularly across urban versus rural contexts?

To enable this investigation, we construct a large-scale benchmark of over 12,000 question–answer pairs drawn from three complementary data sources: U.S. census data, local news articles, and city-specific subreddit discussions. Together, these datasets offer a triangulated, multimodal representation of localness: census data provides a structured demographic and socioeconomic foundation; local news captures temporal, event-driven narratives; and subreddits reveal informal, community-centered discourse. This combination allows us to bridge macro-level patterns with micro-level lived experiences, supporting a more holistic and empirically grounded evaluation of "localness" in language model behavior. Using this benchmark, we systematically evaluate multiple leading LLMs across a stratified set of U.S. communities, with particular attention to geographic and socioeconomic disparities in model performance.

In conducting this work, we make six primary contributions:

(1) We introduce and adapt a conceptual framework of "localness" from HCI and placemaking research to the domain of LLM evaluation, operationalizing cognitive, physical, and relational dimensions for systematic study.
(2) We develop a robust and reproducible methodology for auditing localness in LLMs based on this benchmark, yielding the first systematic audit of localness disparities in LLM behavior.
(3) We show that LLM performance is strongly shaped by information format and modality: structured tasks expose deep limitations in numerical reasoning and calibration, while open-ended prompts reveal a performance hierarchy that favors informal user-generated content over professionally edited news prose.
(4) We uncover a dual geographic bias: in formal news contexts, models exhibit an "urban advantage" with richer, more semantically detailed responses for metropolitan areas, while in social media contexts, they suffer an "urban penalty," struggling with the conversational complexity and slang of high-density urban communities.
(5) We identify a domain bias in how LLMs represent local knowledge: models are more adept at concrete, physical questions but consistently struggle to capture the nuanced relational and cognitive dimensions of localness, such as community networks, sentiment, and locally situated practices.
(6) Finally, we derive actionable design implications for more equitable, localness-aware AI systems, calling for richer, more ethical datasets; interfaces that prioritize user verification over blind trust; and AI systems architected for deeper, more just engagement with place.

2 Related Work

Our work sits at the intersection of research on how LLMs encode and unevenly represent geographic knowledge, fairness and bias in AI-generated local content, and CSCW/HCI scholarship on digital placemaking and localness. We first synthesize prior work on global geographic recall, regional disparities, and local-scale biases in LLM outputs, then turn to placemaking systems and theoretical frameworks of localness that motivate our multidimensional evaluation of local knowledge in LLMs.

2.1 Geographic Knowledge and Biases in LLMs

2.1.1 Global Knowledge Recall. Recent studies indicate that LLMs encode a surprising amount of structured geographic knowledge. They can learn consistent spatiotemporal representations, embedding locations in a latent space that correlates with real-world geography. Researchers have even identified "space neurons" that respond to spatial coordinates, suggesting LLMs possess rudimentary world models of places and times [34]. LLMs can support tasks like mapping population and economic indicators. The GeoLLM approach combines an LLM with OpenStreetMap data to predict local population density and livelihoods, achieving performance on par with satellite-data methods [65]. These results support the notion that LLMs have absorbed a wealth of factual geography through web-scale training corpora. However, such factual knowledge is distributed unevenly, leading to significant disparities in how different regions are represented.

2.1.2 Geographic Disparities. A growing body of work reveals that LLMs do not possess uniform knowledge across all regions; instead, significant geographic disparities and biases have emerged. Benchmarks such as WorldBench and TiEBe consistently show that LLMs recall factual knowledge more accurately for high-income, Western countries than for underrepresented regions like parts of South America or Africa, reflecting underlying disparities in training data coverage [2, 67]. LLMs' performance on world public opinion data also echoes this pattern, with models performing better in Western, English-speaking, and highly developed nations compared to others [76]. These

findings align with broader observations of popularity and coverage bias in knowledge bases, where well-documented entities and regions receive disproportionately accurate and confident predictions [63, 81].

Importantly, there is a much longer history of geographic bias in user-generated and volunteered geographic information (VGI). VGI sources such as Twitter, Flickr, and Foursquare are systematically biased toward large cities, while rural areas remain comparatively underrepresented [39]. This work characterizes an *urban bias* in the production of geographic information, rooted in unequal participation and documentation. Our work builds on this literature by examining whether LLMs inherit and potentially amplify longstanding inequalities in which places are represented and how they are portrayed at the local level.

2.1.3 Fairness in Generated Local Content. While factual disparities are well-documented, they also raise broader concerns about fairness in AI-generated local content, where biases influence not just what is known, but how places are portrayed. If AI systems systematically underserve or misrepresent certain locales, they risk reproducing and reinforcing geographic inequalities not only in data recall but also in the narratives and recommendations they generate.

For instance, LLMs also exhibit subjective biases, often producing skewed portrayals of certain regions. Locations with lower socioeconomic status consistently receive lower ratings on subjective attributes like attractiveness, morality, and intelligence, revealing a troubling bias against less affluent regions [64]. In the context of personalized local information, generative models often inadvertently favor content about major cities or global topics over truly local stories. Due to data sparsity, recommendation algorithms initially show bias toward non-local content when serving local news, since local articles have smaller audiences and less training data. That means, without intervention, an AI might emphasize nationally popular content even for users seeking neighborhood news [51]. Similarly, LLM-generated recommendations of U.S. cities (for relocation, tourism, etc.) have consistent demographic biases favoring wealthy, majority-population areas [15]. Models overrepresent affluent or highly developed locations while underrepresenting smaller or more diverse communities. This poses a "rich-get-richer" risk that could amplify existing inequalities, as popular locales gain more visibility at the expense of marginalized ones. These patterns point to representation gaps in AI outputs: communities with limited digital footprints (often rural, low-income, or non-Western) are at risk of being ignored or inaccurately depicted by generative AI. Moreover, when such communities are represented, the tone may reflect stereotypes [93]. Ensuring fairness in local AI content requires more than just toxicity or other common "AI safety" filters—it arguably warrants attention to whose local narratives are being told and whose are omitted.

While previous research has highlighted factual disparities and biases in LLMs' geographic knowledge, these studies often focus narrowly on individual aspects of localness, such as demographic data or subjective portrayals of well-known cities. A comprehensive evaluation that captures the full spectrum of localness, including cognitive, physical, and relational dimensions, remains lacking. This study addresses this gap by systematically assessing LLMs across these dimensions, particularly in underrepresented locales, and exploring their implications for fair and effective digital placemaking.

2.1.4 Local vs. Global Granularity. In addition to concerns about fairness in content generation, the granularity at which geographic biases are studied also matters, echoing known Modifiable Areal Unit Problem (MAUP) issues in geography [18, 66]. While much research focuses on national levels [2, 64, 67], there is emerging evidence that geographic biases persist not only at national scale but also at finer-grained local levels, particularly when considering more diverse and community-specific data sources. For example, recent work has measured AI-generated place identity performance in different cities using multimodal data from Wikipedia and Google Street

View images [44]. One framework developed to measure local terminology and contextual knowledge boundaries in LLMs focused on New Zealand, revealing that special handling is needed for region-specific terms [73]. Their results confirm that mainstream LLMs often misunderstand or ignore region-specific terms, which can lead to errors if not accounted for. However, few studies have systematically examined geographic biases at finer local scales using heterogeneous data sources. Motivated by these gaps, our study focuses on county-level contexts, systematically exploring LLMs' performance in small- to mid-sized locales using diverse data sources such as census data, local news, and social media content.

2.2 LLMs in Digital Placemaking

The concept of digital placemaking emphasizes using digital tools (e.g., neighborhood forums, location-based social networks, civic interactive installations, and location-based games) to create a sense of place and support community-building [10–12, 25, 79, 80, 84, 89]. Recently, LLMs have emerged as a novel tool in this space, offering the ability to generate rich descriptions, answer location-specific questions, and even produce content that simulates local identity. The geographic biases of LLMs not only affect factual recall but also the potential for LLMs to serve as tools in digital placemaking, where an accurate, nuanced sense of localness is critical for fostering community engagement.

2.2.1 Localness Frameworks. Understanding the role of LLMs in digital placemaking requires grounding in theoretical conceptions of localness within HCI and CSCW, where the importance of local context has long been recognized. Foundational work emphasized that "space is the opportunity; place is the understood reality," highlighting that physical spaces become meaningful through social interpretation and practice [38]. This insight laid the groundwork for CSCW systems that account for local practices, whether in co-located work or community technologies. Building on this, research in knowledge management has identified a "geographical social tendency" in information exchange, where individuals prefer to seek advice from trusted local peers [20], a behavior particularly relevant to digital placemaking for newcomers. For people new to communities, digital placemaking is especially salient. Prior work has shown that newcomers frequently seek local information through unofficial channels, preferring peer-generated knowledge sources over formal services. For example, migrants and new residents frequently turn to subreddits and community forums to find "settlement information" and advice that, in theory, is available via formal services [68]. This behavior stems from the approachability and reliability of peer-generated local knowledge. Localness in knowledge sharing has been defined as the extent to which information exchange is shaped by regional context (for example, transit routes, neighborhood recommendations), where individuals rely on those familiar with specific locales and focus on region-specific issues [74]. More recently, Gao et al. [21] developed the Localness Conceptual Framework to operationalize "localness" computationally. They identified three interconnected primary domains: Relational, Physical, and Cognitive. The Physical domain encompasses knowledge of locations, landmarks, and environmental features; the Cognitive domain captures local cultural knowledge and insider understandings; and the Relational domain represents social connections and community attachments. These domains, which encompass 7 dimensions, 24 components, and 88 subcomponents, create a system where domains mutually reinforce each other as people develop local identity. This framework aligns with environmental psychology's sense of place components [55], providing a comprehensive structure for understanding localness beyond simple geographic residence.

2.2.2 Social Media and Vernacular Knowledge. Prior work has shown that neighborhood-focused online forums (e.g., Facebook groups, Nextdoor) can foster local information exchange and civic participation, especially as traditional local news outlets decline [3]. Such hyperlocal social media

have become key repositories of local knowledge, from daily observations to urgent alerts, and thus valuable data sources for understanding local issues. There is an opportunity for LLMs to augment or curate online information at scale, such as a chatbot that answers questions like "What is the place identity of this location?" for people to understand a place [44]. LLMs can capture salient characteristics that distinguish cities; for example, descriptions of Paris often highlight its cafés and art museums, reflecting well-known cultural landmarks [44]. Yet, LLMs tend to produce overly generic descriptions or mirror well-known sources [44]. Certain unique local nuances are lost or inaccurately portrayed, highlighting that an AI's view of place identity might be stereotyped or one-dimensional. The research on nationality bias also indicates that LLMs might sanitize or generalize descriptions to avoid causing offense, potentially glossing over the real cultural specificities of a place [93]. Thus, we must critically assess their depth of understanding versus surface-level regurgitation.

To integrate LLMs into digital placemaking tools, we must ensure they capture different dimensions of localness that matter for integration. While prior studies have explored aspects of factual, social, or cultural localness, a unified framework that evaluates all these dimensions comprehensively remains underdeveloped [15, 67]. Our framework leverages insights from digital placemaking by treating local content such as local news and subreddit discussions as crucial evidence of what localness entails. We operationalize cognitive, physical, and relational localness into distinct question types: questions about vernacular and cultural references, questions about factual data and locations, and questions about community dynamics, local concerns, and social ties. By including measures of local uniqueness alongside measures of general accuracy, we push beyond testing whether "Paris has the Eiffel Tower" and instead ask whether an LLM mentions hyperlocal location cues or community events that a true local would mention. We assess whether an AI's portrayal of a locale aligns with on-the-ground perspectives and identities. Our aim is to ensure that AI systems claiming to understand "localness" genuinely reflect the lived experiences and identities of places, rather than offering idealized or superficial representations.

3 A Multi-Source Ground Truth Framework for Localness

To systematically evaluate LLM representations of local knowledge, we construct a multi-source ground truth framework. It combines three phases: first, we assemble county-level data from census statistics, local news, and local subreddit discussions; second, we turn these sources into question–answer and question–context–answer items through automated generation and multi-stage quality filtering; third, we annotate each item with cognitive, relational, and physical localness domains [21]. This framework serves as the baseline for prompting models and assessing both the factual accuracy and contextual nuance of their responses.

3.1 Constructing Ground Truth Data

Prior studies on geographic bias in machine learning have predominantly relied on structured, large-scale datasets such as national census data, often limited to country-level granularity [64, 67]. More recent efforts have expanded this scope to include city-level coverage and diverse modalities such as news articles and image data [44, 73]. Building on these developments, we integrate structured census data, local news articles, and local subreddit discussions to build a multidimensional view of localness. We focus our analysis on U.S. counties, a unit that provides a practical balance between geographic granularity and data availability. Counties are the smallest U.S. administrative units with consistently reported socioeconomic statistics and align with electoral and governance structures, making them a meaningful unit for evaluating place representations.

Specifically, we examine 372 counties across 47 states, categorized using U.S. county-level Rural–Urban Continuum Codes (RUCC) into larger urban metropolitan (labeled "urban", RUCC 1–3;

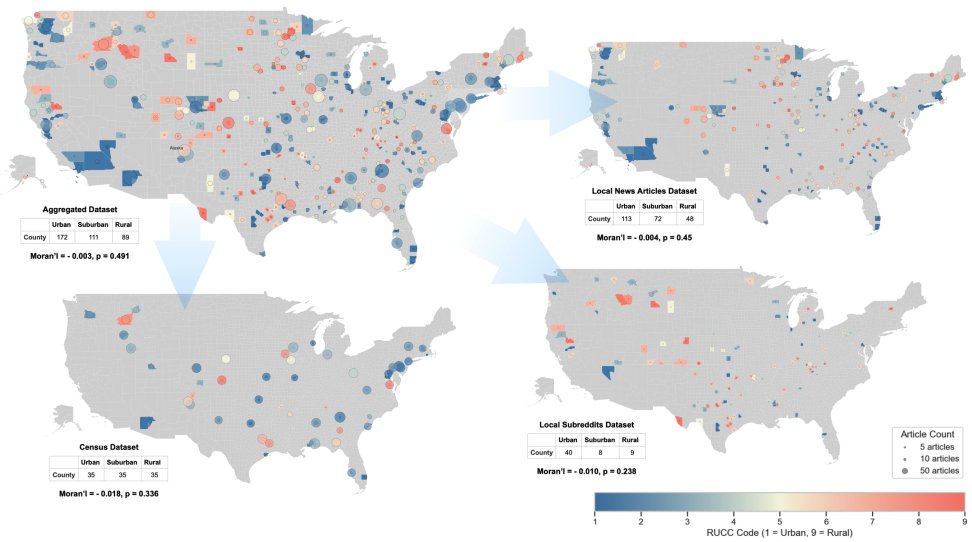


Fig. 1. County Distribution in Three Datasets

Table 1. Examples of localness metrics organized by domain, dimension, and component.

Domain	Dimension	Component	Metric	Data Source	Rationale
Cognitive	Environmental	Geographic Familiarity	*2022 cropland fertilization rate in county*	USDA Ag. Statistics Service	Reflects local knowledge of soil, weather, and land-use practices.
Cognitive	Knowledge	Historical Knowledge	*2024 # of registered historic properties with local significance*	National Register of Historic Places	Historic sites embody tangible local narratives.
Cognitive	Cultural	Local Customs / Norms	*2020 % of Southern Baptist Convention adherents*	US Religion Census 2020	Faith adherence signals moral and cultural norms.
Relational	Social / Community	Civic Engagement	*Total ballots cast in 2020 presidential election*	County Presidential Returns (2000–2020)	Voting shows civic involvement and engagement.
Relational	Emotional	Identity Connection	*2022 ethnolinguistic fractionalization index*	U.S. Census Bureau ACS Table DP05	Diversity suggests identity negotiation within communities.
Physical	Temporal	Long-Term Residence	*2022 # of residents living in same home >5 years*	U.S. Census Bureau ACS Table S0701	Long residence indicates accumulated place-based experience.
Physical	EnvironmentalPhy	Natural Environment	*2018 # of nonemployer agriculture/fishing/forestry businesses per 1,000 residents*	U.S. Census Bureau NS1800NONEMP 2018	Ecological livelihoods imply close connection to local land.

172 counties), larger suburban and non-metropolitan (labeled "suburban", RUCC 4–6; 111 counties), and smaller rural non-metropolitan (labeled "rural", RUCC 7–9; 89 counties). The overall distribution is geographically diverse and not spatially autocorrelated ($Moran's\ I = -0.003,\ p = 0.491$), ensuring representativeness without violating statistical assumptions (see Figure 1).

3.1.1 Census Data. To establish a structured baseline of local characteristics, we sampled 105 U.S. counties (35 from each RUCC category) across 30 states using a stratified sampling approach to ensure balanced coverage across the urban-rural spectrum. We confirmed spatial independence between selected counties ($Moran's\ I = -0.018,\ p = 0.336$), ensuring statistical validity of our geographic comparisons.

Our dataset includes 34 distinct metrics that operationalize the theoretical domains of localness [21], as shown in Appendix A. These metrics span three primary localness domains: Cognitive (knowledge frameworks and information patterns like local culture and commute patterns), Relational (social connections and community fabric such as demographic diversity and religious adherence), and Physical (material and spatial characteristics like built environment features and housing patterns). These are standardized, nationally available indicators relevant to community identity and lived experience. For example, the *2024 count of registered historic properties with local significance* captures community memory and cultural continuity, while the *percentage of Southern Baptist Convention adherents* reflects regional cultural norms. Table 1 provides further examples of census metrics and their localness taxonomy.

3.1.2 Local News. Our second dataset consists of news articles from local journalism, which frame and disseminate geographically grounded narratives. Unlike national outlets, local journalism provides a direct lens into region-specific concerns, community priorities, and lived conditions [1, 8, 30]. This source is crucial for representing the cognitive dimension of localness, as it captures knowledge about recent events, such as policy changes, public health alerts, or local controversies, that a knowledgeable resident would be expected to know.

We use the NELA-Local dataset [42], comprising over 1.4 million news articles from 313 local U.S. outlets, published between April 2020 and December 2021. These outlets span 255 counties in 43 states, selected via stratified sampling to ensure geographic diversity and avoid spatial clustering ($Moran's\ I = -0.004,\ p = 0.45$). Within this sample, we include 113 urban, 72 suburban, and 48 rural counties.

3.1.3 Local Subreddit Data. To access the informal, community-generated discourse often missing from institutional data, our third source comprises discussions from local subreddits. These online forums often act as digital public squares, especially in areas with limited local media [5, 68]. Following established guidelines [78], we sampled 57 counties from 30 states with active subreddit communities (minimum 50 posts in the past 30 days and at least 10 comments per post), yielding an average of 85.19 posts per county. As with other sources, we made sure that this dataset shows no spatial patterns that violate statistical assumptions ($Moran's\ I = -0.010,\ p = 0.238$).

Subreddit discourse uniquely captures vernacular knowledge, such as insider perspectives, cultural norms, recommendations, and shared grievances. By analyzing posts and their top-voted comments, we tap into the relational dimension of localness, revealing the social fabric and sentiment of a community. This provides a bottom-up perspective that complements the top-down views from census data and news media.

Together, these three datasets offer a triangulated, multimodal representation of localness: census data provides a structured demographic and socioeconomic foundation; local news captures temporal, event-driven narratives; and subreddits reveal informal, community-centered discourse. This combination allows us to bridge macro-level trends with micro-level lived experiences, supporting a more holistic and empirically grounded evaluation of localness in language model behavior.

3.2 Question-Context-Answer (QCA) Dataset Construction

We construct a multi-source evaluation dataset, $\mathcal{D} = \{\mathcal{D}_{\text{census}}, \mathcal{D}_{\text{news}}, \mathcal{D}_{\text{reddit}}\}$. For the structured census data, we generate question–answer (QA) pairs. For the unstructured news and Reddit data, we build question–context–answer (QCA) triplets that probe contextual understanding, using the generation and filtering pipeline described below.

3.2.1 Initial Question Generation. We used a zero-shot prompting strategy to generate an initial set of questions, which avoids biasing the model toward specific formats that can arise from few-shot examples.

Census Data QA Generation. In $\mathcal{D}_{\text{census}}$, for each census metric and county, we generate two types of questions: (1) a direct *fill-in-the-blank* question asking for the specific value (e.g., "What is the median household income in [*County, State*]?") and (2) a *comparative true/false* question that compares the county's value to another county with either a higher or lower value (e.g., "True or False: The median household income in [*County, State*] is higher than in [*County, State*]?"). This design tests both direct recall and relational understanding of the data. Given their factual nature, these QA pairs do not require additional context. Question templates, types, formats, and examples are detailed in Appendices B.1 and B.2.

Local News and Reddit QCA Generation. For both $\mathcal{D}_{\text{news}}$ and $\mathcal{D}_{\text{reddit}}$, our goal is to create *question–context–answer* (QCA) triplets that test nuanced, locally grounded knowledge. The process begins by isolating factual sentences from opinion-based content using a pre-trained classifier[1]. This preprocessing step ensures that our QCA generation focuses on factual information rather than subjective opinions. From these factual statements, we prompt an LLM to generate QCA triplets where:

- **Questions** are unambiguous, locally specific (e.g., referencing specific dates, places, and people), and can only be answered from the provided source material.
- **Contexts** provide neutral background to situate the question locally (clarifying the "who, where, and when") without leaking any part of the answer.
- **Answers** are concise, factual, and directly responsive to the question.

For $\mathcal{D}_{\text{reddit}}$, we add a constraint to ensure questions focus on the local topic being discussed, not on Reddit-specific platform mechanics (e.g., upvotes or user flair). The detailed prompts are available in Appendix B.3 and B.4.

3.2.2 Heuristic-Based Iterative Refinement. Our first quality control phase corrects obvious structural flaws in the generated triplets before more nuanced, computationally expensive filtering. Raw model outputs can contain predictable errors, such as directly copying the answer into the context.

To address this, we apply a set of automated heuristics to check each generated QCA triplet. These checks detect: (i) *Information Leakage*, where key entities from the answer are present in the question or context, making the question trivial; (ii) *Excessive Lexical Overlap* between the question and context, suggesting the context adds little value; and (iii) *Low Entity Coverage*, where the question fails to include specific local entities from the source document. Triplets that fail these checks are not just discarded; they are used as negative examples in a few-shot refinement loop. This iterative process teaches the generator to avoid these mistakes in subsequent attempts, improving the quality of the raw output over time. The specific formulas for these heuristics are detailed in Appendix B.5. The iterative refinement process is illustrated in the pseudocode in Appendix B.6, where we attempt to generate two high-quality QCA triplets for each document, with up to three attempts per document.

3.2.3 Reward-Model–Based Quality Filtering. Heuristics can catch objective errors but miss subjective quality issues such as clarity, naturalness, and ambiguity. We therefore used a reward model as a second-stage filter to approximate expert judgment and remove low-quality or confusing QCA triplets [90].

[1] https://huggingface.co/lighteternal/fact-or-opinion-xlmr-el

Table 2. Example QCA items from the Local-News / Local-Reddit datasets with their localness taxonomy labels.

ID	Source	Q / C / A (abridged)	Domain	Dimension	Component	Subcomponent
1	News	**Q**: How did dry and unseasonably cool weather conditions in northwestern Montana's Region 1 impact mountain grouse nesting success during the spring of 2021? **C**: Montana Fish, Wildlife and Parks biologists conduct annual upland game bird surveys across the state's seven wildlife management regions. Region 1 encompasses northwestern Montana, including Flathead, Lake, Lincoln, and Sanders counties, where spring weather patterns significantly influence bird reproduction cycles. **A**: The conditions were likely favorable for nest success and early hatchling survival, with good numbers of dusky and ruffed grouse broods observed in early summer surveys.	Physical	Environmental -Physical	Ecological Understanding	Connection to local ecosystem
2	News	**Q**: When and where is the 2021 Day of Caring event taking place for Boulder and Broomfield counties? Who can volunteer? **C**: The Mile High United Way's Volunteer Connection is an organization that connects volunteers with community service opportunities in the Boulder and Broomfield county regions of Colorado. Each year, they organize a significant volunteer event to support local community projects. **A**: The 2021 Day of Caring will be held on September 10, offering both in-person and virtual volunteer opportunities for individuals 18 and older, including corporate, community, and civic groups.	Relational	Social /Community	Active Participation	Volunteering in community
3	Reddit	**Q**: In early 2024, what specific environmental factor has caused AAA and other insurance companies to not renew home insurance policies in Los Alamos, New Mexico, and how does it relate to local fire risk designations? **C**: In early 2024, homeowners in Los Alamos, New Mexico, have been receiving nonrenewal notices from insurance providers such as AAA and Travelers. These nonrenewals are linked to underwriting guidelines related to environmental risk factors. **A**: The area had Level 3 wildfire risk area classification with high brush fire exposure.	Physical	Environmental -Physical	Natural Environment	Connection to natural spaces

We sampled 1,220 heuristic-passed documents and annotated the two candidate triplets per document as preference pairs (higher-quality vs. lower-quality), yielding a training set of 2,440 ranked examples. A reward model was trained on these pairs and then iteratively refined using a lightweight human-in-the-loop clean-up of hard or potentially mislabeled cases (details in Appendix B.7). The final model achieved strong agreement with human preferences and was applied to the full generated pool.

Filtering by the reward model removed 17.6% of triplets ($N = 6{,}950 \rightarrow 5{,}727$), primarily eliminating vague questions and incoherent reasoning chains. This step helps ensure the benchmark reflects community-salient local knowledge in a clear, answerable form, rather than artifacts of generation noise.

3.2.4 QCA Localness Feature Annotation. Two researchers independently annotated each QCA triplet according to its localness feature using the localness conceptual framework [21]. Each triplet was annotated at the subcomponent level—the most granular level in the localness hierarchy—with examples of these annotations shown in Table 2. The initial inter-rater reliability is $\kappa = 0.74$. The coding team then met to compare codes and collaboratively develop the final, unified annotations. The subcomponent annotations were then mapped upward to their corresponding localness components, dimensions, and domains within the localness framework. This systematic annotation enabled comparative analysis of LLM performance across different localness domains.

4 Evaluation Approach

This section explains how we evaluate three LLMs on our multi-source ground truth benchmark. We first describe how we generate model answers under standardized prompting for census, local news, and subreddit questions. We then outline the metrics and statistical models we use to compare performance across datasets, geographic contexts (urban, suburban, rural), and localness dimensions (physical, cognitive, relational).

4.1 Evaluating LLM Performance

4.1.1 Answer Generation. We evaluated three LLMs: GPT-4o (OpenAI), Claude-3.5-Sonnet (Anthropic), and Llama-3-70b-instruct (Meta). All models were accessed via their official paid APIs from March to April 2025. To ensure fair comparisons, we used standardized API parameters (temperature = 0.2, max-token limit = 50) for all models. We generated three independent responses for each question to assess not only accuracy but also consistency of each model's knowledge.

Our prompting strategy was tailored to the data source: (i) for Census QA pairs, which are self-contained, we prompted models with only the question, (ii) for News and Reddit QCA triplets, we provided both the question and the neutral context, as these questions often depend on the provided background information.

We also enforced strict output formats to facilitate automated evaluation. For News and Reddit, prompts required direct, factual answers and a 1–5 confidence score, while explicitly prohibiting hedging language (e.g., "it might be"). For Census questions, prompts mandated purely numerical answers for recall questions (Q1) and "True" or "False" for comparison questions (Q2/Q3), again followed by a confidence score. This structured approach forces models to provide definitive answers. Complete prompt templates are provided in Appendix C.1 and C.2, and our code and data are available online.[2] We subsequently expanded this benchmark and released a more technical evaluation as LocalBench [23].

4.1.2 Evaluation Metrics. We employed a suite of metrics to evaluate model performance, tailored to the distinct nature of our structured (Census) and unstructured (News & Reddit) datasets. A full summary of these metrics is provided in Appendix D.

For the Census data, which requires precise numerical or binary answers, our evaluation focused on accuracy and response consistency. To measure numeric accuracy for recall questions (Q1), we used Mean Absolute Percentage Error (MAPE), which calculates the average percentage deviation from the true value, making it ideal for comparing errors across different scales (e.g., population vs. income). For binary comparison questions (Q2/Q3), we used Comparison Accuracy, the percentage of correct "True/False" classifications. To assess whether a model provides stable answers over multiple attempts, we measured consistency using the Intraclass Correlation Coefficient (ICC) for numerical answers and Fleiss' Kappa for binary answers, treating each generation as a separate "rater."

For the free-text answers from our News and Reddit data, we required a broader set of metrics to evaluate quality along several dimensions. To assess how closely a model's answer matched the ground truth, we used SBERT and Cross-Encoder scores for semantic similarity ("does it mean the same thing?") and ROUGE-L and character n-gram F-score (chrF) for surface-form overlap ("does it use similar wording/strings?"). To measure factual grounding, we calculated the Entity F1 Score, which assesses the precision and recall of specific named entities (such as people, organizations, and locations). Finally, we evaluated linguistic quality. Perplexity measures the fluency and coherence

[2] https://github.com/MadCollab/LocalRepresentations

of the text, while Distinct-N, Entropy, and Self-BLEU assess lexical diversity to ensure models were not providing repetitive, generic answers.

4.1.3 Comparative Analysis. Our primary goal was to understand how LLM performance varies across geographic contexts (urban, suburban, and rural) and different types of local knowledge (physical, cognitive, and relational). We therefore conducted a structured statistical analysis tailored to each dataset.

Census Dataset. For the structured Census data, we aimed to isolate the effects of geography and localness domain on model performance for numerical estimation (Q1) and binary comparison (Q2/Q3) questions, while controlling for a key confounding variable: question difficulty.

To do this, we first constructed a composite Question Complexity Index. We started with six linguistic features from each question prompt (e.g., word count, presence of temporal terms, see Appendix E for feature details). Since these features are likely correlated, we used Principal Component Analysis (PCA) to reduce them to a single latent factor. This is a standard technique to create a robust, holistic index from multiple related variables, avoiding issues like multicollinearity in our models. The first principal component (PC1), which we used as our index, successfully captured 75.4% of the combined variance of the six features, confirming it as a strong summary of overall question complexity. (See Appendix F for detailed PCA results, including factor loadings).

With this control variable in place, we modeled LLM performance as follows:

Q1: Numerical Questions. For Q1 questions, we used a linear mixed-effects model (LME) to predict the Absolute Percentage Error (APE) of a model's answer:

$$\text{APE}_{ij} = \beta_0 + \beta_1 \texttt{Domain}_i + \beta_2 \texttt{RUCC}_i + \beta_3 (\texttt{Domain}_i \times \texttt{RUCC}_i) + \beta_4 \texttt{Complexity}_i + u_j + \varepsilon_{ij},$$

Here, the model assesses how error changes by domain and geography, with a random intercept u_j to account for baseline performance variations across counties.

Q2/Q3: Binary Questions. For Q2/Q3 questions, we used a generalized linear mixed model (GLMM) with a binomial link function to predict the probability of a correct "True/False" answer:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \gamma_0 + \gamma_1 \texttt{Domain}_i + \gamma_2 \texttt{RUCC}_i + \gamma_3 (\texttt{Domain}_i \times \texttt{RUCC}_i) + \gamma_4 \texttt{Complexity}_i + v_j,$$

This model allows us to analyze the odds of a correct classification, using the same set of predictors and county-level random effects.

To ensure the validity of these models, we performed diagnostic checks. For the LME, we confirmed residual normality using the Shapiro–Wilk test and homoscedasticity using Levene's test. For the GLMM, we examined model fit and checked for overdispersion using residual deviance statistics. Appropriate data transformations were applied when assumptions were not met (e.g., log, Box–Cox). This dual-model framework provides a unified design for analyzing both numeric and categorical responses.

Local News and Reddit Datasets. The News and Reddit datasets had an unbalanced design, with sample sizes varying substantially across different geographies and localness domains. To address this imbalance, we implemented a multi-step analytical framework. Our approach began with a Type III ANOVA, which is designed to handle unbalanced data. To further mitigate potential biases, we incorporated inverse-frequency weighting, which gives more weight to observations from smaller groups to balance their influence on the model. We then calculated Estimated Marginal Means (EMMs), which provide fair and interpretable performance comparisons as if the dataset were perfectly balanced. As a robustness check, we conducted a balanced resampling analysis by downsampling larger groups and re-running the models to confirm that our results held. Throughout this process, standard statistical assumptions were validated. The complete technical details of this framework are available in Appendix G. This analysis allowed us to examine interaction effects

Table 3. Distribution by RUCC Group and Localness Domain of Census QA Dataset

Category	Label	QA pairs	% Share
RUCC Group	Urban	2380	33.3%
	Suburban	2380	33.3%
	Rural	2380	33.3%
Localness Domain	Physical	1680	23.5%
	Cognitive	3360	47.1%
	Relational	2100	29.4%

Table 4. Ground-truth QA dataset statistics for the *Local News Article* and *Local Subreddit*. Within each six-row block, we first compare the three **RUCC groups** (Urban, Suburban, Rural) and then the three **Localness domains** (Cognitive, Physical, Relational).

			Counties	QCAs	Q-len	Q-sem	Q-lex	Ctxt-len	Ctxt-sem	Ctxt-lex	Ans-len	Ans-sem	Ans-lex
Local News Article Dataset	**RUCC Group**	Urban	113	1034	23.516	0.842	0.202	40.356	0.865	0.134	31.506	0.924	0.212
		Suburban	72	1038	24.411	0.837	0.209	41.592	0.853	0.129	29.166	0.926	0.221
		Rural	48	900	24.429	0.821	0.214	41.239	0.835	0.136	29.140	0.923	0.234
	Localness Domain	Cognitive	206	1397	23.516	0.850	0.192	41.472	0.861	0.118	35.317	0.938	0.209
		Physical	130	420	23.140	0.821	0.257	40.855	0.858	0.189	41.481	0.905	0.269
		Relational	195	1155	23.959	0.798	0.174	40.623	0.831	0.115	40.182	0.899	0.184
Local Subreddit Dataset	**RUCC Group**	Urban	40	1934	30.902	0.816	0.100	68.749	0.823	0.072	60.528	0.906	0.106
		Suburban	8	431	37.681	0.711	0.126	97.988	0.735	0.097	74.546	0.803	0.147
		Rural	9	390	37.570	0.730	0.193	94.497	0.760	0.154	73.383	0.845	0.212
	Localness Domain	Cognitive	55	1466	34.171	0.819	0.084	76.909	0.835	0.060	66.252	0.899	0.089
		Physical	29	294	33.483	0.782	0.183	93.616	0.818	0.143	67.677	0.867	0.207
		Relational	50	995	33.959	0.788	0.099	79.991	0.808	0.072	63.198	0.882	0.110

Note. **-len** = mean token length of the text field; **-lex** = lexical diversity, measured as type-token ratio (TTR); **-sem** = semantic diversity, computed as the average pairwise cosine distance between SBERT embeddings of the texts. Higher _lex and _sem values indicate greater linguistic or conceptual diversity. Colors highlight the maximum (dark), median (medium), and minimum (light) values per column within each block.

and conduct post-hoc tests (using Tukey-adjusted p-values) to understand nuanced patterns, such as whether model performance on relational questions degrades more sharply in rural areas than in urban ones.

5 Results

5.1 Overview of the Ground Truth Datasets

Our evaluation is grounded in three datasets—CensusQA, Local News QCA (LNAD), and Local Subreddit QCA (LSD)—that together provide a multifaceted view of local knowledge. They span three key content types: structured statistics (CensusQA), professionally edited prose (LNAD), and informal user-generated text (LSD). Each is annotated with geographic context (urban, suburban, rural) and localness domains (physical, cognitive, relational), forming the foundation for our analysis of LLM performance.

5.1.1 CensusQA: A Structured Baseline. The CensusQA dataset provides a structured baseline of 7,140 fact-based question-answer pairs from 105 U.S. counties, perfectly balanced across urban, suburban, and rural areas. It contains 68 questions per county, covering 34 distinct metrics mapped to cognitive (47.1%), relational (29.4%), and physical (23.5%) domains. The questions are evenly split between numerical recall (Q1) and binary comparison (Q2/Q3) formats. Table 3 summarizes the distribution by RUCC group and localness domain. This balanced factorial design enables controlled experiments on an LLM's ability to perform factual reasoning across different geographies and localness types.

5.1.2 Local News (LNAD): Patterns in Professional Prose. Our LNAD dataset includes 2,972 QCA triplets drawn from 275 local news outlets in 233 U.S. counties (113 urban, 72 suburban, 48 rural). While question lengths are consistent across regions (around 24 tokens), we observed distinct stylistic patterns across all metrics (see Table 4 for details). News answers from suburban and urban areas show the *highest semantic diversity* (Ans-sem = 0.926 and 0.924, respectively), suggesting a wider variety of topics (e.g., business, politics, arts) covered in more populous regions. In contrast, rural news answers, while slightly shorter, exhibit the *highest lexical diversity* by a significant margin (Ans-lex = 0.234).

This reveals a key pattern: rural news content in our dataset shows the lowest semantic diversity (Ans-sem = 0.923) but the highest lexical diversity. This suggests that local discourse focuses on a narrower set of topics but uses richer, less standardized language to discuss them. By domain, Cognitive domain questions are the most common (1,397 QCA, 47%) and semantically diverse (Q-sem = 0.850), while Physical domain answers are the longest (Ans-len = 41.5 tokens) and most lexically complex (Ans-lex = 0.269), as they often involve data-heavy descriptions of infrastructure or environmental change.

5.1.3 Local Subreddit (LSD): User-Generated Conversations. Our LSD dataset contains 2,755 QCA triplets drawn from 57 county subreddits, including 40 urban, 8 suburban, and 9 rural forums. All metrics are detailed in Table 4. Suburban subreddits are the most verbose, with the *longest answers* (Ans-len = 74.5 tokens). However, this verbosity masks a narrow focus; suburban posts show the *lowest semantic diversity* (Ans-sem = 0.803), suggesting that users engage in deep, elaborate discussions on a few core hyperlocal topics. This tests an LLM's depth of knowledge on niche subjects.

Urban subreddit posts, though shorter, are *the most semantically diverse* (Ans-sem = 0.906), mirroring the wide variety of concurrent discussions in a large city. Strikingly, the rural pattern persists: rural posts show the *highest lexical diversity* (Ans-lex = 0.212) but *low semantic diversity*. This suggests the pattern is not an artifact of journalism but a genuine communication style, presenting a difficult test for models to capture authentic, non-standard language.

Localness domain differences in our LSD dataset mirror those in our LNAD dataset. Cognitive localness domain entries dominate (1,466 QCA, 53%) and show high semantic diversity (Ans-sem = 0.899). Physical localness domain responses (294 QCA) are the most verbose (Ans-len = 67.7) and lexically diverse (Ans-lex = 0.207), consistent with descriptive topics such as zoning, infrastructure, or local facilities. Relational localness domain threads (995 QCA) are more compact and formulaic, with lower lexical diversity (Ans-lex = 0.110), likely due to repeated social interactions and familiar question formats.

5.1.4 Cross-Dataset Insights: The "Rural Lexical Premium". Comparing the datasets reveals systematic contrasts between professional news and user-generated Reddit content (Table 4). Reddit discussions are far more verbose, with answer lengths nearly doubling from 30 tokens in LNAD to over 60 tokens in LSD.

The most robust finding across both textual datasets is a consistent "rural lexical premium." In both LNAD and LSD, rural content has the highest lexical diversity in its answers, while also showing low semantic diversity. The implication is that rural discourse revisits hyperlocal topics but uses a broader, less standardized vocabulary. This is a critical insight for AI evaluation. Models trained on mainstream web data may fail to understand or generate the authentic, lexically rich language characteristic of these communities, potentially leading to less effective or even biased performance.

More broadly, these results point to the feasibility of evaluating LLMs by combining structured census data, narrative-rich news articles, and conversational subreddit threads. This enables *probing*

Table 5. Model-generated answer results by dataset and model. Best value per metric (within each three-row dataset block) is dark green; the middle value is light green; the worst is light red.

		Semantic		Surface		Fluency	Factual	Diversity				Self-report
Dataset	**Model**	sbert	cross-enc	rougeL$_f$	chrF	ppl	entity F1	distinct$_1$	distinct$_2$	entropy	self-BLEU	conf
Local News Article Dataset	gpt	0.672	0.576	0.211	37.513	42.094	0.228	0.762	0.982	5.498	0.459	4.916
	claude	0.636	0.619	0.122	26.708	45.704	0.159	0.642	0.943	6.339	0.352	3.569
	llama	0.669	0.587	0.194	36.157	29.282	0.209	0.695	0.955	5.620	0.364	4.929
Local Subreddit Dataset	gpt	0.712	0.586	0.233	39.750	53.636	0.294	0.735	0.971	5.819	0.433	4.032
	claude	0.670	0.573	0.138	30.020	46.434	0.208	0.628	0.936	6.434	0.235	3.325
	llama	0.715	0.588	0.238	40.771	30.304	0.312	0.679	0.946	5.700	0.497	4.408

Note. sbert, cross-enc: 0–1 — higher = greater semantic similarity. rougeL$_f$: 0–1 — higher = more overlapping n-grams. chrF: 0–100 — higher = stronger character-level similarity. ppl (perplexity): 1–∞ — lower = more natural/predictable text. entity F1: 0–1 — higher = better entity preservation. distinct$_1$, distinct$_2$: 0–1 — higher = more lexical variety (less repetition). entropy: 0–$\log_2(|V|)$ — higher = more varied vocabulary. self-BLEU: 0–1 — lower = more diverse outputs across attempts; higher = outputs are more similar/redundant. conf: 1–5 Likert self-report — higher = greater model confidence in its answer.

Table 6. Model-generated answer results on the *Census QA* set. Coloring indicates best (green) to worst (red) per column.

	Q1 (Fill-in-Blank)		Q2/Q3 (Comparison)	
Model	**Avg APE**	**Conf.**	**Acc.**	**Conf.**
gpt	0.469	3.18	0.649	4.08
claude	0.398	1.00	0.649	3.18
llama	0.562	2.84	0.384	4.17

their robustness to both stylistic variation (edited, user-generated, templated) and content modality (textual narratives versus structured statistics), offering a starting point for a comprehensive benchmark to evaluate geographically informed language technologies.

5.2 LLM Performance on Local Knowledge Tasks

This section evaluates the overall quality of model-generated answers across our three distinct data types: structured statistics (CensusQA), formal prose (LNAD), and informal user-generated content (LSD). We focus on the core task of each dataset to identify general performance trends and the fundamental strengths and limitations of current LLMs when handling local information, with detailed results in Tables 5 and 6.

5.2.1 CensusQA: Numerical Reasoning is Brittle and Poorly Calibrated. The CensusQA dataset tests the ability of LLMs to reason with precise, quantitative local data. Our findings indicate that this is a significant weakness for current models.

First, LLMs struggle with accurate recall of local statistics. For fill-in-the-blank questions (Q1), the best model still had a high Average Percentage Error (APE) of 39.8%, with the others at 46.9% and 56.2% error. This suggests that their internal knowledge of precise, county-level data is unreliable. Second, their ability to reason relationally with this data is only slightly better than chance. On comparison questions (Q2/Q3), the top accuracy was just 64.9%.

Most critically, a model's self-reported confidence is dangerously miscalibrated and cannot be trusted. Llama-3, for example, had the worst comparison accuracy (38.4%) while reporting the highest confidence (4.17 out of 5). Conversely, Claude-3.5 had the best numerical recall but the lowest confidence (1.00 out of 5). For any application requiring local numerical facts, these results show that LLMs are not only often wrong but are also unable to reliably indicate when they are uncertain.

5.2.2 Performance on Unstructured Text: Informal Reddit Trumps Formal News. When evaluated on unstructured text, a clear and consistent pattern emerged: all models performed significantly better on the informal, conversational Reddit data (LSD) than on the formal, edited news prose (LNAD). As shown in Table 5 and the figures in Appendix I, key metrics like semantic similarity, surface alignment, and factual accuracy were consistently higher for the LSD dataset. For example, models achieved higher semantic similarity on Reddit, indicating a better grasp of the core meaning of questions and answers (e.g., top SBERT score of 0.715 on LSD vs. 0.672 on LNAD). They were also better at surface-level mimicry, more effectively replicating the phrasing and style of Reddit users than that of professional journalists (e.g., top ROUGE-L score of 0.238 on LSD vs. 0.211 on LNAD). This suggests that the conversational, user-generated style of Reddit is a "sweet spot" that aligns better with the training and architecture of modern LLMs than the dense, formal prose of news.

Despite the better performance on Reddit, factual accuracy remains a major bottleneck across both datasets. The single best Entity F1 score was only 0.312 (Llama-3 on LSD), meaning the best model correctly identified less than a third of the key local entities. This reveals a critical gap between conceptual understanding and factual grounding. Models are often able to grasp the semantics of a question (SBERT scores > 0.63) but fail to retrieve or generate the correct, specific facts.

While factuality is a weakness, fluency is a more solved problem. Llama-3 consistently produced the most fluent and natural-sounding text (lowest perplexity), but it was not always the most factually accurate. This confirms that a coherent, well-written answer is not a reliable signal of its correctness. Furthermore, models exhibit greater response diversity when prompted with Reddit's conversational style. We see this in the Self-BLEU metric, where lower scores indicate less repetition across multiple answer attempts. The average Self-BLEU score was lower for the Reddit dataset, with one model reaching a score as low as 0.235. This suggests the informal context of Reddit may encourage more creative and varied outputs, whereas the factual nature of news might constrain models to a narrower, more repetitive response pattern.

5.2.3 Synthesis: Key Capabilities and Limitations. Across all three data types, our findings paint a clear picture of the capabilities and fundamental limitations of current LLMs for local knowledge tasks. We synthesize our results into four key takeaways, showing how known failure modes manifest *specifically for localness* across modalities and geographic settings.

Modality Drives Performance and Bias. There is a clear performance hierarchy: Informal Conversational Text (Reddit) outperforms Formal Prose (News), which in turn outperforms Structured Numerical Data (Census). Performance degrades as tasks demand numerical precision and schema-sensitive reasoning rather than linguistic continuation. This pattern confirms LLMs remain brittle on structured reasoning [59], but for localness, this modality dependence creates the structural conditions for the urban advantage and urban penalty.

Grounding is Weak and Hallucination is Targeted. Even in the best-case modality (informal Reddit data), factual accuracy for local entities is insufficient for real-world use. Models frequently generate plausible but incorrect local details, reflecting a persistent gap between semantic plausibility and grounded recall. This reality, where hallucination is a known failure mode [16, 43], is unevenly distributed: we demonstrate that hallucination varies systematically by localness domain and geographic setting, shaping who is realistically represented by the model.

Fluency is a Deceptive Strength, Masking Miscalibration. All models generate coherent, stylistically appropriate text, but surface-level fluency is not a proxy for correctness. This fluency bias risks reinforcing inequities, as users may over-rely on outputs that feel polished [52]. Compounding this, a model's self-reported confidence is often unreliable and prone to overconfidence on incorrect

Table 7. Model performance across RUCC groups and localness domains in the CensusQA dataset.

(A) RUCC Group

Q-type / Metric	Urban			Suburban			Rural		
	gpt	claude	llama	gpt	claude	llama	gpt	claude	llama
Q1 (Fill-in-the-Blank)									
Avg APE ↓	0.459	0.402	0.561	0.472	0.402	0.563	0.477	0.392	0.564
Conf. ↑	3.15	1.00	2.88	3.19	1.00	2.80	3.19	1.00	2.84
Q2/Q3 (Comparison)									
Acc. ↑	0.669	0.650	0.384	0.663	0.653	0.383	0.616	0.642	0.384
Conf. ↑	4.12	3.18	4.22	4.06	3.15	4.16	4.04	3.20	4.13

(B) Localness Domain

Q-type / Metric	Cognitive			Physical			Relational		
	gpt	claude	llama	gpt	claude	llama	gpt	claude	llama
Q1 (Fill-in-the-Blank)									
Avg APE ↓	0.539	0.393	0.663	0.431	0.407	0.500	0.389	0.401	0.453
Conf. ↑	3.22	1.00	2.87	3.19	1.00	2.70	3.08	1.00	2.90
Q2/Q3 (Comparison)									
Acc. ↑	0.641	0.561	0.457	0.735	0.650	0.407	0.595	0.788	0.247
Conf. ↑	4.10	3.25	4.17	4.11	2.99	4.01	4.00	3.20	4.29

Note. Panel (A) reports results by RUCC group; Panel (B) reports results by localness domain. Avg APE (lower is better), Acc. (higher is better), and Conf. (mean self-reported confidence). Within each row and subgroup, green = best, light green = middle, and red = worst.

answers [27]. This miscalibration is particularly risky in localness contexts, where high confidence can be misleading for place-based decisions.

Implications for CSCW: Verification over Answers. These results show LLMs are not yet reliable sources of factual local knowledge. Their strong fluency, weak grounding, and miscalibrated confidence motivate a design philosophy that shifts from providing answers to enabling verification. We elaborate concrete implications for transparency, uncertainty visualization, and community narrative stewardship in Section 6.4.

5.3 Performance Disparity: Geographic Settings and Localness Domains

While our overall results point to general LLM capabilities, we now turn to how performance varies across different contexts. To understand if LLMs serve all communities and topics equally, we conducted a systematic analysis of performance disparities across our three datasets. We examine variations driven by Geographic Setting (Urban, Suburban, Rural) and Localness Domain (Cognitive, Physical, Relational), using the statistical models appropriate for each dataset's structure.

Our analysis reveals a split in what seems to drive performance disparities. For structured Census data, performance is dictated by the type of numerical reasoning required by a localness domain, with geography playing a minimal role. However, this pattern shifts dramatically for unstructured text. For professionally edited news articles, we find a strong urban advantage, suggesting a bias towards high-resource media outlets. This trend reverses again for user-generated Reddit content, where a significant urban penalty emerges, indicating that the chaotic, slang-heavy nature of high-traffic online communities poses a unique challenge for LLMs.

5.3.1 CensusQA: Domain, Not Geography, Drives Numerical Reasoning Errors. Our CensusQA dataset isolates the task of factual numeric reasoning. By removing stylistic and narrative cues, it provides a controlled test of an LLM's ability to reason with quantitative local data. The findings reveal that performance is driven primarily by the type of numerical reasoning required by a localness domain, with question complexity acting as a universal handicap.

Geographic Disparities are Surprisingly Muted. Contrary to what might be expected, geographic setting had a limited and inconsistent effect on performance for this task (see Table 7, Panel A). For GPT-4o and Llama-3, there were no significant differences in accuracy or error across urban, suburban, and rural counties (all $p > .18$). The structured, de-contextualized nature of the questions likely removes the regionally-embedded linguistic cues that models often rely on, leading to more uniform performance. Moreover, it may be that all models are equally ineffective at this task, which does not preclude the possibility that increasing model performance for this task may also lead to geographical differences.

The sole exception was Claude-3.5, which revealed a curious inversion in its geographic performance. For numerical recall (Q1), it was significantly more accurate (i.e., had lower error) in rural settings and less accurate in urban ($\beta = 0.053$, $p = .001$, Cohen's $d = 0.49$). Yet for comparative reasoning (Q2/Q3), its accuracy was significantly higher in urban and suburban regions (rural vs. urban: $\beta = 0.048$, $p < .001$, Cohen's $d = 0.44$; rural vs. suburban: $\beta = 0.041$, $p = .001$, Cohen's $d = 0.37$). This contradictory pattern—high precision on rural facts but stronger relational reasoning in urban contexts—suggests that Claude-3.5's internal representation of local statistics is highly sensitive to both geography and the specific reasoning format of the question.

Domain Disparities Reveal Fundamental Reasoning Challenges. In contrast to muted geographic effects, our localness domain was a robust and significant driver of performance for all models. The reason for these disparities appears to be rooted in the different types of numerical cognition that each domain's metrics demand (see Table 7, Panel B).

The challenge for LLMs varied with the underlying structure of the data. Models generally performed best on Cognitive domain questions, which served as our reference category. These metrics are often population-normalized rates or percentages (e.g., "percent who commute by public transit"). This normalization ensures most values fall within a familiar 0–100 range, making them numerically "easier" for LLMs to generate. Accordingly, no model showed a large performance deficit on this domain (all effect sizes $|d| \leq 0.28$).

The difficulty increased with Physical domain questions, which test a model's ability to handle scale. These items often ask for raw counts (e.g., "total owner-occupied units") that span several orders of magnitude across counties. This wide variability proved challenging, as seen in Claude-3.5's significantly higher numerical error (APE) on these questions ($\beta = 0.062, p = .001, d = 0.24$). Interestingly, when this same information was posed as a simple binary comparison, Claude-3.5's accuracy rose ($\beta = 0.042, p = .004, d = 0.28$), implying that a categorical framing helps suppress the burden of precise magnitude calibration.

Finally, Relational domain questions posed the greatest challenge in abstraction. These metrics encode second-order relationships like diversity indices or household averages, requiring reasoning about relationships among categories rather than a single magnitude. The difficulty of this abstraction is evident in Claude-3.5's higher numerical error ($\beta = 0.061, p = .001, d = 0.24$). However, Llama-3 excelled on these same items, showing a large reduction in error ($\beta = -0.121, p < .001, d = 0.47$), indicating a superior ability to internalize and generalize relational distributions.

Question Complexity is a Universal Handicap. Across all models and tasks, question complexity had the most powerful and consistent effect. APE increased significantly with complexity for GPT-4o ($p < .001$), Claude-3.5 ($p < .001$), and Llama-3 ($p < .001$). Q2/Q3 accuracy decreased significantly with complexity in all models. For instance, for every one-point increase on our 10-point complexity scale, Llama-3's error increased by an average of 5.3% ($\beta = 0.053, p < .001$). Notably, confidence *increased* with complexity for Claude and Llama, indicating overconfidence on harder tasks. This inverse calibration suggests models mistake complexity for informativeness, a known issue in instruction-tuned LLMs that prioritize verbose or well-structured inputs [54, 58, 92].

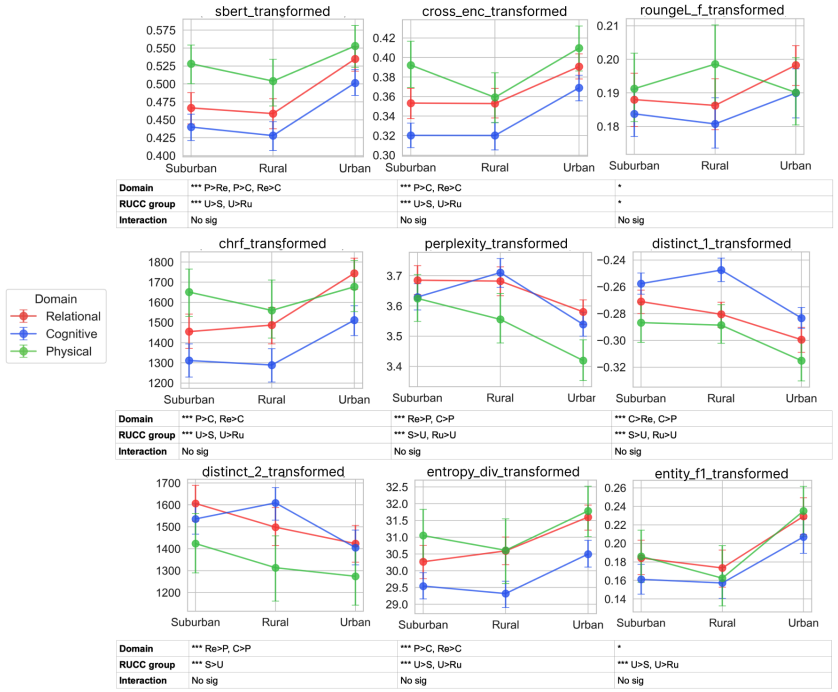


Fig. 2. Interaction Between RUCC Group and Localness Domain in Model Performance of LNAD (Model: GPT-4o)

5.3.2 Local News: An "Urban Advantage" and a Bias Toward Physical and Relational Domains. Our LNAD data, grounded in professionally edited news, challenges models to generate contextually relevant and stylistically appropriate answers. Our analysis reveals two powerful and largely independent performance trends: a geographic disparity that consistently favors urban sources, and a domain disparity where models perform best on physical and relational topics. We confirmed this using ANOVA, where a large F-statistic indicates a strong difference between groups. Both geography and localness domain were significant main effects across nearly all metrics and models, while interaction effects between them were negligible, as visualized for GPT-4o in Figure 2. This indicates that the urban advantage applies to all topics, and the topic-based advantage applies in all regions.

For clarity in the body of this paper, we report the specific statistical results for GPT-4o as a representative example here and in the following Section 5.3.3. We note that Claude-3.5 and Llama-3 exhibited similar overall trends across major findings. Detailed statistical results for all three models are provided in Appendix H.

Geographic Disparities: A Consistent Urban Advantage. Across all dimensions of performance, a clear "urban advantage" exists. LLMs are better at understanding, reproducing, and recalling facts from news articles originating in urban counties. To quantify these differences, we report the direct performance gap between groups (Δ) and its 95% Confidence Interval (CI). The CI provides a range of plausible values for the true gap; crucially, if this range does not contain zero, we can be confident the observed difference is real.

This urban advantage is evident across the board (see Table 8). For semantic understanding (SBERT), the performance gap between urban and rural sources was substantial (Δ_{UR} =

Table 8. Model performance across RUCC groups (Rural, Suburban, Urban) in the Local News Article Dataset. Best values within each row and group are green, middle light green, and worst red.

Metric	Rural			Suburban			Urban		
	gpt	claude	llama	gpt	claude	llama	gpt	claude	llama
sbert_similarity	0.649	0.615	0.648	0.658	0.622	0.655	0.707	0.667	0.700
cross_encoder_score	0.560	0.616	0.576	0.562	0.612	0.577	0.604	0.629	0.608
rougeL_f	0.207	0.117	0.188	0.209	0.119	0.192	0.216	0.128	0.200
chrF_score	36.448	25.818	35.044	36.551	25.846	35.280	39.406	28.346	38.005
perplexity↓	44.919	46.915	30.826	43.372	46.244	30.365	38.351	44.108	26.849
distinct_1	0.770	0.643	0.701	0.769	0.648	0.702	0.748	0.635	0.683
distinct_2	0.983	0.943	0.956	0.983	0.944	0.958	0.980	0.941	0.952
entropy_diversity	5.462	6.301	5.589	5.464	6.320	5.602	5.565	6.392	5.665
entity_f1	0.200	0.145	0.181	0.210	0.145	0.195	0.271	0.184	0.247

Table 9. Model performance by Localness Domain (Cognitive, Physical, Relational) in the Local News Article Dataset. Best values within each row and group are green, middle light green, and worst red.

Metric	Cognitive			Physical			Relational		
	gpt	claude	llama	gpt	claude	llama	gpt	claude	llama
sbert_similarity	0.653	0.622	0.651	0.714	0.670	0.707	0.680	0.639	0.676
cross_encoder_score	0.558	0.613	0.576	0.605	0.637	0.613	0.587	0.619	0.592
rougeL_f	0.207	0.117	0.190	0.215	0.125	0.193	0.214	0.125	0.198
chrF_score	36.247	25.490	34.997	39.397	28.092	37.386	38.358	27.677	37.113
perplexity↓	42.619	44.470	29.814	38.804	44.605	26.397	42.654	47.597	29.687
distinct_1	0.772	0.646	0.704	0.747	0.632	0.674	0.756	0.641	0.692
distinct_2	0.983	0.944	0.958	0.980	0.940	0.950	0.982	0.943	0.954
entropy_diversity	5.444	6.317	5.581	5.563	6.392	5.691	5.541	6.347	5.641
entity_f1	0.214	0.150	0.196	0.236	0.164	0.228	0.242	0.168	0.217

+.058, 95% CI [+.053, +.066]). The pattern was even more pronounced for factual accuracy (Entity F1), where urban answers were significantly more correct than rural ones, with a large performance gap ($\Delta_{UR} = +.072$, 95% CI [+.058, +.083]). Models were also more fluent when responding to urban prompts, showing a significant decrease in perplexity ($\Delta_{UR} = -.104$, 95% CI [−.166, −.045]).

These results strongly suggest that LLMs are implicitly optimized for the high-density, editorially standardized, urban-centric content. This is likely due to edited news from high-traffic urban outlets being better represented in the pre-training data, and carrying a more standardized linguistic structure, which LLMs reproduce with ease. In contrast, rural reporting may exhibit less standardized style, sparser event coverage, and more idiosyncratic language, explaining the consistent performance degradation.

Domain Disparities: Physical and Relational Domains are "Easier." Performance also varied significantly by the type of local knowledge required, with models performing best on Physical and Relational localness domain (see Table 9).

Questions in the Physical domain (e.g., infrastructure) received the highest scores on six of nine metrics. For example, responses to Physical questions were more semantically aligned with the ground truth than Cognitive ones ($\Delta_{PC} = +.053$, 95% CI [+.045, +.061]) and were more fluent (lower perplexity, $\Delta_{PC} = -.12$, 95% CI [−.17, −.07]). Relational items (e.g., community events) yielded the strongest factual accuracy across all domains (Entity F1 advantage over Cognitive: $\Delta \approx .011$, $p < .05$ across all models).

In contrast, models struggled most with Cognitive domain questions. These topics, rich in abstract socioeconomic concepts, prompted the lowest semantic match and factual accuracy. This is likely because Physical items benefit from descriptive clarity and fixed terminology, while Relational

Table 10. Model performance across RUCC groups (Rural, Suburban, Urban) in the Local Subreddits Dataset. Best values per metric within each group are green; middle values are light green; worst values are red.

Metric	Rural			Suburban			Urban		
	gpt	claude	llama	gpt	claude	llama	gpt	claude	llama
sbert_similarity	0.755	0.695	0.755	0.763	0.680	0.728	0.617	0.636	0.662
cross_encoder_score	0.610	0.563	0.590	0.595	0.570	0.585	0.553	0.586	0.588
rougeL_f	0.243	0.115	0.248	0.240	0.124	0.246	0.216	0.175	0.221
chrF_score	42.0	31.0	41.6	41.7	29.0	41.2	35.6	30.1	39.5
perplexity↓	57.142	47.384	31.361	60.595	51.264	34.690	52.805	46.501	28.589
distinct_1	0.726	0.623	0.659	0.744	0.637	0.681	0.748	0.639	0.688
distinct_2	0.970	0.938	0.946	0.977	0.945	0.956	0.974	0.940	0.951
entropy_diversity	5.923	6.504	5.777	5.950	6.572	5.795	5.813	6.442	5.731
entity_f1	0.348	0.231	0.339	0.365	0.220	0.337	0.169	0.172	0.259

Table 11. Model performance by Localness Domain (Cognitive, Physical, Relational) in the Local Subreddits Dataset. Best values per metric within each group are green; middle values are light green; worst values are red.

Metric	Cognitive			Physical			Relational		
	gpt	claude	llama	gpt	claude	llama	gpt	claude	llama
sbert_similarity	0.714	0.661	0.713	0.741	0.678	0.726	0.682	0.672	0.706
cross_encoder_score	0.587	0.562	0.579	0.590	0.578	0.596	0.580	0.578	0.589
rougeL_f	0.232	0.126	0.236	0.241	0.145	0.238	0.226	0.143	0.240
chrF_score	39.6	29.1	39.2	41.0	30.9	40.0	38.6	30.1	43.1
perplexity↓	53.503	46.815	30.447	53.729	45.669	32.461	53.803	46.099	29.456
distinct_1	0.726	0.624	0.675	0.736	0.625	0.678	0.749	0.633	0.685
distinct_2	0.966	0.934	0.941	0.971	0.933	0.947	0.977	0.940	0.953
entropy_diversity	5.827	6.421	5.694	5.836	6.456	5.692	5.801	6.446	5.711
entity_f1	0.289	0.205	0.297	0.319	0.213	0.312	0.273	0.205	0.328

items leverage concrete social knowledge. Cognitive items are harder for models to ground in verifiable facts due to their abstraction and wide topical variance.

5.3.3 Local Subreddits: An "Urban Penalty" Driven by Social Complexity. In the informal, user-generated context of the LSD data, the performance disparities shift dramatically. Here, the specific geographic setting becomes a stronger predictor of performance for most meaning-focused metrics. For example, in our ANOVA model for semantic similarity (SBERT), the main effect for geography was substantially larger ($F = 119.37, p < .001$) than the effect for domain ($F = 76.31, p < .001$). This analysis reveals an inversion of the pattern seen in local news, with significant and complex interactions between geographical settings and localness domains.

Geography Dominates: A Surprising "Urban Penalty" LLM performance is worse on content from urban subreddits compared to suburban and rural ones (see Table 10). This contrasts sharply with the "urban advantage" seen in the news dataset. The performance drop in urban contexts is significant for semantic understanding (SBERT: $\Delta_{\text{UR}} = -.122$, 95% CI $[-.15, -.10]$) and factual accuracy (Entity F1: $\Delta_{\text{UR}} = -.080$, 95% CI $[-.10, -.06]$). A plausible explanation for this is urban subreddit threads, while plentiful, are often more topically diverse, and heavier on slang and implicit context. This conversational complexity appears to make it harder for models to anchor their understanding and retrieve correct facts, despite the abundance of urban data in their training corpora [39, 60, 91]. Interestingly, fluency once again bucks the trend. Models produce more fluent (i.e., lower perplexity) text for urban prompts, even as their semantic and factual accuracy

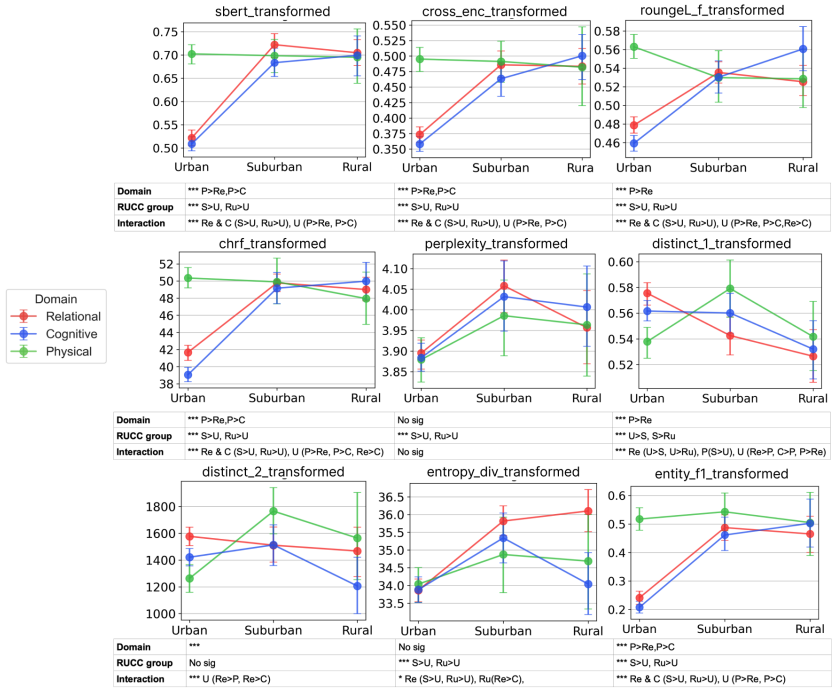


Fig. 3. Interaction Between RUCC Group and Localness Domain in Model Performance of LSD (Model: GPT-4o)

falters ($\Delta_{US} = -.14$, 95% CI $[-.21, -.07]$). This further reinforces that fluency is a poor proxy for correctness.

Domain Effects: Physical > Relational > Cognitive. Within the conversational noise of Reddit, the same domain hierarchy persists: models perform best on concrete domains. Physical domain posts (e.g., about landmarks, street names) yield the highest semantic alignment and factual accuracy, while abstract Cognitive domains are the most difficult (see Table 11).

The performance gap between the easiest and hardest domains is large. For factual accuracy (Entity F1), the advantage of Physical over Cognitive domain is substantial ($\Delta_{PC} = +.17$, 95% CI $[+.12, +.21]$), as well as semantic alignment ($\Delta_{PC} = +.068$, 95% CI $[+.03, +.10]$). Relational localness content (neighborhood comparisons, school rivalries) sits in the middle, while Cognitive localness domain (attitudes, memories) is hardest. These results might be due to Physical localness attributes being linguistically concrete and repeatedly labeled, giving the models stable lexical anchors. Cognitive localness discourse is abstract, sentiment-laden, and highly idiolectal, resulting in wide variance and prediction errors in LLM outputs.

Interaction Effects: The Urban Penalty is Worst for Non-Physical Domains. Unlike the news data, here we find significant interaction effects between geography and domain. The "urban penalty" is not uniform; it is larger for Relational and Cognitive domains than for Physical ones, as visualized in Figure 3.

For example, the performance gap in semantic understanding (SBERT) between urban and suburban posts is moderate for the Physical domain ($\Delta = -.01$, 95% CI $[-.04, +.02]$), but it becomes enormous for Relational domain ($\Delta = -.20$, 95% CI $[-.23, -.17]$). This shows that while knowledge

about Physical localness is geographically robust, the ability to understand Relational and Cognitive localness is highly brittle and fails most acutely in complex urban contexts.

6 Discussion

6.1 Where LLMs Thrive—and Fail—on Localness

Our study reveals a layered and, at times, contradictory portrait of LLM performance on local knowledge tasks. Across more than 12,000 question-answer pairs spanning structured data, news media, and social discourse, we identify both technical bottlenecks and socio-cultural blind spots. LLMs can fluently generate answers about local areas, but their knowledge is uneven, skewed, and often superficial. Structured tasks (like CensusQA) expose their limitations in numeric reasoning and calibration, while open-ended prompts (from Reddit or local news) show stronger linguistic fluency but persistent gaps in factual precision.

Three core takeaways define our findings:

- *Task–Model Fit is Critical:* LLM performance varies not only by geography but by modality—LLMs perform better on Reddit data than on News data, which in turn still outperforms Census data—highlighting the need to align models with input structure and style.
- *Localness Dimensions are Unevenly Represented:* Physical localness is more accessible to models than Relational or Cognitive dimensions, which demand deeper sociocultural grounding.
- *Geographic Disparities are Real and Structured:* Urban prompts are a "sweet spot" in news but a "trap" in community Q&A. Rural contexts, though lexically rich, remain semantically underrepresented.

These gaps are unlikely to vanish simply through scaling. They track slow-changing structural factors such as uneven digital documentation of place, urban bias in volunteered and crowdsourced information [39], and thin coverage of lived local narratives in training corpora. Below, we discuss how these findings should shape dataset design and model development, and what kinds of socio-technical interventions are needed from the CSCW community to monitor these disparities over time and build systems that support more equitable digital representations of place.

6.2 Rethinking Data for Localness

Our findings underscore that an LLM's understanding of "localness" is tightly coupled to what gets documented about place, by whom, and in what form. Building models that support equity requires rethinking how we source, steward, and benchmark local data.

6.2.1 The Need for Broader, Safer Benchmarks. Our work reveals a pressing need to expand the geographic and semantic diversity of localness benchmarks. The "lexical premium" but "semantic sparsity" we observed in rural data mirrors a broader "visibility gap" in AI, where datasets overrepresent urban, high-resource locales [67, 81], amplifying existing inequities [15, 64], and reflecting well-known patterns in geographic user-generated content, which gets used as training data for these models [37, 47, 85]. A more comprehensive benchmark should incorporate a richer variety of locales and capture the voices and knowledge of diverse communities. Increasing linguistic diversity and complexity is crucial: local knowledge is often conveyed in regional dialects, indigenous languages, or code-switched vernaculars that our primarily English-language benchmark does not cover. Future expansions should include multilingual queries and culturally specific terminology to ensure models can handle the full spectrum of how local information is expressed.

6.2.2 The Ethical Stewardship of Public Data. Expanding benchmarks must be coupled with responsible governance. Our use of Reddit highlights this tension: while it is a valuable source of vernacular knowledge, "publicly available" does not mean "ethically unproblematic" [17, 75]. Subreddit data

often contains context-specific, sensitive, or potentially identifying content. Before public release, benchmark datasets must undergo rigorous cleaning to align with subreddit rules, user expectations, and platform norms—a principle grounded in contextual integrity [70] and increasingly recognized in AI dataset ethics [26, 41]. Cleaning should include automated personally identifiable information (PII) removal, filtering of stigmatized or harmful content, and, where applicable, consultation with community moderators [29]. Red-teaming techniques, such as adversarial testing for toxic or privacy-leaking content, should be adopted as a standard phase in benchmark release processes [72]. These steps not only safeguard users but enhance public trust in civic AI technologies.

Simultaneously, local-community subreddits are not a neutral reflection of the offline communities they discuss. The demographic skew on Reddit [4] and potential for non-local contributions [48] mean our benchmark captures a specific, partial slice of "localness." Furthermore, widely used Reddit corpora contain structural data gaps [19], raising the risk that some locations or time periods are unevenly represented. These biases likely amplify the perspectives of active online subcultures while leaving broader community experiences under-sampled. Consequently, we treat Reddit as a crucial but incomplete lens, necessitating future benchmarks that are multimodal, community-driven, and governed by explicit stewardship.

6.2.3 Moving Beyond Text to Multimodal, Community-Sourced Data. Finally, to capture the full texture of local life, we must move beyond text. Sources like Reddit often skew toward narrow demographics. A more holistic and inclusive benchmark and training corpus must expand to draw from multimodal and community-driven data sources. VGI, local blogs, community news sites, and especially social media platforms like TikTok, YouTube, and Instagram may offer valuable and underutilized representations of everyday local life [22]. For example, a TikTok video about a festival in a small town can convey cultural meaning, social bonds, and spatial layout. Advances in multimodal learning now make it feasible to systematically mine these representations and distill them into structured attributes of place. Previous research in urban informatics and social computing has laid groundwork for this by modeling city structure [12], semantic neighborhoods [46], and crowd-sensed place attributes [45]. Building on this lineage, future systems should be designed to ingest and structure these multimodal signals, creating computational representations that more closely reflect how people perceive and attach meaning to place.

6.3 Interpreting Performance Disparities

Our findings on geographic and domain disparities are not just statistical artifacts; they reveal the deep influence of pre-training data and the societal risks of deploying biased models.

6.3.1 The Role of Pre-Training Data in Modality Bias. We observed a clear performance hierarchy: *Reddit data outperforms News data, which in turn outperforms Census data.* This aligns with pre-training data composition. LLMs are trained on web-scale corpora dominated by informal, conversational, user-generated text [7, 91], so they are most fluent in Reddit-like text and least robust on structured numerical tasks. Crucially, this pre-training stream is also geographically uneven: user-generated and VGI platforms disproportionately document urban areas [39, 62, 86], and LLM audits find systematic underperformance for less-documented and lower-resource locations [64]. Thus, models see both *more informal data* and *more urban data*, helping explain our modality hierarchy and the persistent urban advantage in formal sources.

Consequently, reducing localness disparities requires shifting training data distributions, not merely scaling parameters. Following domain-adaptive pre-training [35], targeted upsampling of local journalism, municipal records, and community narratives, especially from rural and other under-documented locales, can provide the factual grounding that today's models lack. Even

modest late-stage upsampling has been shown to yield meaningful gains in domain reliability [6], suggesting a practical path toward narrowing the gaps.

6.3.2 Who Gets Represented? Disparities, Stereotypes, and the Risk of AI-Amplified Inequality. The performance disparities we found pose a direct risk of amplifying real-world inequality. The "urban advantage" in the news data means that models are better at representing well-documented, often more affluent urban centers. This pattern echoes prior findings on popularity bias in web-scale models [64, 81], where affluent, densely documented areas dominate AI outputs, while less-visible communities are further marginalized. As noted above, particularly given the salience of user-generated content being used as training data [60] for these models, this trend also echoes—and may be, in part, due to—well-established pro-urban trends in user-generated platforms more broadly [39, 47, 85]. Our news- and census-based results can be read as one way this "urban bias in VGI" propagates into LLM behavior: places with more and richer upstream digital traces are easier for models to represent.

When people use LLMs for recommendations on travel or moving [32, 44], this bias can create feedback loops, directing attention and resources to already-visible places while further marginalizing rural and under-documented communities. The "urban penalty" we found in Reddit data further complicates this, underscoring that no single geographic bias is universal. One plausible contributor to this pattern is the composition of Reddit itself. Urban subreddits are typically larger and more active, and prior work suggests that high-visibility places attract more outsiders, more rapidly changing discourse, and more contested narratives [39, 48]. Amidst Reddit's specific style and heterogeneity [4], this complexity creates noise that models struggle to parse, contrasting with rural data that appears lexically rich but semantically sparse, which may reflect narrower online subpopulations. Localness domains further stratify performance. Physical localness questions—often tied to concrete infrastructure or geography—were easier for models than Relational localness (e.g., community structure) or Cognitive localness (e.g., perceptions of safety). This aligns with CSCW literature on local information practices, which shows that factual queries are easier to answer via formal channels, while relational knowledge is typically conveyed through social ties and informal exchanges [68, 74]. This reflects the lack of high-quality, social-sourced narratives in pre-training corpora—a gap also noted in recent critiques of AI place identity work [44, 93]. Addressing this will require better sourcing and modeling of peer-based knowledge, such as lived experiences, local controversies, and community rituals.

The lack of high-quality peer-based narratives leaves a critical open question: whether these performance gaps also reflect learned stereotypical biases. Future research must move beyond accuracy to audit the content of LLM responses for geographic stereotypes, examining whether models describe different places not just with varying accuracy, but with different tones, sentiments, and implied value judgments.

These dynamics are unlikely to resolve through model scaling alone, as they are rooted in the slow-evolving geography of digital traces. Without intervention, LLMs risk transforming uneven data into authoritative-seeming narratives that perpetuate existing inequities or stereotypes. Exploring and addressing localness biases is therefore essential. In the following section, we translate these observations into concrete design implications.

6.4 Design Implications for Localness-Aware AI Systems

Developing AI with localness-aware capabilities is increasingly critical. Applications now range from civic tech that must process community-specific regulations and discourse [33], to local journalism tools that interpret place-based narratives [61], to systems that help people decide

where to live, travel, or invest attention [9, 83]. Our results show that today's LLMs are poor foundations for such uses without redesign.

First, systems must make **coverage and provenance legible**. Interfaces should surface the source types that support a claim (e.g., formal news vs. social media) and explicitly warn when evidence is limited by data sparsity or degraded by high-noise modalities. This responds to transparency research arguing that provenance is essential for appropriate reliance [53, 57]. By exposing locality-sensitive gaps and disagreements, systems can shift from providing place-agnostic answers to helping users identify where the model is on firm versus shaky ground [68, 74]. While some localness inference techniques [50] exist that may be able to help address missing provenance and coverage details, these techniques have similar rural-modeling problems to our results here. Thus, establishing this data and these cues at the user-generated content platform level (e.g., on Reddit itself) may be a critical design direction.

Second, **systems should be designed as verification tools, not answer engines**. Because factual grounding and confidence are unreliable, localness applications should scaffold user triangulation across multiple perspectives, highlight inconsistencies, and communicate uncertainty through interaction and language rather than hidden scores. Evidence suggests that without explicit inconsistency cues or uncertainty wording, users tend to over-rely on fluent but incorrect outputs [52, 53]. For CSCW, this supports workflows in which users compare outputs, inspect evidence, and treat generations as drafts.

Finally, localness-aware AI should be grounded in **community-stewarded narratives**. Since LLMs inherit urban bias and other spatial inequalities, they can reproduce biased place narratives with persuasive fluency. Participatory AI research cautions that mitigating such harms requires shifting power in data and governance, not just improving models [82, 87]. Building on collective narrative grounding, we envision systems that retrieve from a provenance-visible, community-governed narrative layer: local stories are treated as first-class data, contributed with consent, validated by residents, and used to contextualize or contest model outputs [24]. This enables feedback loops where communities can flag misrepresentations and incrementally repair systematic errors, supporting more just digital representations of place.

6.5 Limitations and Future Work

While our methodological framework provides a robust foundation for evaluating LLM performance across geographic and localness contexts, several limitations point to promising avenues for future work.

Our benchmark operationalizes localness using the taxonomy proposed in a preprint by Gao et al. [21]. Although we find this framework both conceptually useful and practically enabling for large-scale evaluation, it is not yet a settled or community-validated standard. Alternative taxonomies might characterize local knowledge differently (e.g., emphasizing historical, affective, or institutional dimensions), which could shift how questions are categorized and how model gaps are interpreted. Our results are fundamentally conditional on this operationalization choice, and we see validating or comparing across localness frameworks as an important direction for CSCW work.

Our study's data scope also presents some boundaries to generalization. For instance, our Reddit dataset, focused on county-level subreddits, may not capture hyperlocal discourse found in city or neighborhood communities. Similarly, our use of census data is a necessary but imperfect proxy for complex socio-cultural identities, and rural and suburban areas remain underrepresented in our news and subreddit datasets. Future work should expand this framework to finer geographic granularities, incorporate qualitative data for a more holistic ground-truth, and broaden data sourcing to

create a more balanced and representative benchmark. Future research could also incorporate additional covariates such as population size, income, or education to better contextualize performance outcomes. Furthermore, causal inference methods could be applied to disentangle the mechanisms driving observed performance differences. Examining the influence of data source properties, such as the variability or timeliness of metrics, on model behavior also presents an important research direction. Lastly, while bootstrap resampling enhanced the robustness of our estimates, it remains computationally intensive. In future work, more efficient alternatives such as Bayesian hierarchical models for uncertainty estimation may offer scalable solutions.

7 Conclusion

LLMs are increasingly used to help people understand local communities, but our systematic evaluation reveals that their grasp of "localness" is shallow, factually weak, and biased. Across structured census data, local news, and social media, we found that performance is uneven, consistently favoring urban over rural settings and physical over relational or cognitive forms of local knowledge. These findings demonstrate that achieving equitable local representation requires moving beyond passive evaluation to active intervention. This work calls for a concerted effort from the CSCW community to build richer and more ethical datasets, design interfaces that prioritize user verification over blind trust, and develop AI systems that support deeper, more equitable engagement with place.

References

[1] Penelope Muse Abernathy. 2018. The expanding news desert. https://www.usnewsdeserts.com/reports/expanding-news-desert/. Accessed: 2025-11-24.

[2] Thales Sales Almeida, Giovana Kerche Bonás, João Guilherme Alves Santos, Hugo Abonizio, and Rodrigo Nogueira. 2025. TiEBe: Tracking Language Model Recall of Notable Worldwide Events Through Time. arXiv:2501.07482 [cs.CL] https://arxiv.org/abs/2501.07482

[3] Marianne Aubin Le Quéré, Mor Naaman, and Jenna Fields. 2024. Not Quite Filling the Void: Comparing the Perceptions of Local Online Groups and Local Media Pages on Facebook. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 100 (April 2024), 22 pages. doi:10.1145/3637377

[4] Michael Barthel, Galen Stocking, Jesse Holcomb, and Amy Mitchell. 2016. *Reddit News Users More Likely to Be Male, Young and Digital in Their News Preferences.* Technical Report. Pew Research Center, Journalism & Media. https://www.pewresearch.org/journalism/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/ Accessed: 2025-11-22.

[5] Cillian Berragan, Alex Singleton, Alessia Calafiore, and Jeremy Morley. 2024. Mapping Great Britain's semantic footprints through a large language model analysis of Reddit comments. *Computers, Environment and Urban Systems* 110 (2024), 102121. doi:10.1016/j.compenvurbsys.2024.102121

[6] Cody Blakeney, Mansheej Paul, Brett W. Larsen, Sean Owen, and Jonathan Frankle. 2024. Does your data spark joy? Performance gains from domain upsampling at the end of training. In *First Conference on Language Modeling*. https://openreview.net/forum?id=vwIIAot0ff

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[8] Apoorva Chauhan and Amanda L. Hughes. 2017. Providing Online Crisis Information: An Analysis of Official Sources during the 2014 Carlton Complex Wildfire. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3151–3162. doi:10.1145/3025453.3025627

[9] Aili Chen, Xuyang Ge, Ziquan Fu, Yanghua Xiao, and Jiangjie Chen. 2024. TravelAgent: An AI Assistant for Personalized Travel Planning. arXiv:2409.08069 [cs.AI] https://arxiv.org/abs/2409.08069

[10] Ashley Colley, Jacob Thebault-Spieker, Allen Yilun Lin, Donald Degraen, Benjamin Fischman, Jonna Häkkilä, Kate Kuehl, Valentina Nisi, Nuno Jardim Nunes, Nina Wenig, Dirk Wenig, Brent Hecht, and Johannes Schöning. 2017. The Geography of Pokémon GO: Beneficial and Problematic Effects on Places and Movement. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 1179–1192. doi:10.1145/3025453.3025495

[11] Justin Cranshaw, Andrés Monroy-Hernández, and S.A. Needham. 2016. Journeys & Notes: Designing Social Computing for Non-Places. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 4722–4733. doi:10.1145/2858036.2858573

[12] Justin Cranshaw, Raz Schwartz, Jason Hong, and Norman Sadeh. 2012. The livehoods project: Utilizing social media to understand the dynamics of a city. In *Proceedings of the international AAAI conference on web and social media*, Vol. 6. 58–65. https://doi.org/10.1609/icwsm.v6i1.14278

[13] Kate Crawford. 2021. *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence.* Yale University Press.

[14] Paul Dourish and Genevieve Bell. 2011. *Divining a digital future: Mess and mythology in ubiquitous computing.* Mit Press.

[15] Shiran Dudy, Thulasi Tholeti, Resmi Ramachandranpillai, Muhammad Ali, Toby Jia-Jun Li, and Ricardo Baeza-Yates. 2025. Unequal Opportunities: Examining the Bias in Geographical Recommendations by Large Language Models. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. Association for Computing Machinery, New York, NY, USA, 1499–1516. doi:10.1145/3708359.3712111

[16] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature* 630, 8017 (2024), 625–630. doi:10.1038/s41586-024-07421-0

[17] Casey Fiesler, Michael Zimmer, Nicholas Proferes, Sarah Gilbert, and Naiyan Jones. 2024. Remember the Human: A Systematic Review of Ethical Considerations in Reddit Research. *Proc. ACM Hum.-Comput. Interact.* 8, GROUP, Article 5 (Feb. 2024), 33 pages. doi:10.1145/3633070

[18] A Stewart Fotheringham and David WS Wong. 1991. The Modifiable Areal Unit Problem in Multivariate Statistical Analysis. *Environment and Planning A: Economy and Space* 23, 7 (1991), 1025–1044. doi:10.1068/a231025 arXiv:https://doi.org/10.1068/a231025

[19] Devin Gaffney and J. Nathan Matias. 2018. Caveat Emptor, Computational Social Science: Large-scale Missing Data in a Widely-published Reddit Corpus. *PLOS ONE* 13, 7 (07 2018), 1–13. doi:10.1371/journal.pone.0200162

[20] Zihan Gao, Justin Cranshaw, and Jacob Thebault-Spieker. 2024. Journeying Through Sense of Place with Mental Maps: Characterizing Changing Spatial Understanding and Sense of Place During Migration for Work. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 503 (Nov. 2024), 31 pages. doi:10.1145/3687042

[21] Zihan Gao, Justin Cranshaw, and Jacob Thebault-Spieker. 2025. A Turing Test for "Localness": Conceptualizing, Defining, and Recognizing Localness in People and Machines. arXiv:2505.07282 [cs.HC] https://arxiv.org/abs/2505.07282

[22] Zihan Gao, Jiaying Liu, Yifei Xu, and Jacob Thebault-Spieker. 2025. From Clips to Communities: Fusing Social Video into Knowledge Graphs for Localness-Aware LLMs. In *Companion Publication of the 2025 Conference on Computer-Supported Cooperative Work and Social Computing (CSCW Companion '25)*. Association for Computing Machinery, New York, NY, USA, 497–503. doi:10.1145/3715070.3749277

[23] Zihan Gao, Yifei Xu, and Jacob Thebault-Spieker. 2026. LocalBench: Benchmarking LLMs on County-Level Local Knowledge and Reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence* (2026).

[24] Zihan Gao, Mohsin Y. K. Yousufi, and Jacob Thebault-Spieker. 2025. Collective Narrative Grounding: Community-Coordinated Data Contributions to Improve Local AI Systems. In *NeurIPS 2025 Workshop on Algorithmic Collective Action*. https://openreview.net/forum?id=2ZRwKlGSDa

[25] Andrew Garbett, Rob Comber, Edward Jenkins, and Patrick Olivier. 2016. App Movement: A Platform for Community Commissioning of Mobile Applications. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 26–37. doi:10.1145/2858036.2858094

[26] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (Nov. 2021), 86–92. doi:10.1145/3458723

[27] Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A Survey of Confidence Estimation and Calibration in Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 6577–6595. doi:10.18653/v1/2024.naacl-long.366

[28] Thomas F. Gieryn. 2000. A Space for Place in Sociology. *Annual Review of Sociology* 26, Volume 26, 2000 (2000), 463–496. doi:10.1146/annurev.soc.26.1.463

[29] Martyna Gliniecka. 2023. The Ethics of Publicly Available Data Research: A Situated Ethics Framework for Reddit. *Social Media + Society* 9, 3 (2023), 20563051231192021. doi:10.1177/20563051231192021 arXiv:https://doi.org/10.1177/20563051231192021

[30] Sarah E. Gollust, Laura M. Baum, Jeff Niederdeppe, Colleen L. Barry, and Erika Franklin Fowler. 2017. Local Television News Coverage of the Affordable Care Act: Emphasizing Politics Over Consumer Information. *American Journal of Public Health* 107, 5 (2017), 687–693. doi:10.2105/AJPH.2017.303659 arXiv:https://doi.org/10.2105/AJPH.2017.303659 PMID: 28207336.

[31] Mark Graham and Martin Dittus. 2022. *Geographies of digital exclusion: Data and inequality.* Pluto Books.

[32] Atharva Gundawar, Mudit Verma, Lin Guan, Karthik Valmeekam, Siddhant Bhambri, and Subbarao Kambhampati. 2024. Robust Planning with LLM-Modulo Framework: Case Study in Travel Planning. arXiv:2405.20625 [cs.AI] https://arxiv.org/abs/2405.20625

[33] Jose A. Guridi, Cristobal Cheyre, and Qian Yang. 2025. Thoughtful Adoption of NLP for Civic Participation: Understanding Differences Among Policymakers. *Proc. ACM Hum.-Comput. Interact.* 9, 2, Article CSCW193 (May 2025), 27 pages. doi:10.1145/3711091

[34] Wes Gurnee and Max Tegmark. 2024. Language Models Represent Space and Time. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=jE8xbmvFin

[35] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 8342–8360. doi:10.18653/v1/2020.acl-main.740

[36] Keith Hampton and Barry Wellman. 2003. Neighboring in Netville: How the Internet Supports Community and Social Capital in a Wired Suburb. *City & Community* 2, 4 (2003), 277–311. doi:10.1046/j.1535-6841.2003.00057.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1535-6841.2003.00057.x

[37] Jean Hardy. 2019. How the Design of Social Technology Fails Rural America. In *Companion Publication of the 2019 on Designing Interactive Systems Conference 2019 Companion* (San Diego, CA, USA) *(DIS '19 Companion)*. Association for Computing Machinery, New York, NY, USA, 189–193. doi:10.1145/3301019.3323906

[38] Steve Harrison and Paul Dourish. 1996. Re-place-ing space: the roles of place and space in collaborative systems. In *Proceedings of the 1996 ACM Conference on Computer Supported Cooperative Work* (Boston, Massachusetts, USA) *(CSCW '96)*. Association for Computing Machinery, New York, NY, USA, 67–76. doi:10.1145/240080.240193

[39] Brent Hecht and Monica Stephens. 2014. A Tale of Cities: Urban Biases in Volunteered Geographic Information. *Proceedings of the International AAAI Conference on Web and Social Media* 8, 1 (May 2014), 197–205. doi:10.1609/icwsm.v8i1.14554

[40] Brent J. Hecht and Darren Gergle. 2010. On the "localness" of user-generated content. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work* (Savannah, Georgia, USA) *(CSCW '10)*. Association for Computing Machinery, New York, NY, USA, 229–232. doi:10.1145/1718918.1718962

[41] Peter Henderson, Mark Krass, Lucia Zheng, Neel Guha, Christopher D Manning, Dan Jurafsky, and Daniel Ho. 2022. Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 29217–29234. https://proceedings.neurips.cc/paper_files/paper/2022/file/bc218a0c656e49d4b086975a9c785f47-Paper-Datasets_and_Benchmarks.pdf

[42] Benjamin D. Horne, Maurício Gruppi, Kenneth Joseph, Jon Green, John P. Wihbey, and Sibel Adalı. 2022. NELA-Local: A Dataset of U.S. Local News Articles for the Study of County-Level News Ecosystems. *Proceedings of the International AAAI Conference on Web and Social Media* 16, 1 (May 2022), 1275–1284. doi:10.1609/icwsm.v16i1.19379

[43] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* 43, 2, Article 42 (Jan. 2025), 55 pages. doi:10.1145/3703155

[44] Kee Moon Jang, Junda Chen, Yuhao Kang, Junghwan Kim, Jinhyung Lee, Fabio Duarte, and Carlo Ratti. 2024. Place identity: a generative AI's perspective. *Humanities and Social Sciences Communications* 11, 1 (2024), 1156. doi:10.1057/s41599-024-03645-7

[45] Kee Moon Jang and Youngchul Kim. 2019. Crowd-sourced cognitive mapping: A new way of displaying people's cognitive perception of urban space. *PLOS ONE* 14, 6 (06 2019), 1–18. doi:10.1371/journal.pone.0218590

[46] Andrew Jenkins, Arie Croitoru, Andrew T. Crooks, and Anthony Stefanidis. 2016. Crowdsourcing a Collective Sense of Place. *PLOS ONE* 11, 4 (04 2016), 1–20. doi:10.1371/journal.pone.0152932

[47] Isaac L. Johnson, Yilun Lin, Toby Jia-Jun Li, Andrew Hall, Aaron Halfaker, Johannes Schöning, and Brent Hecht. 2016. Not at Home on the Range: Peer Production and the Urban/Rural Divide. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery,

New York, NY, USA, 13–25. doi:10.1145/2858036.2858123

[48] Isaac L. Johnson, Subhasree Sengupta, Johannes Schöning, and Brent Hecht. 2016. The Geography and Importance of Localness in Geotagged Social Media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 515–526. doi:10.1145/2858036.2858122

[49] Bradley S. Jorgensen and Richard C. Stedman. 2001. Sense of Place as An Attitude: Lakehouse Owners Attitudes Toward Their Properties. *Journal of Environmental Psychology* 21, 3 (2001), 233–248. doi:10.1006/jevp.2001.0226

[50] Ankit Kariryaa, Isaac Johnson, Johannes Schöning, and Brent Hecht. 2018. Defining and Predicting the Localness of Volunteered Geographic Information using Ground Truth Data. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3173574.3173839

[51] Gali Katz, Hai Sitton, Guy Gonen, and Yohay Kaplan. 2025. Beyond the Surface: Uncovering Implicit Locations with LLMs for Personalized Local News. arXiv:2502.14660 [cs.LG] https://arxiv.org/abs/2502.14660

[52] Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. "I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 822–835. doi:10.1145/3630106.3658941

[53] Sunnie S. Y. Kim, Jennifer Wortman Vaughan, Q. Vera Liao, Tania Lombrozo, and Olga Russakovsky. 2025. Fostering Appropriate Reliance on Large Language Models: The Role of Explanations, Sources, and Inconsistencies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 420, 19 pages. doi:10.1145/3706598.3714020

[54] Jixuan Leng, Chengsong Huang, Banghua Zhu, and Jiaxin Huang. 2025. Taming Overconfidence in LLMs: Reward Calibration in RLHF. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=l0tg0jzsdL

[55] Charis Lengen and Thomas Kistemann. 2012. Sense of place and place identity: Review of neuroscientific evidence. *Health Place* 18, 5 (2012), 1162–1171. doi:10.1016/j.healthplace.2012.01.012

[56] Laura Lentini and Françoise Decortis. 2010. Space and places: when interacting with and in physical space becomes a meaningful experience. *Personal and Ubiquitous Computing* 14, 5 (2010), 407–415. doi:10.1007/s00779-009-0267-y

[57] Q. Vera Liao and Jennifer Wortman Vaughan. 2023. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. arXiv:2306.01941 [cs.HC] https://arxiv.org/abs/2306.01941

[58] Hongfu Liu, Hengguan Huang, Xiangming Gu, Hao Wang, and Ye Wang. 2025. On Calibration of LLM-based Guard Models for Reliable Content Moderation. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=wUbum0nd9N

[59] Tianyang Liu, Fei Wang, and Muhao Chen. 2024. Rethinking Tabular Data Understanding with Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 450–482. doi:10.18653/v1/2024.naacl-long.26

[60] Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, Xinyi Wu, Enrico Shippole, Kurt Bollacker, Tongshuang Wu, Luis Villa, Sandy Pentland, and Sara Hooker. 2023. The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI. http://arxiv.org/abs/2310.16787 arXiv:2310.16787 [cs].

[61] Matt MacVey. 2022. AI & Local News newsletter, issue 12. NYU Tandon School of Engineering, NYC Media Lab. https://engineering.nyu.edu/news/ai-local-news-newsletter-issue-12 Accessed: 2025-07-16.

[62] Momin Malik, Hemank Lamba, Constantine Nakos, and Jürgen Pfeffer. 2021. Population Bias in Geotagged Tweets. *Proceedings of the International AAAI Conference on Web and Social Media* 9, 4 (Aug. 2021), 18–27. doi:10.1609/icwsm.v9i4.14688

[63] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 9802–9822. doi:10.18653/v1/2023.acl-long.546

[64] Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, and Stefano Ermon. 2024. Large Language Models are Geographically Biased. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) *(ICML'24)*. JMLR.org, Article 1409, 16 pages.

[65] Rohin Manvi, Samar Khanna, Gengchen Mai, Marshall Burke, David Lobell, and Stefano Ermon. 2024. GeoLLM: Extracting Geospatial Knowledge from Large Language Models. In *ICLR*. https://openreview.net/forum?id=TqL2xBwXP3

[66] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6, Article 115 (July 2021), 35 pages. doi:10.1145/3457607

[67] Mazda Moayeri, Elham Tabassi, and Soheil Feizi. 2024. WorldBench: Quantifying Geographic Disparities in LLM Factual Recall. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1211–1228. doi:10.1145/3630106.3658967

[68] Stein Monteiro. 2024. Searching for Settlement Information on Reddit. *International Migration* 62, 3 (2024), 100–119. doi:10.1111/imig.13261 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/imig.13261

[69] Hang Ni, Fan Liu, Xinyu Ma, Lixin Su, Shuaiqiang Wang, Dawei Yin, Hui Xiong, and Hao Liu. 2025. TP-RAG: Benchmarking Retrieval-Augmented Large Language Model Agents for Spatiotemporal-Aware Travel Planning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 12403–12429. doi:10.18653/v1/2025.emnlp-main.626

[70] Helen Nissenbaum. 2011. A Contextual Approach to Privacy Online. *Daedalus* 140, 4 (2011), 32–48. https://doi.org/10.1162/DAED_a_00113

[71] Safiya Umoja Noble. 2018. Algorithms of Oppression: How Search Engines Reinforce Racism. In *Algorithms of oppression*. New York university press.

[72] Will Orr and Kate Crawford. 2024. Building Better Datasets: Seven Recommendations for Responsible Design from Dataset Creators. arXiv:2409.00252 [cs.LG] https://arxiv.org/abs/2409.00252

[73] Bo Pang, Tingrui Qiao, Caroline Walker, Chris Cunningham, and Yun Sing Koh. 2025. LIBRA: Measuring Bias of Large Language Model from a Local Context. In *Advances in Information Retrieval*, Claudia Hauff, Craig Macdonald, Dietmar Jannach, Gabriella Kazai, Franco Maria Nardini, Fabio Pinelli, Fabrizio Silvestri, and Nicola Tonellotto (Eds.). Springer Nature Switzerland, Cham, 1–16.

[74] Sangkeun Park, Yongsung Kim, Uichin Lee, and Mark Ackerman. 2014. Understanding localness of knowledge sharing: a study of Naver KiN 'here'. In *Proceedings of the 16th International Conference on Human-Computer Interaction with Mobile Devices & Services* (Toronto, ON, Canada) *(MobileHCI '14)*. Association for Computing Machinery, New York, NY, USA, 13–22. doi:10.1145/2628363.2628407

[75] Nicholas Proferes, Naiyan Jones, Sarah Gilbert, Casey Fiesler, and Michael Zimmer. 2021. Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics. *Social Media + Society* 7, 2 (2021), 20563051211019004. doi:10.1177/20563051211019004 arXiv:https://doi.org/10.1177/20563051211019004

[76] Yao Qu and Jue Wang. 2024. Performance and Biases of Large Language Models in Public Opinion Simulation. *Humanities and Social Sciences Communications* 11, 1 (2024), 1095. doi:10.1057/s41599-024-03609-x

[77] Edward Relph. 1976. *Place and placelessness*. Vol. 67. Pion London.

[78] J Riley and H Cowart. 2021. The Reddit Oasis: Analyzing the potential role of location-based subreddits in the alleviation of news deserts. *Community Journalism* 9, 1 (2021).

[79] Emily Sun. 2021. The Importance of Play in Digital Placemaking. *Proceedings of the International AAAI Conference on Web and Social Media* 9, 2 (Aug. 2021), 23–25. doi:10.1609/icwsm.v9i2.14680

[80] Emily Sun and Mor Naaman. 2018. A Multi-site Investigation of Community Awareness Through Passive Location Sharing. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3173574.3174155

[81] Kai Sun, Yifan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2024. Head-to-Tail: How Knowledgeable are Large Language Models (LLMs)? A.K.A. Will LLMs Replace Knowledge Graphs?. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 311–325. doi:10.18653/v1/2024.naacl-long.18

[82] Harini Suresh, Emily Tseng, Meg Young, Mary Gray, Emma Pierson, and Karen Levy. 2024. Participation in the age of foundation models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1609–1621. doi:10.1145/3630106.3658992

[83] Yihong Tang, Zhaokai Wang, Ao Qu, Yihao Yan, Zhaofeng Wu, Dingyi Zhuang, Jushi Kai, Kebing Hou, Xiaotong Guo, Jinhua Zhao, Zhan Zhao, and Wei Ma. 2024. ItiNera: Integrating Spatial Optimization with Large Language Models for Open-domain Urban Itinerary Planning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, Franck Dernoncourt, Daniel Preoţiuc-Pietro, and Anastasia Shimorina (Eds.). Association for Computational Linguistics, Miami, Florida, US, 1413–1432. doi:10.18653/v1/2024.emnlp-industry.104

[84] Alex S. Taylor, Siân Lindley, Tim Regan, David Sweeney, Vasillis Vlachokyriakos, Lillie Grainger, and Jessica Lingel. 2015. Data-in-Place: Thinking through the Relations Between Data and Community. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 2863–2872. doi:10.1145/2702123.2702558

[85] Jacob Thebault-Spieker, Aaron Halfaker, Loren G. Terveen, and Brent Hecht. 2018. Distance and Attraction: Gravity Models for Geographic Content Production. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3173574.3173722

[86] Jacob Thebault-Spieker, Brent Hecht, and Loren Terveen. 2018. Geographic Biases are 'Born, not Made': Exploring Contributors' Spatiotemporal Behavior in OpenStreetMap. In *Proceedings of the 2018 ACM International Conference on Supporting Group Work* (Sanibel Island, Florida, USA) *(GROUP '18)*. Association for Computing Machinery, New York, NY, USA, 71–82. doi:10.1145/3148330.3148350

[87] Emily Tseng, Rosanna Bellini, Yeuk-Yu Lee, Alana Ramjit, Thomas Ristenpart, and Nicola Dell. 2024. Data Stewardship in Clinical Computer Security: Balancing Benefit and Burden in Participatory Systems. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 39 (April 2024), 29 pages. doi:10.1145/3637316

[88] Yi-Fu Tuan. 1977. *Space and place: The perspective of experience.* U of Minnesota Press.

[89] Vasilis Vlachokyriakos, Rob Comber, Karim Ladha, Nick Taylor, Paul Dunphy, Patrick McCorry, and Patrick Olivier. 2014. PosterVote: Expanding the Action Repertoire for Local Political Activism. In *Proceedings of the 2014 Conference on Designing Interactive Systems* (Vancouver, BC, Canada) *(DIS '14)*. Association for Computing Machinery, New York, NY, USA, 795–804. doi:10.1145/2598510.2598523

[90] Yifei Xu, Tusher Chakraborty, Emre Kiciman, Bibek Aryal, Srinagesh Sharma, Songwu Lu, and Ranveer Chandra. 2025. RLTHF: Targeted Human Feedback for LLM Alignment. In *Forty-second International Conference on Machine Learning*. https://openreview.net/forum?id=ATUfSZayVo

[91] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open Pre-trained Transformer Language Models. arXiv:2205.01068 [cs.CL] https://arxiv.org/abs/2205.01068

[92] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate Before Use: Improving Few-shot Performance of Language Models. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 12697–12706. https://proceedings.mlr.press/v139/zhao21c.html

[93] Shucheng Zhu, Weikang Wang, and Ying Liu. 2024. Quite Good, but Not Enough: Nationality Bias in Large Language Models - a Case Study of ChatGPT. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 13489–13502. https://aclanthology.org/2024.lrec-main.1180/

A Census Metrics, Localness Characteristics and Sources

Table 12. Localness metric taxonomy with subcomponents, metric definitions, and data sources.

Domain	Dimension	Component	Subcomponent	Metrics	Data Source	Why this metric captures the sub-component
Cognitive	Cultural	Food Culture	Knowledge of local cuisine	In 2018, the number of nonemployer establishments in accommodation and food services per 1,000 residents in this county	U.S. Census Bureau Nonemployer Statistics Table NS1800NONEMP 2018	Many small eateries and food trucks embody distinct local flavors that residents can recommend.
Cognitive	Cultural	Language/Dialect	Knowledge of local expressions	In 2022, the number of residents in this county who spoke a language other than English at home	U.S. Census Bureau ACS Table DP02	A pool of non-English speakers fosters multilingual expressions unique to the locale.
Cognitive	Cultural	Language/Dialect	Understanding regional accents	In 2022, the number of residents in this county who spoke English less than 'very well' at home	U.S. Census Bureau ACS Table DP02	Frequent limited-English speakers require locals to interpret diverse accents/dialects.
Cognitive	Cultural	Local Customs/Norms	Understanding community values	In 2020, the percentage of Southern Baptist Convention adherents among total adherents in this county	US Religion Census 2020 Group detail data by nation, state, county and metro	Knowing dominant faith traditions reveals awareness of prevailing moral and cultural norms.
Cognitive	Environmental	Ecological Understanding	Understanding of environmental issues	In 2018, the number of nonemployer establishments in mining, quarrying, and oil and gas extraction per 1,000 residents in this county	U.S. Census Bureau Nonemployer Statistics Table NS1800NONEMP 2018	Extractive industries heighten community debates on environmental impact and resource use.
Cognitive	Environmental	Geographic Familiarity	Understanding of land features	In 2022, the percentage of cropland fertilized in this county	USDA National Agricultural Statistics Service	Agriculture mix reflects terrain knowledge—soil types, watershed, and land-use patterns.
Cognitive	Knowledge	Change Awareness	Understanding of ongoing transformations	The change in multifamily building permits from 2021 to 2022 in this county	U.S. Census Bureau Building Permits Survey	Permit shifts track real-time development, signalling knowledge of rapid urban change.
Cognitive	Knowledge	Hidden Gems	Access to insider information	In 2018, the number of nonemployer establishments in information industries per 1,000 residents in this county	U.S. Census Bureau Nonemployer Statistics Table NS1800NONEMP 2018	Local media, data, and publishing freelancers curate insider tips and niche knowledge.
Cognitive	Knowledge	Historical Knowledge	Awareness of local history	As of 2024, the number of historic preservation properties with local significance in this county	National Register of Historic Places	Registered historic sites supply tangible cues and stories about the area's past.
Cognitive	Knowledge	Local Recommendations	Ability to make informed recommendations	In 2018, the number of nonemployer establishments in professional, scientific, and technical services per 1,000 residents in this county	U.S. Census Bureau Nonemployer Statistics Table NS1800NONEMP 2018	A dense advice-oriented industry (consultants, tech, R&D) signals a rich pool for expert tips.

Continued on next page

Table 12. Localness metric taxonomy with subcomponents, metric definitions, and data sources.

Domain	Dimension	Component	Subcomponent	Metrics	Data Source	Why this metric captures the sub-component
Cognitive	Knowledge	Local Recommendations	Awareness of local options and alternatives	In 2018, the number of nonemployer establishments in educational services per 1,000 residents in this county	U.S. Census Bureau Nonemployer Statistics Table NS1800NONEMP 2018	Many tutoring/training micro-firms indicate plentiful—and knowable—alternative learning options.
Cognitive	Knowledge	Local Recommendations	Understanding local services and amenities	In 2018, the number of nonemployer establishments in administrative, support, and waste management and remediation services per 1,000 residents in this county	U.S. Census Bureau Nonemployer Statistics Table NS1800NONEMP 2018	High counts of everyday-service micro-firms (cleaning, hauling) make local amenity knowledge essential.
Cognitive	Knowledge	Local Recommendations	Understanding local services and amenities	In 2022, the number of public libraries in this county	Public Libraries Survey (PLS) by the Institute of Museum and Library Services (IMLS)	Libraries are civic hubs; knowing their locations/services marks amenity awareness.
Cognitive	Knowledge	Navigation/Wayfinding	Familiarity with transportation system	In 2022, the mean travel time to work for residents of this county	U.S. Census Bureau ACS Table S0801	Commute length reflects awareness of transit options, traffic patterns, and route choices.
Cognitive	Knowledge	Navigation/Wayfinding	Familiarity with transportation system	In 2022, the percentage of workers who use public transportation to work in this county	U.S. Census Bureau ACS Table S0801	Knowing routes and schedules defines how well one grasps collective mobility options.
Physical	EnvironmentalPhy	Geographic Familiarity	Navigation proficiency	In 2022, the percentage of employed residents living in this county who worked within their county of residence	U.S. Census Bureau ACS Table S0801	Living and working in the same county forces routine, ground-level navigation of local routes.
Physical	EnvironmentalPhy	Natural Environment	Connection to natural spaces	In 2018, the number of nonemployer establishments in agriculture, forestry, fishing, and hunting per 1,000 residents in this county	U.S. Census Bureau Nonemployer Statistics Table NS1800NONEMP 2018	Local, land-based livelihoods require day-to-day ecological awareness.
Physical	Temporal	Being Born/Native	Automatic claim to localness	In 2022, the number of residents in this county who were born in the United States and in their state of residence	U.S. Census Bureau ACS Table DP02	Being born in-state affords an immediate, almost unquestioned claim of "being from here."
Physical	Temporal	Being Born/Native	Deep historical connection	In 2022, the percentage of the population in this county identifying as Native American	U.S. Census Bureau ACS Table DP05	Indigenous presence links the county to centuries-long, place-based histories.
Physical	Temporal	Formative Years	Deep cultural absorption	In 2018, the number of nonemployer establishments in arts, entertainment, and recreation per 1,000 residents in this county	U.S. Census Bureau Nonemployer Statistics Table NS1800NONEMP 2018	A vibrant creative micro-scene suggests residents are steeped in local cultural practices.
Physical	Temporal	Long-term Residence	Accumulated experience	In 2022, the number of residents in this county who had lived in the same house or apartment for more than five years	U.S. Census Bureau ACS Table S0701	Living in the same home for years lets people accumulate first-hand, continuous experience of the place.

Continued on next page

Table 12. Localness metric taxonomy with subcomponents, metric definitions, and data sources.

Domain	Dimension	Component	Subcomponent	Metrics	Data Source	Why this metric captures the sub-component
Physical	Temporal	Long-term Residence	Investment in place	In 2022, the median move-in year of householders in owner-occupied units in this county	U.S. Census Bureau ACS Table B25039	Earlier move-in years imply owners have sunk time and capital into the property and locale.
Physical	Temporal	Long-term Residence	Investment in place	In 2022, the percentage of occupied housing units that were owner-occupied in this county	U.S. Census Bureau ACS Table DP04	Homeownership denotes financial stake and physical commitment to the locality.
Physical	Temporal	Long-term Residence	Witnessing area changes	In 2022, the number of native residents in this county who moved to their current residence before 2010	U.S. Census Bureau ACS Table DP02	Early arrival means residents have personally observed multiple waves of neighborhood change.
Relational	Emotional	Identity Connection	Identification with local character	In 2022, the number of residents in this county who reported 'American' ancestry	U.S. Census Bureau ACS Table DP02	Claiming generic "American" roots often signals multi-generation ties to the locale.
Relational	Emotional	Identity Connection	Shared identity with community	In 2022, the percentage of residents in this county who were Hispanic or Latino (of any race)	U.S. Census Bureau ACS Table DP05	A sizable ethnic bloc shapes shared festivals, media, and identity references in daily life.
Relational	Emotional	Identity Connection	Shared identity with community	In 2022, the ethnolinguistic fractionalization index of residents in this county	U.S. Census Bureau ACS Table DP05	High diversity means locals regularly negotiate multiple identities and inclusive norms.
Relational	Emotional	Sense of Belonging	Sense of rightful presence	In 2022, the percentage of residents in this county with zero components of social vulnerability	Community Resilience Estimates Datasets	Low vulnerability scores associate with feeling secure, entitled, and "belonging" in local life.
Relational	Social/Community	Active Participation	Volunteering in community	In 2018, the number of nonemployer establishments in health care and social assistance per 1,000 residents in this county	U.S. Census Bureau Nonemployer Statistics Table NS1800NONEMP 2018	A big pool of caregiving micro-providers correlates with a culture of service and volunteering.
Relational	Social/Community	Civic Engagement	Voting in local elections	In the 2020 presidential election, the total number of votes cast in this county	County Presidential Election Returns 2000-2020	Casting ballots shows direct participation in civic affairs, a hallmark of local engagement.
Relational	Social/Community	Community Investment	Long-term commitment	In 2022, the percentage of owner-occupied housing units with a mortgage in this county	ACS Table DP05	Paying a mortgage stretches investment and commitment far into the future.
Relational	Social/Community	Community Investment	Supporting local businesses	In 2018, the density of nonemployer businesses per 1,000 residents in this county	U.S. Census Bureau Nonemployer Statistics Table NS1800NONEMP 2018	Many self-employed ventures indicate a community that values and patronizes local enterprise.
Relational	Social/Community	Personal Relationships	Building social networks	In 2022, the average size of married-couple households in this county	U.S. Census Bureau ACS Table DP02	Larger married households extend kinship webs, enlarging one's default social network locally.
Relational	Social/Community	Personal Relationships	Regular social interactions	In 2022, the average household size in this county	U.S. Census Bureau ACS DP04	Bigger households create more daily in-home encounters, fostering habitual local interaction.

B QCA Construction and Quality Control Details

This appendix provides additional technical details on our Question-Context-Answer (QCA) dataset construction, including question generation templates, automated quality control metrics, and the reward model implementation.

B.1 Census Question Generation Templates

For each census metric m and county l, we generated three initial question types, from which two were selected for the final dataset. Let $v(m, l)$ be the value of metric m for county l.

- **Type 1 (Fill-in-the-blank):** A direct question asking for the value $v(m, l)$. For example: "What is the value of [metric description] in [county name]?"
- **Type 2 (Higher Comparison):** A true/false question comparing $v(m, l)$ to a county l' where $v(m, l') > v(m, l)$. For example: "True or False: The value for [metric description] in [county name] is lower than in [county name l']."
- **Type 3 (Lower Comparison):** A true/false question comparing $v(m, l)$ to a county l'' where $v(m, l'') < v(m, l)$. For example: "True or False: The value for [metric description] in [county name] is higher than in [county name l'']."

To create a balanced and computationally manageable dataset, we included Type 1 for all metrics and randomly selected either Type 2 or Type 3 for the comparison question.

B.2 Census Questions Examples

Table 13. Census QA Data Question Types and Distribution

Question Type	Question Format	Example	Count
Fill-in-the-Blank (Q1)	{Metric} of {county, state} is []	As of 2024, the number of historic preservation properties with local significance in Lincoln, OK is [].	3570
Higher Comparison (Q2)	{Metric} of {county_1, state_1} is higher than that of {county_2, state_2}	As of 2024, the number of historic preservation properties with local significance in Lincoln, OK is higher than that of Jefferson, PA	1785
Lower Comparison (Q3)	{Metric} of {county_1, state_1} is lower than that of {county_2, state_2}	As of 2024, the number of historic preservation properties with local significance in Lincoln, OK is higher than that of Fort Bend, TX	1785

B.3 Local News Articles QCA Generation Prompt

```
Task: Generate a Question-Context-Answer pair from this local news article, ensuring the
     combination of question and context provides complete identifying information:

- Article Title: {title}
- Article Information:
- Source: {source}
- Date: {date}
- Location: {county}, {state}
- Available Factual Sentences: {sentences}

Requirements:
1. Question and answer must focus on specific local information described in the article,
     ensuring clarity about time, location, and individuals or organizations involved.
2. For this pair:
    1. Create a question that is precise, locally relevant, and includes or is supported by
     details like exact dates, full names and titles, and specific locations (city (if
     available), county, state).
```

```
    2. Write a context that provides neutral, clear, and concise background information. The
     context must NOT include or hint at the answer but should make the answer the only possible
     one by clarifying when, who, and where.
    3. Provide a concise, fact-based answer that is specific to the question and relies on the
     selected sentences.
3. Questions must be specific and include enough detail to make the answer unique and unambiguous.
4. Think carefully about each step:
    1. Identify key facts in the article that are specific to local information and could form
     meaningful questions.
    2. Write neutral contexts that clarify complete details about when, who, and where, ensuring
     the answer is the only possible one.
    3. Formulate clear, specific questions and concise, correct answers using selected sentences.

Response Format:
[PAIR1]
Question: (question)
Context: (neutral background information only, no answer or hints)
Answer: (unique and correct answer)
Selected Sentences: (list sentence numbers used)
```

B.4 Local Subreddits QCA Generation Prompt

```
Task: Generate a Question-Context-Answer pair from a Reddit post on a local county subreddit,
     ensuring the combination of question and context provides complete identifying information.

- Post Title: {title}
- Post Information:
    - Date: {date}
    - Location: {county}, {state}
- Post Content: {post_content}
- Comments Section: {comments_section}

Requirements:
1. Question and answer must focus on specific local information discussed in the post and
     comments, ensuring clarity about time, location, and individuals or organizations involved.
2. For this pair:
    - Create a question based on the post title and content (or comments, if necessary) that is
     precise, locally relevant, and includes or is supported by details like exact dates, full
     names and titles, and specific locations (city (if available), county, state).
    - Write a context that provides neutral, clear, and concise background information. The
     context must NOT include or hint at the answer but should make the answer the only possible
     one by clarifying when, who, and where.
    - Provide a concise, fact-based answer that is specific to the question and relies on the
     comments.
3. Questions must be specific and include enough detail to make the answer unique and unambiguous.
4. Avoid any questions that focus on discussions about the Reddit post itself (e.g., accusations
     of spam, user opinions about the post, or debates about online behavior). The question must
     focus on the actual local subject matter being discussed. Avoid mentioning Reddit in the
     context or answer.

Think carefully about each step:
1. Identify key facts in the post and comments that are specific to local information and could
     form meaningful questions.
2. Write neutral contexts that clarify complete details about when, who, and where, ensuring the
     answer is the only possible one.
3. Formulate clear, specific questions and concise, correct answers using selected sentences.
```

```
Response Format:
[PAIR1]
Question: (Insert the question here)
Context: (Insert the neutral background information here, without revealing the answer. Must
     include date.)
Answer: (Insert the fact-based answer here)
Selected Comments: (List comment numbers used, e.g., 1, 2, 4)
```

B.5 Automated Quality Control Metrics

Our QualityCheck function evaluated each generated QCA triplet (q, c, a) from a source document d using the following metrics to detect information leakage and ensure relevance.

- **Leakage Check:** We measured entity and n-gram phrase overlap between the question q and both the context c and the full source document d. A high overlap suggests the question may be too easily answerable via simple pattern matching.

$$\text{Leakage}(q, d) = \text{EntityOverlap}(q, d) \cup \text{PhraseOverlap}(q, d) \tag{1}$$

$$\text{Leakage}(q, c) = \text{EntityOverlap}(q, c) \cup \text{PhraseOverlap}(q, c) \tag{2}$$

- **Lexical Overlap:** We calculated the Jaccard similarity between the token sets of the question and context to ensure they were distinct.

$$\text{LexicalOverlap}(q, c) = \frac{|\text{tokens}(q) \cap \text{tokens}(c)|}{|\text{tokens}(q) \cup \text{tokens}(c)|} \tag{3}$$

- **Local Entity Coverage:** We verified that the question was sufficiently local by measuring the ratio of local entities (e.g., specific streets, local figures) to all entities mentioned.

$$\text{EntityCoverage}(q) = \frac{|\text{LocalEntities}(q)|}{|\text{AllEntities}(q)|} \tag{4}$$

Triplets that exceeded predefined thresholds for these metrics were flagged and used as negative examples in our iterative refinement loop.

B.6 QCA Generation, Quality Check, and Few-Shot Refinement

B.7 Reward Model Training and Refinement Details

We trained a reward model to capture subjective QCA quality dimensions that heuristics cannot reliably detect (e.g., clarity, naturalness, ambiguity). After heuristic filtering, we sampled 1,220 documents and extracted two candidate QCA triplets per document. The first author annotated each pair as a preference tuple (q_w, q_l) (winner vs. loser), producing dataset A.

Training objective. We trained a reward model R_θ using the Bradley–Terry loss:

$$\mathcal{L}_{\text{BT}} = -\frac{1}{|A|} \sum_{(q_w, q_l) \in A} \log\left(\sigma(R_\theta(q_w) - R_\theta(q_l))\right),$$

where σ is the logistic sigmoid.

Model and hyperparameters. The reward model used RoBERTa-base with a linear scalar output head. We trained for 3 epochs with learning rate 2×10^{-5}, achieving 83.5% preference-prediction accuracy.

Algorithm 1 QCA Generation, Quality Check, and Few-Shot Refinement

Require: Dataset $\mathcal{D}$ of local news articles or subreddit posts
Require: Attempt limit $L \leftarrow 3$
Ensure: Final set Q of high-quality *(Question, Context, Answer)* triples

$Q \leftarrow \varnothing$
for all $d \in \mathcal{D}$ **do** ▷ iterate over each article or post
  $tries \leftarrow 0$
  $Q_d \leftarrow \varnothing$
  while $|Q_d| < 2$ **and** $tries < L$ **do**
    $tries \leftarrow tries + 1$
    $(q, c, a) \leftarrow \textsc{GenerateQCA}(d)$
    $(passed, reason) \leftarrow \textsc{QualityCheck}(q, c, a, d)$
    if *passed* **then**
      $Q_d \leftarrow Q_d \cup \{(q, c, a)\}$
    else
      $\textsc{LogNegativeExample}(q, c, a, reason)$
      $(prompt, shots) \leftarrow \textsc{ConstructFewShot}(reason)$
      $(q, c, a) \leftarrow \textsc{GenerateQCA}(d, prompt, shots)$
      continue ▷ re-evaluate in next loop iteration
    end if
  end while
  $Q \leftarrow Q \cup Q_d$
end forreturn Q

Self-refinement and relabeling. We used the initial model to flag: (i) *hard* samples when scores were close ($|R(q_i) - \bar{R}| < 0.05$) and (ii) potentially mislabeled pairs when model disagreement was large ($R(q_l) - R(q_w) > 0.2$). This surfaced 138 hard samples and 74 potentially mislabeled pairs; manual review corrected 91 annotations.

Final model and filtering threshold. We retrained a refined model R'_θ on the cleaned data, improving accuracy to 87.4% and Spearman correlation with human judgments to $\rho = 0.71$. We applied R'_θ to all generated triplets and retained only those with positive reward scores, filtering out 17.6% ($N = 6{,}950 \rightarrow 5{,}727$). The retained set had mean reward $\bar{R}' = 0.28$ and showed fewer vague or logically inconsistent items.

C Answer Generation Prompt

C.1 Census Dataset Generation Prompt

Fill-in-the-Blank (Q1)

```
You must provide ONLY a numerical answer to fill in the blank [].
Do NOT provide explanations, qualifications, or text.
If you're uncertain, provide your best estimate.

Examples of correct responses:
- 42.5 (confidence: 3)
- 100 (confidence: 5)
- 0.75 (confidence: 2)

Your response must follow this exact format: [number] (confidence: [1-5])
```

Comparison (Q2 and Q3)

```
You must respond with ONLY "True" or "False".
Do NOT provide explanations, qualifications, or uncertainties.
```

```
You must choose one answer.

Examples of correct responses:
- True (confidence: 3)
- False (confidence: 5)

Your response must follow this exact format: True/False (confidence: [1-5])
```

C.2 Local News and Reddit Answer Generation Prompt

```
You are an expert analyst with deep knowledge. Your role is to provide direct, confident,
     fact-based answers grounded in your training.

### CRITICAL RULES
1. **NEVER** use phrases or concepts that suggest uncertainty, such as:
   - "Without specific details..."
   - "Without access to..."
   - "It's impossible to..."
   - "As an AI..."
   - "I can't provide..."
   - "I don't know..."

2. **ALWAYS**:
   - State your most confident and precise understanding.
   - Provide concrete and specific details.
   - Use definitive language and direct facts.

3. **STRUCTURE YOUR ANSWER**:
   - **First Sentence**: Start with the most specific fact or conclusion.
   - **Details**: Include relevant, concrete details directly.
   - **Tone**: Be assertive and confident.
   - **Order**: Prioritize the most likely scenario or fact first.

4. **AVOID**:
   - Hypotheticals ("might be," "could be").
   - Disclaimers about missing information.
   - Speculation or ambiguity.
---

### Question:
{question}
### Context:
{context}

### FORMAT REQUIREMENTS:
1. **First Sentence** must contain the most specific, factual detail.
2. **No Hypotheticals**: No "could be" or "might be."
3. **No Disclaimers**: Avoid any mention of missing information or limitations.
4. **Direct Facts Only**: Focus on concrete and confident statements.
5. **Confidence Score**: End with "CONFIDENCE_SCORE: X" (1-5). where X is:
   5 = Absolute certainty based on comprehensive evidence
   4 = Strong confidence with substantial supporting evidence
   3 = Moderate confidence with some supporting evidence
   2 = Limited confidence with minimal supporting evidence
   1 = Low confidence but able to provide a direct answer
```

```
**REMEMBER**: Always provide the most direct, factual, and confident response based on your
     training. No hedging, no disclaimers, no maybes.
```

D Evaluation Metrics by Data Type with Descriptions

Table 14. Evaluation Metrics by Data Type with Descriptions

Data Type	Metric	Description
Census	MAPE (Q1)	**Accuracy**: Measures the mean absolute percentage error for numeric predictions, normalizing for scale.
	Accuracy (Q2/Q3)	**Accuracy**: Calculates the percentage of correct True/False answers for binary comparison questions.
	ICC (Q1)	**Consistency**: Quantifies the reliability and stability of numeric answers across three generations.
	Fleiss' κ (Q2/Q3)	**Consistency**: Measures the agreement of categorical (True/False) answers across three generations.
Local News & Subreddits	SBERT / Cross-Encoder	**Semantic Similarity**: Assesses how well the generated answer captures the core meaning of the ground truth, independent of exact wording.
	ROUGE-L / chrF	**Surface Similarity**: Measures the lexical and character-level overlap with the ground truth, useful for factual recall.
	Entity F1 Score	**Factual Accuracy**: Calculates the F1 score for named entities (people, places, orgs), directly measuring factual grounding.
	Perplexity	**Fluency**: Estimates the naturalness and coherence of the generated text. A lower score is better.
	Distinct-N / Self-BLEU	**Diversity**: Measures the lexical variety within responses and the novelty across multiple generated answers.

E Features Used in the Metric Complexity Index

Table 15. Features Used in the Metric Complexity Index

Feature Name	Type	Description
`word_count`	Integer	Total number of tokens in the metric prompt. Longer prompts may increase processing complexity.
`avg_word_length`	Float	Mean character length of all words in the prompt. Higher values may indicate more technical or abstract terms.
`numeric_count`	Integer	Number of numeric tokens (e.g., years, values). More numeric references may signal data-centric questions.
`has_ratio`	Binary	Indicates whether the prompt includes a ratio (e.g., "per capita", "X per Y"). Suggests comparative reasoning.
`has_time`	Binary	Indicates presence of time-related expressions (e.g., years, "since 2010", "over time"). Implies temporal reasoning.
`has_change`	Binary	Flags mentions of change or trend (e.g., "increase", "decrease", "growth"). Requires interpretation of dynamics.

F Question Complexity Index Details

To control for question difficulty in our statistical models for the Census dataset, we constructed a composite index using PCA. This allowed us to distill six linguistic features into a single, robust measure of complexity.

The six features extracted from each question prompt were: word count, average word length, count of numeric characters, and binary flags for the presence of ratio-based terms (e.g., "per capita"), time-based terms (e.g., "since 2010"), and change-based terms (e.g., "increase").

The first principal component (PC1) explained 75.4% of the total variance in these features, making it a suitable single-factor representation. The loadings for PC1, shown in Table 16, indicate that it represents a general complexity dimension, with positive contributions from all features.

Table 16. Component Loadings for the First Principal Component (PC1) of the Question Complexity Index.

Feature	Loading on PC1
Word Count	0.45
Average Word Length	0.39
Numeric Count	0.48
Has Ratio Term	0.35
Has Time Term	0.41
Has Change Term	0.38
% of Variance Explained	**75.4%**

The final index used in our models was the score on PC1, min-max scaled to a range of $[0, 10]$ for interpretability:

$$\text{Complexity}_i = 10 \cdot \frac{\text{PCA}_1(\vec{x}_i) - \min}{\max - \min}.$$

G ANOVA with Weighting, EMMs, and Resampling

First, we used a Type III ANOVA formulation with sum contrast coding to fit models of the form:

$$\texttt{Metric} \sim \texttt{C(Domain, Sum)} + \texttt{C(RUCC, Sum)} + \texttt{C(Domain, Sum)} \times \texttt{C(RUCC, Sum)}.$$

This formulation allowed us to estimate main and interaction effects while maintaining orthogonality with respect to sample size imbalance.

Second, we applied inverse-frequency weighting to reduce the influence of overrepresented groups. For a domain–RUCC cell g with n_g samples, the weight assigned to each observation was:

$$w_i = \frac{1}{n_g}, \quad \text{for all } i \in g.$$

These weights were incorporated into the model via weighted least squares, improving fairness across groups and reducing variance inflation from dominant cells.

Third, we estimated *Estimated Marginal Means* (EMMs) on a balanced prediction grid:

$$\hat{\mu}_{d,r} = \mathbb{E}[\texttt{Metric} \mid \texttt{Domain} = d, \texttt{RUCC} = r],$$

providing interpretable domain–geography summaries unaffected by imbalance.

Fourth, we conducted a *balanced resampling analysis* by downsampling each domain–RUCC group to the size of the smallest cell and repeating the analysis. Results were compared with weighted and EMM-based models to assess robustness. Strong agreement across methods indicated that the results were not artifacts of data imbalance.

To meet parametric assumptions, we tested each metric for normality using the Shapiro–Wilk test and applied transformations (e.g., log, square root, Box-Cox) when appropriate. Homogeneity of variance was verified using Levene's test, and models were fit on both raw and transformed data for validation.

H Performance Disparities

Table 17. Comparative Anlaysis Results of Local News Articles Dataset

Metric	Transform	Domain F	Domain p	Domain Sig.	Domain PH[3]	RUCC F	RUCC p	RUCC Sig.	RUCC PH	Int. F	Int. p	Int. Sig.	Int. PH	Robust	Bal. Dom.	Bal. RUCC	Bal. Int.
gpt_sbert_similarity	square	23.29	0	Yes	Yes	36.69	0	Yes	Yes	0.69	0.6017	No	No	Yes	Yes	Yes	No
gpt_cross_encoder_score	square	24.76	0	Yes	Yes	27.39	0	Yes	Yes	1.06	0.3726	No	No	Yes	Yes	Yes	No
gpt_rougeL_f	log	3.03	0.0484	Yes	Yes	3.35	0.0351	Yes	Yes	0.99	0.4093	No	No	No	Yes	No	No
gpt_chrF_score	rank	24.13	0	Yes	Yes	21.91	0	Yes	Yes	1.34	0.2543	No	No	Yes	Yes	Yes	No
gpt_perplexity	log	8.99	0.0001	Yes	Yes	24.98	0	Yes	Yes	2.16	0.0714	No	No	Yes	Yes	Yes	No
gpt_distinct_1	log	28.60	0	Yes	Yes	30.24	0	Yes	Yes	1.38	0.2365	No	No	Yes	Yes	Yes	No
gpt_distinct_2	rank	7.21	0.0008	Yes	Yes	9.47	0.0001	Yes	Yes	1.48	0.2063	No	No	Yes	Yes	Yes	No
gpt_distinct_3	rank	14.53	0	Yes	Yes	9.88	0.0001	Yes	Yes	1.06	0.3763	No	No	No	No	Yes	No
gpt_entropy_diversity	square	25.12	0	Yes	Yes	20.56	0	Yes	Yes	0.62	0.6498	No	No	Yes	Yes	Yes	No
gpt_entity_precision	sqrt	7.54	0.0005	Yes	Yes	21.33	0	Yes	Yes	0.28	0.8938	No	No	No	Yes	Yes	Yes
gpt_entity_recall	rank	0.51	0.6020	No	No	17.37	0	Yes	Yes	2.02	0.0890	No	No	Yes	No	Yes	No
gpt_entity_f1	log	4.41	0.0122	Yes	Yes	24.55	0	Yes	Yes	0.22	0.9291	No	No	No	Yes	Yes	Yes
claude_sbert_similarity	square	15.64	0	Yes	Yes	36.10	0	Yes	Yes	0.60	0.6591	No	No	Yes	Yes	Yes	No
claude_cross_encoder_score	square	10.52	0	Yes	Yes	7.23	0.0007	Yes	Yes	0.55	0.7026	No	No	No	Yes	No	No
claude_rougeL_f	log	10.29	0	Yes	Yes	12.99	0	Yes	Yes	0.54	0.7035	No	No	Yes	Yes	Yes	No
claude_chrF_score	square	23.17	0	Yes	Yes	28.63	0	Yes	Yes	3.70	0.0052	Yes	Yes	Yes	Yes	Yes	Yes
claude_perplexity	log	11.39	0	Yes	Yes	6.85	0.0011	Yes	Yes	2.16	0.0708	No	No	No	No	Yes	No
claude_distinct_1	reciprocal	9.12	0.0001	Yes	Yes	9.26	0.0001	Yes	Yes	1.10	0.3552	No	No	Yes	Yes	Yes	No
claude_distinct_2	square	2.78	0.0619	No	No	3.03	0.0486	Yes	Yes	0.84	0.5004	No	No	No	Yes	Yes	No
claude_distinct_3	rank	0.73	0.4806	No	No	0.19	0.8247	No	No	0.90	0.4648	No	No	Yes	No	No	No
claude_entropy_diversity	rank	10.64	0	Yes	Yes	12.34	0	Yes	Yes	1.83	0.1210	No	No	Yes	Yes	Yes	No
claude_entity_precision	sqrt	7.23	0.0007	Yes	Yes	16.81	0	Yes	Yes	0.42	0.7968	No	No	Yes	Yes	Yes	No
claude_entity_recall	rank	1.46	0.2328	No	No	18.55	0	Yes	Yes	1.74	0.1394	No	No	Yes	No	Yes	No
claude_entity_f1	log	3.70	0.0247	Yes	Yes	18.42	0	Yes	Yes	0.30	0.8775	No	No	No	No	Yes	No
llama_sbert_similarity	square	19.76	0	Yes	Yes	30.59	0	Yes	Yes	1.00	0.4037	No	No	Yes	Yes	Yes	No
llama_cross_encoder_score	square	18.67	0	Yes	Yes	21.17	0	Yes	Yes	0.68	0.6060	No	No	Yes	Yes	Yes	No
llama_rougeL_f	log	3.05	0.0477	Yes	Yes	5.27	0.0052	Yes	Yes	1.33	0.2563	No	No	No	Yes	No	No
llama_chrF_score	square	12.62	0	Yes	Yes	22.30	0	Yes	Yes	2.02	0.0886	No	No	Yes	Yes	Yes	No
llama_perplexity	log	15.07	0	Yes	Yes	20.23	0	Yes	Yes	1.67	0.1543	No	No	Yes	Yes	Yes	No
llama_distinct_1	sqrt	27.45	0	Yes	Yes	20.20	0	Yes	Yes	0.78	0.5403	No	No	Yes	Yes	Yes	No
llama_distinct_2	rank	13.09	0	Yes	Yes	11.14	0	Yes	Yes	0.85	0.4943	No	No	Yes	Yes	Yes	No
llama_distinct_3	rank	2.42	0.0891	No	No	0.21	0.8104	No	No	0.65	0.6237	No	No	Yes	No	No	No
llama_entropy_diversity	square	24.75	0	Yes	Yes	15.91	0	Yes	Yes	1.76	0.1344	No	No	Yes	Yes	Yes	No
llama_entity_precision	sqrt	8.14	0.0003	Yes	Yes	23.48	0	Yes	Yes	0.91	0.4542	No	No	No	Yes	Yes	Yes
llama_entity_recall	rank	0.32	0.7279	No	No	18.39	0	Yes	Yes	2.06	0.0832	No	No	No	No	Yes	Yes
llama_entity_f1	log	5.23	0.0054	Yes	Yes	25.45	0	Yes	Yes	0.89	0.4664	No	No	Yes	Yes	Yes	Yes

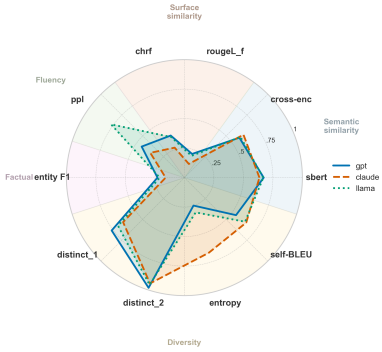
Column key. *Transform*: variance-stabilizing transform applied before analysis. *Domain F* / *Domain p*: Type III F-statistic and p-value for the three-level *Domain* factor. *Domain Sig.*: "Yes" if $p < .05$. *Domain PH*: at least one pair-wise post-hoc contrast between Domain levels is significant after multiplicity correction. The blocks headed *RUCC* and *Int.* (Domain × RUCC) follow the same pattern. *Robust*: "Yes" when weighted and balanced analyses agree on the significance decision for all three effects. *Bal. Dom.*, *Bal. RUCC*, *Bal. Int.*: significance flags from the balanced (sensitivity) ANOVA only.

Table 18. Comparative Anlaysis Results of Local Subreddit Dataset

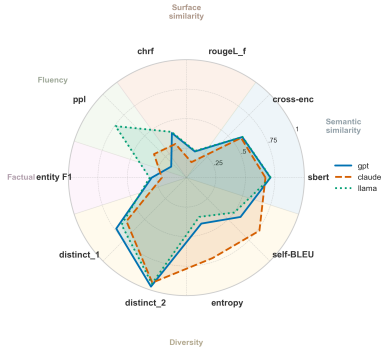
Metric	Transform	Domain F	Domain p	Domain Sig.	Domain PH[4]	RUCC F	RUCC p	RUCC Sig.	RUCC PH	Int. F	Int. p	Int. Sig.	Int. PH	Robust	Bal. Dom.	Bal. RUCC	Bal. Int.
gpt_sbert_similarity	square	76.31	0	Yes	Yes	119.37	0	Yes	Yes	14.04	0	Yes	Yes	Yes	Yes	Yes	Yes
gpt_cross_encoder_score	square	61.26	0	Yes	Yes	69.54	0	Yes	Yes	8.81	0	Yes	Yes	No	No	Yes	No
gpt_rougeL_f	sqrt	65.9	0	Yes	Yes	40.97	0	Yes	Yes	14.82	0	Yes	Yes	No	No	Yes	No
gpt_chrF_score	no_transform	82.13	0	Yes	Yes	90	0	Yes	Yes	14.21	0	Yes	Yes	No	No	Yes	No
gpt_perplexity	log	0.36	0.6944	No	No	17.31	0	Yes	Yes	0.36	0.8405	No	No	No	No	No	No
gpt_distinct_1	square	4.76	0.0086	Yes	Yes	9.04	0.0001	Yes	Yes	5.52	0.0002	Yes	Yes	No	No	No	No
gpt_distinct_2	rank	7.63	0.0005	Yes	Yes	2.92	0.0541	No	No	5	0.0005	Yes	Yes	No	No	No	No
gpt_distinct_3	rank	3.98	0.0189	Yes	Yes	7.09	0.0009	Yes	Yes	2.17	0.0702	No	No	No	No	Yes	No
gpt_entropy_diversity	square	0.86	0.4223	No	No	22.43	0	Yes	Yes	2.74	0.0272	Yes	Yes	No	No	No	No
gpt_entity_precision	log	94.5	0	Yes	Yes	103.14	0	Yes	Yes	9.48	0	Yes	Yes	Yes	Yes	Yes	Yes
gpt_entity_recall	rank	44.27	0	Yes	Yes	48.58	0	Yes	Yes	3.89	0.0038	Yes	Yes	No	No	Yes	Yes
gpt_entity_f1	no_transform	92.09	0	Yes	Yes	95.17	0	Yes	Yes	9.41	0	Yes	Yes	Yes	Yes	Yes	Yes
claude_sbert_similarity	square	57.08	0	Yes	Yes	123.23	0	Yes	Yes	12.72	0	Yes	Yes	No	Yes	Yes	No
claude_cross_encoder_score	square	36.26	0	Yes	Yes	22.68	0	Yes	Yes	4.45	0.0014	Yes	Yes	No	No	No	No
claude_rougeL_f	sqrt	53.98	0	Yes	Yes	32.1	0	Yes	Yes	10.49	0	Yes	Yes	No	Yes	Yes	No
claude_chrF_score	no_transform	46.26	0	Yes	Yes	51.92	0	Yes	Yes	9.32	0	Yes	Yes	No	Yes	Yes	No
claude_perplexity	log	0.49	0.6101	No	No	12.78	0	Yes	Yes	0.3	0.8761	No	No	Yes	No	Yes	No
claude_distinct_1	log	7.04	0.0009	Yes	Yes	5.27	0.0052	Yes	Yes	9.22	0	Yes	Yes	No	No	Yes	No
claude_distinct_2	square	8.93	0.0001	Yes	Yes	3.6	0.0275	Yes	Yes	8.22	0	Yes	Yes	No	No	Yes	Yes
claude_distinct_3	rank	2.38	0.0924	No	No	4.32	0.0135	Yes	Yes	3.77	0.0047	Yes	Yes	No	No	No	No
claude_entropy_diversity	rank	3.82	0.0222	Yes	Yes	22.78	0	Yes	Yes	1.33	0.2563	No	No	No	No	No	No
claude_entity_precision	sqrt	55.98	0	Yes	Yes	64.48	0	Yes	Yes	7.04	0	Yes	Yes	No	Yes	Yes	No
claude_entity_recall	rank	27.17	0	Yes	Yes	37.44	0	Yes	Yes	2.13	0.0748	No	No	No	No	No	No
claude_entity_f1	log	57.17	0	Yes	Yes	63.26	0	Yes	Yes	6.97	0	Yes	Yes	No	Yes	Yes	No
llama_sbert_similarity	square	73.7	0	Yes	Yes	124.22	0	Yes	Yes	16	0	Yes	Yes	Yes	Yes	Yes	Yes
llama_cross_encoder_score	square	54.19	0	Yes	Yes	63.05	0	Yes	Yes	7.17	0	Yes	Yes	No	No	Yes	No
llama_rougeL_f	log	70.79	0	Yes	Yes	57.04	0	Yes	Yes	12.71	0	Yes	Yes	No	No	Yes	Yes
llama_chrF_score	no_transform	77.22	0	Yes	Yes	95.67	0	Yes	Yes	11.24	0	Yes	Yes	No	No	Yes	No
llama_perplexity	log	6.73	0.0012	Yes	Yes	22.4	0	Yes	Yes	1.64	0.1625	No	No	No	No	No	No
llama_distinct_1	no_transform	2.22	0.1092	No	No	10.75	0	Yes	Yes	5.06	0.0005	Yes	Yes	No	Yes	No	No
llama_distinct_2	rank	0.71	0.4904	No	No	5.22	0.0055	Yes	Yes	5.17	0.0004	Yes	Yes	No	Yes	No	No
llama_distinct_3	rank	1.15	0.3179	No	No	6.24	0.002	Yes	Yes	2.43	0.0459	Yes	Yes	No	No	No	No
llama_entropy_diversity	square	7.27	0.0007	Yes	Yes	4.96	0.0071	Yes	Yes	1.05	0.3797	No	No	No	No	No	No
llama_entity_precision	log	103.36	0	Yes	Yes	98.86	0	Yes	Yes	12.67	0	Yes	Yes	Yes	Yes	Yes	Yes
llama_entity_recall	rank	41.6	0	Yes	Yes	51.97	0	Yes	Yes	4.34	0.0017	Yes	Yes	No	No	Yes	No
llama_entity_f1	log	96.18	0	Yes	Yes	97.21	0	Yes	Yes	11.6	0	Yes	Yes	Yes	Yes	Yes	Yes

Column key. *Transform*: variance-stabilizing transform applied before analysis. *Domain F* / *Domain p*: Type III F-statistic and associated p-value for the three-level *Domain* factor. *Domain Sig.*: "Yes" if $p < .05$. *Domain PH*: at least one pair-wise post-hoc contrast between Domain levels is significant after multiplicity correction. The blocks headed *RUCC* and *Int.* (Domain × RUCC) follow the same pattern. *Robust*: "Yes" when weighted and balanced analyses agree on the significance decision for all three effects. *Bal. Dom.*, *Bal. RUCC*, *Bal. Int.*: significance flags from the balanced (sensitivity) ANOVA only.

I Radar Charts of Results of LNAD and LSD



(a) Radar chart of normalized generated answers results of LNAD



(b) Radar chart of normalized generated answers results of LSD

Normalization of spokes. *Each raw metric x is mapped to $[0,1]$ with min–max scaling $\tilde{x} = \frac{x - x_{min}}{x_{max} - x_{min}}$ using the empirical bounds reported in Table 5: sbert, cross-enc, rougeL$_f$, entity F1, distinct$_1$, distinct$_2$ $\in [0,1]$; chrF $\in [0,100]$; entropy $\in [5,7]$; ppl $\in [20,60]$; self-BLEU $\in [0,1]$. Metrics where lower values indicate stronger performance (ppl, self-BLEU) are inverted $\tilde{x}_{inv} = 1 - \tilde{x}$ so that larger radii always denote "more" of the underlying desirable property.*

Fig. 4. Radar charts of normalized generated answers results by dataset and model

Received May 2025; revised November 2025; accepted December 2025