LocalBench: Benchmarking LLMs on County-Level Local Knowledge and Reasoning

Zihan Gao[1], **Yifei Xu**[2], **Jacob Thebault-Spieker**[1]

[1]University of Wisconsin-Madison
[2]University of California, Los Angeles
zihan.gao@wisc.edu, yfxu@ucla.edu, jacob.thebaultspieker@wisc.edu

Abstract

Large language models (LLMs) have been widely evaluated on macro-scale geographic tasks, such as global factual recall, event summarization, and regional reasoning. Yet, their ability to handle hyper-local knowledge remains poorly understood. This gap is increasingly consequential as real-world applications, from civic platforms to community journalism, demand AI systems that can reason about neighborhood-specific dynamics, cultural narratives, and local governance. Existing benchmarks fall short in capturing this complexity, often relying on coarse-grained data or isolated references. We present LocalBench, the first benchmark designed to systematically evaluate LLMs on county-level local knowledge across the United States. Grounded in the Localness Conceptual Framework, LocalBench includes 14,782 validated question-answer pairs across 526 U.S. counties in 49 states, integrating diverse sources such as Census statistics, local subreddit discourse, and regional news. It spans physical, cognitive, and relational dimensions of locality. Using LocalBench, we evaluate 13 state-of-the-art LLMs under both closed-book and web-augmented settings. Our findings reveal critical limitations: even the best-performing models reach only 56.8% accuracy on narrative-style questions and perform below 15.5% on numerical reasoning. Moreover, larger model size and web augmentation do not guarantee better performance, for example, search improves Gemini's accuracy by $+13.6\%$, but reduces GPT-series performance by -11.4%. These results underscore the urgent need for language models that can support equitable, place-aware AI systems: capable of engaging with the diverse, fine-grained realities of local communities across geographic and cultural contexts.

Datasets — https://github.com/zihanngao/LocalBench

Introduction

LLMs have been extensively evaluated on tasks involving macro-scale geographic knowledge, such as factual recall (Moayeri, Tabassi, and Feizi 2024), global event summarization (Almeida et al. 2025), and cross-regional spatiotemporal reasoning (Gurnee and Tegmark 2024). These evaluations demonstrate that LLMs are capable of handling broad geographic contexts and structured global facts. However, this focus on macro-scale capabilities obscures a critical limitation examined in this work: *LLMs continue to struggle with hyper-local knowledge*—the fine-grained, community-specific information essential for grounded, real-world applications that demand geographic and cultural nuance.

The demand for hyper-local AI capabilities is rapidly growing. Civic platforms rely on AI systems that can navigate local governance structures and community dynamics (Guridi, Cheyre, and Yang 2025), while community journalism initiatives need AI tools that contextualize events within local cultural narratives (MacVey 2022). Location-aware services must reason about neighborhood-specific preferences and constraints (Chen et al. 2024; Tang et al. 2024). Despite these needs, current LLMs often default to generic outputs or propagate biases, failing to capture *the multifaceted nature of local knowledge*, which includes not only statistical facts but also cultural norms, vernacular expressions, and community interactions, etc. (Gao, Cranshaw, and Thebault-Spieker 2025).

Existing benchmarks have made important progress in evaluating geographic knowledge, but they fall short in capturing the fine-grained complexity of local reasoning required for real-world, community-centered applications. Prior work focuses largely on global factual recall (Moayeri, Tabassi, and Feizi 2024; Almeida et al. 2025), isolated cultural references (Shi et al. 2025; Dudy et al. 2025), or generative descriptions of place identity in urban settings (Jang et al. 2024), often targeting specific regions, subjective narratives, or coarse-grained indicators. Other efforts highlight geographic biases in model performance (Manvi et al. 2024a; Zhu, Wang, and Liu 2024), or develop tools for geospatial inference (Manvi et al. 2024b; Wu et al. 2024), but do not offer comprehensive frameworks for local knowledge reasoning and evaluation. A core reason for this gap is that existing evaluations lack the geographic granularity and knowledge breadth needed to capture how local knowledge operates in practice. Local reasoning is not just about factual recall or cultural trivia, but requires integrating statistical indicators, cultural narratives, vernacular expressions, and community-specific governance knowledge across diverse social contexts (Gao, Cranshaw, and Thebault-Spieker 2024).

Focusing on *county-level knowledge* offers a tractable yet underexplored path toward evaluating local reasoning: (i) counties are the smallest U.S. administrative unit with consistently reported socio-economic statistics (e.g., ACS, CDC, USDA), (ii) they map cleanly onto electoral and governance structures that drive civic decision-making, making it a practical sweet-spot between granularity and data availability. However, current LLMs systematically overrepresent major metropolitan areas, leaving rural and smaller communities neglected (Manvi et al. 2024a; Sun et al. 2023). Addressing this imbalance requires an evaluation framework that reflects the full spectrum of localities, especially those that are underrepresented due to data scarcity. Moreover, it remains unclear whether LLMs can overcome these limitations through web search augmentation, particularly when community-specific knowledge is fragmented or absent from standard retrieval sources.

In this work, we introduce LOCALBENCH, a benchmark for evaluating LLMs on county-level local knowledge and reasoning. Grounded in the Localness Conceptual Framework (Gao, Cranshaw, and Thebault-Spieker 2025), which defines local knowledge across physical, cognitive, and relational dimensions, LOCALBENCH comprises 14,782 question–answer pairs across 526 U.S. counties. It spans all dimensions of the framework and is validated via expert annotation.

Using this benchmark, we evaluate 13 leading LLMs under both closed-book and web-augmented settings. Our findings reveal clear limitations: the best-performing models achieve only 56.8% accuracy on narrative-style questions and struggle with numerical reasoning, falling below 15.5%. Furthermore, increased model scale and current implementations of web augmentation do not guarantee better performance: web search improves Gemini (+13.6%) but harms GPT-series models (–11.4%), and larger or mixture-of-experts (MoE) architectures show no consistent advantage over smaller non-MoE models.

Our contributions are:

- **LOCALBENCH:** We introduce the first benchmark targeting U.S. county-level local knowledge and reasoning to advance research in place-aware AI. It contains 14,782 QA pairs covering 526 counties, spans all dimensions of the localness, and is validated through expert annotation.
- **Comprehensive LLM evaluation:** We present a systematic evaluation of 13 state-of-the-art LLMs in both closed-book and web-augmented settings reveals how even today's strongest models falter when confronted with fine-grained, place-aware queries.
- **Empirical insights for place-aware LLMs:** We provide evidence that challenges the common assumption that increased model size or retrieval augmentation inherently improves local reasoning, offering actionable insights for developing future place-aware language models.

Related Work

Localness and Community-Centered Knowledge

Digital-placemaking research shows that community platforms (e.g., neighborhood forums, local subreddits) rely on hyper-local information, such as vernacular expressions, governance details, and shared narratives, to foster civic engagement (Aubin Le Quéré, Naaman, and Fields 2024; Park et al. 2014; Gao, Cranshaw, and Thebault-Spieker 2024). LLMs are now used in this space, yet studies reveal that their place descriptions are often generic or stereotyped, especially outside iconic cities (Jang et al. 2024; Zhu, Wang, and Liu 2024). Existing benchmarks probe single facets: global facts (Moayeri, Tabassi, and Feizi 2024), regional term recognition (Shi et al. 2025), or city stereotypes (Jang et al. 2024), but none test whether models grasp the full texture of locality.

To fill this gap, we adopt the Localness Conceptual Framework (Gao, Cranshaw, and Thebault-Spieker 2025), which organizes local knowledge into three interwoven domains: Physical, Cognitive, and Relational, covering 7 dimensions and 88 subcomponents. This structure aligns with "sense of place" theory (Lengen and Kistemann 2012) and provides a conceptual definition that allows us to ask whether an LLM can, for example, (i) cite a neighborhood landmark (Physical), (ii) decode local slang (Cognitive), and (iii) reference a county board's role in a festival (Relational).

By using localness as our evaluation lens we move beyond generalized location-focused knowledge like "The Eiffel Tower is in Paris," and evaluate whether AI systems can surface hyper-local narratives, vernacular, and community dynamics across urban, suburban, and rural contexts. In short, we explore if LLMs and agentic web-search approaches are sufficiently capable to be trusted for community-focused applications.

Geographic Knowledge and Cultural Locality Benchmarks

Existing benchmarks have made progress in evaluating LLMs' geographic knowledge, but they often focus on global or national-level factual recall. *WorldBench* (Moayeri, Tabassi, and Feizi 2024) tests LLMs on factual recall of national indicators such as GDP and literacy rates across over 180 countries. *TiEBe* (Almeida et al. 2025) evaluates temporal event recall at global and regional scales, focusing on historically significant world events. These benchmarks prioritize breadth over locality, without addressing fine-grained community-specific knowledge.

Other efforts target cultural locality or factual transfer in regional contexts, but often with limited geographic scope or narrow task design. *LIBRA* (Shi et al. 2025) focuses on New Zealand-specific cultural terms, probing local bias through term recognition and classification. *LoFTI* (Dudy et al. 2025) examines factuality transfer by prompting LLMs to adapt general knowledge into Indian state- and city-level contexts. *Place Identity* (Jang et al. 2024) evaluates generative descriptions of cities, but focuses exclusively on urban areas, leaving suburban and rural communities unexamined.

These benchmarks address valuable aspects of cultural knowledge but lack systematic evaluation of local knowledge across diverse geographic and social settings.

Geospatial Reasoning and Spatial Bias Probing

Other benchmarks target spatial reasoning and geographic bias, but do not comprehensively evaluate local knowledge reasoning. *TorchSpatial* (Wu et al. 2024) assesses geospatial representation learning using geo-tagged image classification tasks, while *GeoLLM* (Manvi et al. 2024b) probes population inference and location classification via OpenStreetMap and WorldPop data. These efforts focus on spatial embeddings and regression tasks, rather than reasoning over local cultural or governance contexts.

Complementing these tasks, several studies highlight geographic performance biases in LLMs, particularly their tendency to perform better on well-documented, high-resource regions while struggling with underrepresented communities (Manvi et al. 2024a; Zhu, Wang, and Liu 2024). These works document systematic failures in representing less-popular or under-resourced areas, but focus primarily on analyzing model outputs across regions rather than providing structured benchmarks for evaluating local knowledge breadth and depth.

Broader efforts in geoscience and geospatial foundation models address related challenges but remain orthogonal to the problem of local knowledge reasoning. *GeoGPT* (Zhang et al. 2023) introduces a tool-augmented pipeline for executing geospatial tasks and querying geospatial APIs, while *Contrastive Spatial Pretraining (CSP)* (Mai et al. 2023) develops multimodal foundation models via visual-text contrastive learning on remote sensing data. These methods advance spatial understanding but do not target natural language evaluation of local, community-level knowledge.

Benchmark Construction

We introduce LOCALBENCH, a large-scale benchmark designed to evaluate LLMs' ability to reason over county-level local knowledge. The dataset comprises 14,782 question–answer (QA) pairs covering 526 U.S. counties, with a balanced distribution across urban, suburban, and rural regions. Each QA pair is aligned to a ground-truth source and annotated according to the *Localness Conceptual Framework*.

Nature of Localness Evaluation

LOCALBENCH targets a core limitation of current LLMs: reasoning about hyper-local knowledge that is often difficult to retrieve online. Many questions require information from small local media outlets such as county newsletters, or community-level discussions from hyper-local forums like Reddit threads. Other questions demand understanding of fragmented institutional data, such as neighborhood-specific census metrics or municipal reports. These sources are rarely indexed comprehensively by search engines, are inconsistently formatted, and frequently require interpretive aggregation rather than direct retrieval (Gao, Cranshaw, and Thebault-Spieker 2024).

Data Sources

LOCALBENCH integrates three complementary data sources that together capture both structured and unstructured local knowledge. From the U.S. Census Bureau and other census data sources like USDA Agricultural Statistics Service, National Register of Historic Places, and others, we extract 34 localness indicators, spanning cognitive, physical, and relational domains as described by Gao, Cranshaw, and Thebault-Spieker (2025). Examples include the cropland fertilization rate (cognitive), the percentage of residents living in the same home for more than five years (physical), and total ballots cast in the 2020 election (relational).

Using Rural-Urban Continuum Codes (RUCC),we stratified counties into urban (RUCC 1–3), suburban (4–6), and rural (7–9) groups. From the 681 counties with complete data across all 34 indicators, we sample 60 counties per group (N=180). This structured source yields 6,120 QA pairs directly incorporated into the final dataset without quality filtering, evenly split between numerical questions (e.g., "What is the median household income in County X?") and comparison questions (e.g., "How does County X's unemployment rate compare to County Y's?"). Appendix 9 lists the corresponding QA pairs.

To capture unstructured local discourse, we further collect subreddit data from "The Global List of Local Reddits," from r/LocationReddit. We then manually inspected each subreddit and associated subreddits to corresponding counties manually. Data from January 2024 to March 2025 includes posts and the top-50 comments per thread, initially generating 4,210 candidate QA pairs, which produces 4,000 final QA pairs after quality filtering. These final QA pairs focus on narrative and interpretive aspects of local culture, events, and community concerns.

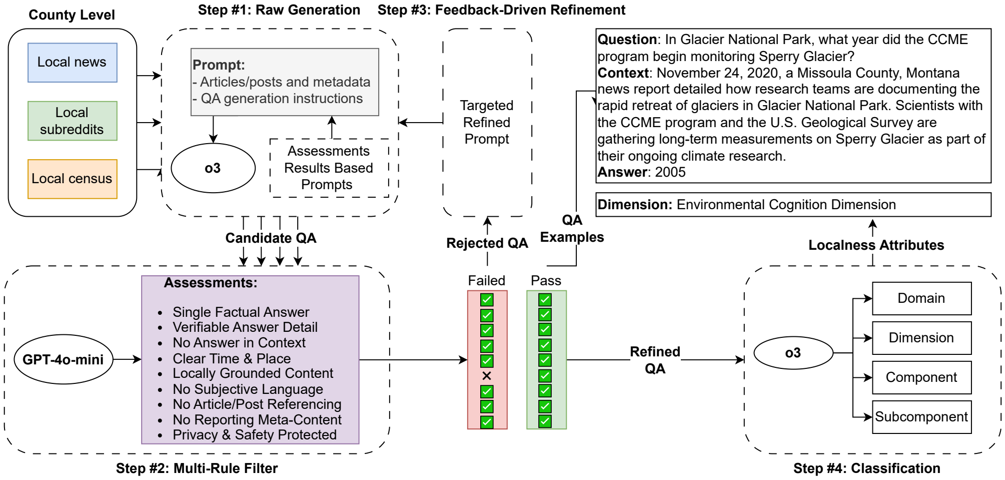
Additionally, we use the NELA-Local corpus (Horne et al. 2022), containing over 1.4 million local news articles from 313 U.S. outlets published between April 2020 and December 2021. Articles are tagged at the county level, initially generating 4,897 candidate QA pairs, which results in 4,662 final QA pairs after quality filtering. These final QA pairs cover governance, civic activities, and hyper-local events that were reported in local news outlets.

Across all sources, the pipeline processes 15,527 initial generation attempts, achieving an overall success rate of 96.9% to produce the final dataset of 14,782 QA pairs covering 526 unique counties across 49 states, with geographic diversity confirmed via Moran's $I = -0.003$ ($p = 0.491$), indicating no spatial autocorrelation.

QA Generation Pipeline

Step #1: Raw Generation Figure 1 illustrates the three-stage pipeline for QA pair generation and validation. In the first stage, we use the OpenAI o3 model to generate candidate QA pairs from the source materials. The generator operates with a temperature parameter of 0.7, top-p parameter of 0.9, and max_tokens parameter of 200, producing 1–3 QAs per document to balance diversity and quality. Importantly, generation is constrained to reason over the given input rather than hallucinate external knowledge, ensuring county-specific grounding.

Step #2: Multi-Rule Filter Census-derived QA pairs bypass the quality control pipeline due to their structured na-



Question: In Glacier National Park, what year did the CCME program begin monitoring Sperry Glacier?
Context: November 24, 2020, a Missoula County, Montana news report detailed how research teams are documenting the rapid retreat of glaciers in Glacier National Park. Scientists with the CCME program and the U.S. Geological Survey are gathering long-term measurements on Sperry Glacier as part of their ongoing climate research.
Answer: 2005

Dimension: Environmental Cognition Dimension

Figure 1: LocalBench construction pipeline. The process involves QA generation using a reasoning model and quality analysis via LLM-based assessment and filtering.

ture and verified accuracy. For Reddit and news sources (9,107 candidate pairs), a QA Quality Analyzer filters the generated pairs using a fine-tuned GPT-4o-mini model trained via Direct Preference Optimization (DPO) (Rafailov et al. 2023). The training dataset consists of 473 human-annotated QA pairs, labeled by two graduate researchers with inter-annotator agreement $\kappa = 0.84$. The analyzer evaluates each pair on nine criteria, including single factual answer validation, geographic grounding, subjectivity filtering, privacy preservation, and temporal consistency, and others (detailed in Appendix). Training parameters include a learning rate of one, batch size of eight, and two epochs.

Step #3: Feedback-Driven Refinement Reddit and news-derived pairs failing any quality criterion enter an iterative re-generation loop, where the QA Generator receives targeted feedback and retries up to three times. If a pair fails all attempts, we remove it from the dataset. After three regeneration rounds, we retain 95.2% of non-census pairs, contributing to the overall 96.9% dataset success rate.

Human Verification. To validate filter accuracy, two independent human annotators annotated 500 randomly sampled QA pairs. The annotators achieved inter-annotator agreement of $\kappa = 0.78$. When comparing with our QA Quality Filters' output, we observed an overall F1-score of 0.94. There is no meaningful correlation across error cases, suggesting that our QA Quality Analyzer produces QA pairs of equivalent quality across different kinds of QA pairs (corelation analysis results are in Appendix 4).

Step #4: Localness Attribute Classification Each validated QA pair undergoes localness classification according to the Localness Conceptual Framework using the o3 model. The classifier receives the question, answer, and original source context, then assigns labels across four hierarchical levels: domain (Physical/Cognitive/Relational), dimension, component, and subcomponent. Classification operates with a temperature parameter of 0.3 and a max_tokens parameter of 150 to ensure consistent, focused outputs. The classification prompt instructs the model to analyze the QA pair's core focus and assign the most specific applicable labels. This hierarchical structure ensures comprehensive coverage while maintaining classification granularity needed for detailed analysis.

Human Verification. To validate classification accuracy, two independent human annotators annotated 200 randomly sampled QA pairs. The annotators achieved inter-annotator agreement of $\kappa = 0.87$. When comparing human annotations against o3 classifications, we observed 94.2% precision across all classification levels, with domain-level precision reaching 98.5%.

Dataset Statistics

Table 1 provides comprehensive statistics for the final LocalBench dataset, broken down by localness components and data sources. The dataset exhibits balanced coverage across all major localness categories while maintaining consistent content complexity across components.

Evaluation Setup

Evaluation Protocol

As noted above, our analysis focuses both prominent proprietary and open source LLMs, both without web-search augmentation, and where possible, with web-search augmentation as well. To ensure reproducibility, all models are run with temperature parameters set to 0.0, and maximum token output parameters of 256 tokens. Each model receives standardized prompts designed to elicit factual, grounded responses while allowing models to express uncertainty when appropriate. Each model is evaluated over three independent runs with different random seeds and we report mean scores. For statistical comparisons, we conduct paired t-tests with

Domain	Dimension	Number	QLen	CLen	ALen	Census	Reddit	News	Rural	Suburban	Urban
Physical	Place Interaction	1,330	28.9	63.9	5.3	180	663	487	306	337	687
	Temporal Presence	2,907	29.2	62.4	4.2	1260	567	1,080	724	818	1,365
Cognitive	Cultural Understanding	1,435	28.5	63.6	4.4	720	286	429	352	411	672
	Environmental Cognition	1,739	30.3	63.2	5.3	540	860	339	519	471	749
	Local Knowledge	3,855	28.5	63.1	4.5	1620	914	1,321	987	1,088	1,780
Relational	Emotional Connection	838	26.8	64.2	3.0	720	54	64	254	267	317
	Social/Community Engagement	2,678	27.5	63.6	4.5	1080	656	942	651	774	1,253
Overall	–	**14,782**	**28.49**	**63.45**	**4.43**	**6,120**	**4,000**	**4,662**	**3,843**	**4,166**	**6,823**

Table 1: Dataset statistics by localness dimensions and data sources, with counts by RUCC group. **QLen**, **CLen**, and **ALen** represent the average token lengths of the *Query*, *Context*, and *Answer*, respectively. Token counts measured using GPT-4o tokenizer.

Bonferroni correction ($\alpha = 0.05$) to control for multiple comparisons.

Models Evaluated

We evaluate a diverse set of models across different capability tiers. Proprietary models include GPT-4o (gpt-4o-2024-08-06), GPT-o3 (o3-2025-04-16) (OpenAI 2025), Gemini-2.5-Pro, Gemini-2.5-Flash (Google 2025), Claude-4-Sonnet (claude-sonnet-4-20250514), Claude-3.7-Sonnet (claude-3-7-sonnet-20250219) (Anthropic 2025), Qwen3-235B-A22B/30B-A3B/32B/14B/8B (Yang et al. 2025). We also test web-augmented configurations, including GPT-4.1 with Search API integration and Gemini-2.5-Pro with Search Grounding.

Evaluation Metrics

We adopt a multi-faceted evaluation framework designed to capture factual correctness, semantic equivalence, numeric reasoning accuracy, and model confidence alignment. This is necessary to address the diverse formats and ambiguity present in locality-grounded questions. In all cases, we pose the question in our QA pairs to the LLM, and collect its generated answer. We then compare the generated answer to our ground-truth answer in our QA pair, based on these metrics.

To evaluate factual correctness, we compute both **Exact Match (EM)**, which enforces strict string equality after normalization, and **ROUGE-1 F1**, which tolerates paraphrasing through unigram overlap. To capture deeper equivalence beyond lexical variation, we compute a **semantic match** score: the cosine similarity between dense embeddings of the generated answer and our ground-truth answer, obtained with OpenAI's `text-embedding-3-small`.

For numerical answers, We report numerical **accuracy**, which counts a numerical prediction as correct if its relative error is under 2% of the gold value (and requires an exact zero when the gold answer is zero).

To handle ambiguous cases where answers may vary in surface form but remain valid, we introduce a **GPT Judge** mechanism. A GPT-4o-mini model receives the full QA context: original question, gold answer, generated answer and any supporting evidence, and outputs a binary judgment of correctness. We also capture the log-probability of the generated judgment token (e.g., "Correct" or "Incorrect"), allowing us to interpret the model's confidence in its assessment. To validate our GPT-based judge, we conduct human annotation on 200 randomly selected samples. The results show that the judge aligns with human annotations in 96% of cases, which demonstrates its reliability for downstream evaluation.

Lastly, we compute the **answer rate** as the proportion of model outputs that provide substantive responses (i.e., excluding empty or "I don't know" answers), reflecting model willingness to engage with local queries.

Results

We evaluate the performance of state-of-the-art LLMs on LocalBench to assess their capacity for local knowledge reasoning. Our analysis reveals substantial challenges that persist across task types, model architectures, and augmentation strategies.

A major finding is the sharp divide between non-numerical and numerical performance. While models such as *Gemini-2.5-Pro+Grounding* reach up to 56.8% GPT Judge accuracy on non-numerical questions, performance on numerical Census tasks remains critically low, with no model exceeding 12.8% accuracy. We observe that models sometimes refuse to answer numerical queries by explicitly stating a lack of knowledge. In particular, *GPT-4o* answers only 39.8% of such questions, whereas all other later-released models respond to at least 75%. This discrepancy may reflect updates in training data or changes in post-training strategies over time, amid the rapidly evolving landscape of LLM development. Web augmentation further exposes architectural differences in retrieval integration. Gemini models benefit substantially from external evidence, while surprisingly, *GPT-4.1+Web* exhibits degraded performance and reduced answer rates, suggesting that retrieval-based grounding can increase model uncertainty when poorly integrated. Model scaling does not consistently improve local knowledge reasoning. Larger models often underperform their smaller counterparts, and mixture-of-experts architectures fail to outperform dense models. In

Model	Non-Numerical QA					Numerical QA	
	EM	ROUGE-1	Semantic	GPT Judge	Ans Rate	Accuracy	Ans Rate
GPT-4o	22.0	30.7	53.0	32.8	99.6	6.2	39.8
GPT-4.1	**32.2**	**52.5**	**74.1**	47.0	100.0	6.2	100.0
GPT-4.1+Web	13.5	27.9	43.2	35.6	92.9	**15.5**	92.0
Gemini-2.5-Pro	28.0	52.0	70.5	52.5	100.0	12.8	100.0
Gemini-2.5-Flash	31.1	46.0	67.6	43.2	100.0	7.5	100.0
Gemini-2.5 Pro+Grounding	21.9	50.1	66.0	**56.8**	91.7	12.8	100.0
Claude-Sonnet-4	23.4	38.5	64.0	39.7	100.0	7.1	97.3
Claude-Sonnet-3.7	21.7	42.5	65.5	43.7	100.0	8.4	91.2
Qwen3-235B-A22B	19.9	29.0	54.0	27.3	99.3	6.6	77.0
Qwen3-30B-A3B	20.5	29.6	54.9	28.0	99.7	2.2	100.0
Qwen3-32B	20.0	29.9	55.4	27.7	99.7	4.9	99.1
Qwen3-14B	19.5	29.8	55.4	27.5	99.6	4.0	100.0
Qwen3-8B	16.3	27.2	54.1	22.9	99.6	3.1	75.2

Table 2: Performance Results Across Non-Numerical and Numerical QA

particular, spatially grounded numerical reasoning remains elusive even for the largest proprietary models, indicating that these limitations stem not from scale but from representational and architectural mismatches.

Taken together, these results highlight the persistent difficulty of local knowledge tasks for current LLMs. Addressing these limitations will require new modeling approaches that explicitly encode geographic reasoning and better handle context-dependent, place-specific information.

Breaking Down Model Performance

While average accuracy offers a basic ranking of models, effect size analysis reveals the magnitude and reliability of performance differences. We quantify performance differences with *Cohen's d*, a standardized mean-difference that expresses how many pooled standard deviations separate two systems; values near 0.20, 0.50, and 0.80 are conventionally interpreted as small, medium, and large effects, respectively (COHEN 1992). To protect the family-wise error rate during the many pairwise contrasts, we adjust each p-value with the Bonferroni procedure (Bland and Altman 1995) (results details in Appendix 5 and 6).

For non-numerical tasks, the *Gemini-2.5-Pro+Grounding* achieves the highest performance with 56.8% accuracy and a large positive effect size ($d = +0.485, p < 0.001$) relative to *GPT-4.1*. *GPT-4o* shows significant performance degradation with 32.8% accuracy and a large negative effect size ($d = -0.412, p < 0.001$). Open-source models lag substantially with large negative effect sizes (Qwen3-30B: $d = -0.581, p < 0.001$).

However, for numerical tasks, overall performance remains critically low across all models. The highest-performing models (*GPT-4.1+Web*) achieve only 15.5% accuracy, representing a substantial but still modest improvement ($d = +0.623, p < 0.001$) over the 6.2% baseline. Most concerning is the dramatic variation in answer rates, while most models maintain answer rates larger than 95%, *GPT-4o* drops to 39.8%, indicating systematic avoidance of numerical reasoning. Notably, open-source models exhibit broadly similar numerical reasoning capability to *GPT-4o*.

Web-Augmented Model Behavior

We find that web augmentation produces opposing effects across model families. For non-numerical QA, *Gemini-2.5-Pro+Grounding* shows substantial improvements (+13.6% in GPT Judge accuracy), achieving the benchmark's highest performance. In stark contrast, *GPT-4.1+Web* shows performance degradation (-11.4%), suggesting fundamental differences in retrieval integration capabilities. Further, we see that patterns in model answer rate suggest retrieval confidence issues. *GPT-4.1+Web* and *Gemini-2.5-Pro+Grounding* shows reduced willingness to answer from 100.0% to 92.9% and 91.7%, respectively, indicating that retrieved information may increase uncertainty in the model, somewhat counterintuitively. For numerical QA, *GPT-4.1+Web* and *Gemini-2.5-Pro+Grounding* both show substantial improvements (+9.3% and +5.3%), while *GPT-4.1+Web* shows reduced willingness to answer from 100.0% to 92.0%. Our results indicate that web augmentation effectiveness is highly model-dependent, with architectural differences in retrieval processing creating divergent outcomes for local knowledge reasoning tasks, depending on the model being used.

Scaling Analysis

We find that local knowledge reasoning shows limited or even negative scaling effects. Within the Qwen family of models, performance gains taper off beyond 32B parameters and even regress at larger scales. This breakdown is particularly evident when comparing mixture-of-experts (MoE) and non-MoE models. Within the Qwen series, the 235B MoE model underperforms not only its distilled 30B counterpart but also the smaller 32B and 14B non-MoE models. These results suggest that MoE architectures may struggle to encode spatial and contextual nuances necessary for local knowledge.

We also find that the limitations of scaling are most acute for numerical local knowledge. Even state-of-the-art proprietary models achieve near-zero accuracy on such questions ranging from 2.2% to 15.5%. This suggests a qualitative limitation in model design rather than a lack of training data or parameter count: current models appear fundamentally constrained in their ability to represent place-specific quantitative reasoning. More broadly, these results suggest that local knowledge reasoning is not constrained by model size itself, but by fundamental mismatches between transformer architectures and the structured, spatially grounded nature of local information. Future progress will require models with explicit geographic reasoning capabilities and place-aware representations.

Discussion

Our evaluation reveals that local knowledge reasoning represents a qualitatively different challenge for current LLMs. The systematic breakdown across numerical reasoning, cultural contextualization, and scaling relationships suggests fundamental architectural limitations rather than simple knowledge gaps.

The Architectural Mismatch Problem

The failure of numerical reasoning may indicate deeper issues than "mere" memory limitations. Current transformer architectures, optimized for next-token prediction on global text corpora, appear fundamentally misaligned with the precision required for quantitative local reasoning. This mismatch is compounded by the situated nature of local knowledge situated nature, where factual claims depend on geographic, temporal, and cultural — to this point, human — context. Current architectures lack explicit mechanisms for spatial reasoning, temporal grounding, or cultural contextualization, these are capabilities essential for local knowledge understanding. The MoE routing failures further demonstrate that scaling expert capacity do not, today, compensate for these representational gaps.

The Retrieval Integration Paradox

The opposing effects of web augmentation across model families reveal that retrieval integration depends heavily on underlying architectural capabilities rather than search quality alone. The instable augmentation effectiveness suggesting differential capacity for filtering noisy content and maintaining uncertainty calibration when external information conflicts with parametric knowledge. This set of results challenge the prevailing assumption that better search algorithms will help solve grounding problems. Instead, it may be that the challenge lies in developing architectures that can effectively discriminate between high- and low-quality local sources of information and synthesize information across heterogeneous community knowledge systems.

Implications for Future Work

These findings argue for fundamental shifts in AI development priorities. Rather than pursuing larger models trained on comprehensive web corpora, the field must invest in architectures specifically designed for situated reasoning, including spatial reasoning capabilities, cultural contextualization mechanisms, and uncertainty estimation methods that handle local knowledge ambiguity. Importantly, the breakdown of conventional scaling laws for local knowledge tasks indicates that current paradigms cannot address place-based reasoning through scale alone. This may necessitate architectural innovations, including geographic reasoning modules and place-aware attention mechanisms.

Limitations

LocalBench is constructed from data spanning 2020 to 2025 across 526 U.S. counties. While this scope enables systematic benchmarking within a well-documented administrative structure, it also introduces important constraints. Most notably, the benchmark reflects only U.S.-based geographic, cultural, and institutional knowledge. As such, its findings may not generalize to local contexts in other countries, where governance structures, data availability, and sociolinguistic norms differ significantly. Extending local knowledge benchmarks internationally will require adapting both the conceptual framework and data sourcing strategies to diverse geopolitical settings.

Within the U.S., uneven digital footprints persist. Rural counties with limited online presence remain underrepresented (Yin, Guo, and Thebault-Spieker 2024; Thebault-Spieker et al. 2018; Thebault-Spieker, Hecht, and Terveen 2018; Johnson et al. 2016), mirroring broader structural disparities that LocalBench cannot fully resolve. In addition, our exclusive reliance on English-language sources restricts coverage of multilingual communities (Hickman et al. 2021), including Indigenous, immigrant, and border populations whose local knowledge may be expressed in Spanish, Native languages, or other vernaculars not captured here.

While efforts were made to ensure data quality, some QA pairs may still contain factual inaccuracies or temporal mismatches, especially in regions with sparse digital activity. Although our evaluation framework integrates multiple metrics and GPT-based judgment (validated against human annotators) it may still miss nuances of cultural appropriateness or locally grounded truth, highlighting the need for complementary human-in-the-loop approaches in future evaluations.

Conclusion

LocalBench demonstrates that local knowledge reasoning poses a fundamental challenge for current LLMs—one that cannot be solved by scale or retrieval alone. Our evaluation across 14,782 QA pairs and 526 U.S. counties uncovers persistent failures in numerical reasoning, cultural understanding, and confidence calibration, revealing architectural and representational mismatches. These limitations point to the need for spatially grounded models, better uncertainty handling, and participatory development practices that center community epistemologies. The observed gap between model confidence and accuracy is especially salient given prevailing conversations about the benefits and societal importance of AI, and our results suggest this is not true for all communities. While LocalBench offers a foundation for diagnosing these gaps, advancing equitable place-aware AI will require new architectures, ethical frameworks, and sustained collaboration between researchers, communities, and policymakers.

References

Almeida, T. S.; Bonás, G. K.; Santos, J. G. A.; Abonizio, H.; and Nogueira, R. 2025. TiEBe: Tracking Language Model Recall of Notable Worldwide Events Through Time. arXiv:2501.07482.

Anthropic. 2025. Models overview – Anthropic. https://docs.anthropic.com/en/docs/about-claude/models/overview. Accessed: 2025-08-02.

Aubin Le Quéré, M.; Naaman, M.; and Fields, J. 2024. Not quite filling the void: Comparing the perceptions of local online groups and local media pages on Facebook. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1): 1–22.

Bland, J. M.; and Altman, D. G. 1995. Multiple significance tests: the Bonferroni method. *Bmj*, 310(6973): 170.

Chen, A.; Ge, X.; Fu, Z.; Xiao, Y.; and Chen, J. 2024. TravelAgent: An AI Assistant for Personalized Travel Planning. *CoRR*, abs/2409.08069.

COHEN, J. 1992. A power primer. *Psychological bulletin*, 112(1): 155–159.

Dudy, S.; Tholeti, T.; Ramachandranpillai, R.; Ali, M.; Li, T. J.-J.; and Baeza-Yates, R. 2025. Unequal Opportunities: Examining the Bias in Geographical Recommendations by Large Language Models. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, 1499–1516.

Gao, Z.; Cranshaw, J.; and Thebault-Spieker, J. 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): 503:1–503:31.

Gao, Z.; Cranshaw, J.; and Thebault-Spieker, J. 2025. A Turing Test for "Localness": Conceptualizing, Defining, and Recognizing Localness in People and Machines. arXiv:2505.07282.

Google. 2025. Gemini models - Gemini API - Google AI for Developers. https://ai.google.dev/gemini-api/docs/models. Accessed: 2025-08-02.

Guridi, J. A.; Cheyre, C.; and Yang, Q. 2025. Thoughtful adoption of nlp for civic participation: Understanding differences among policymakers. *Proceedings of the ACM on Human-Computer Interaction*, 9(2): 1–27.

Gurnee, W.; and Tegmark, M. 2024. Language Models Represent Space and Time. arXiv:2310.02207.

Hickman, M. G.; Pasad, V.; Sanghavi, H. K.; Thebault-Spieker, J.; and Lee, S. W. 2021. Understanding Wikipedia Practices Through Hindi, Urdu, and English Takes on an Evolving Regional Conflict. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 34:1–34:31.

Horne, B. D.; Gruppi, M.; Joseph, K.; Green, J.; Wihbey, J. P.; and Adalı, S. 2022. NELA-Local: A dataset of US local news articles for the study of county-level news ecosystems. In *Proceedings of the international AAAI conference on web and social media*, volume 16, 1275–1284.

Jang, K. M.; Chen, J.; Kang, Y.; Kim, J.; Lee, J.; Duarte, F.; and Ratti, C. 2024. Place Identity: A Generative AI's Perspective. *Humanities and Social Sciences Communications*, 11(1): 1156.

Johnson, I. L.; Lin, Y.; Li, T. J.-J.; Hall, A.; Halfaker, A.; Schöning, J.; and Hecht, B. 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*, 13–25.

Lengen, C.; and Kistemann, T. 2012. Sense of place and place identity: Review of neuroscientific evidence. *Health & place*, 18(5): 1162–1171.

MacVey, M. 2022. AI & Local News Newsletter, Issue 12. https://engineering.nyu.edu/news/ai-local-news-newsletter-issue-12. Accessed July 16, 2025.

Mai, G.; Lao, N.; He, Y.; Song, J.; and Ermon, S. 2023. Csp: Self-supervised contrastive spatial pre-training for geospatial-visual representations. In *International Conference on Machine Learning*, 23498–23515. PMLR.

Manvi, R.; Khanna, S.; Burke, M.; Lobell, D.; and Ermon, S. 2024a. Large Language Models Are Geographically Biased. arXiv:2402.02680.

Manvi, R.; Khanna, S.; Mai, G.; Burke, M.; Lobell, D. B.; and Ermon, S. 2024b. GeoLLM: Extracting Geospatial Knowledge from Large Language Models. In *ICLR*.

Moayeri, M.; Tabassi, E.; and Feizi, S. 2024. WorldBench: Quantifying Geographic Disparities in LLM Factual Recall. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, 1211–1228. New York, NY, USA: Association for Computing Machinery.

OpenAI. 2025. Models – OpenAI API. https://platform.openai.com/docs/models. Accessed: 2025-08-02.

Park, S.; Kim, Y.; Lee, U.; and Ackerman, M. S. 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 (MobileHCI '14)*, 13–22. Association for Computing Machinery.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 53728–53741. Curran Associates, Inc.

Shi, J.; Li, S.; Xu, Y.; Wang, X.; Fu, R.; Ma, Z.; and Wu, T. 2025. Libra: Synergizing CUDA and Tensor Cores for High-Performance Sparse Matrix Multiplication. *arXiv preprint arXiv:2506.22714*.

Sun, K.; Xu, Y. E.; Zha, H.; Liu, Y.; and Dong, X. L. 2023. Head-to-Tail: How Knowledgeable Are Large Language Models (LLMs)? A.K.A. Will LLMs Replace Knowledge Graphs? arXiv:2308.10168.

Tang, Y.; Wang, Z.; Qu, A.; Yan, Y.; Wu, Z.; Zhuang, D.; Kai, J.; Hou, K.; Guo, X.; Zhao, J.; Zhao, Z.; and Ma, W. 2024. ItiNera: Integrating Spatial Optimization with Large Language Models for Open-domain Urban Itinerary Planning. In Dernoncourt, F.; Preoţiuc-Pietro, D.; and Shimorina, A., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 1413–1432. Miami, Florida, US: Association for Computational Linguistics.

Thebault-Spieker, J.; Halfaker, A.; Terveen, L. G.; and Hecht, B. 2018. Distance and Attraction: Gravity Models for Geographic Content Production. In *Proceedings of the 36th Annual ACM Conference on Human Factors in Computing Systems*, 13.

Thebault-Spieker, J.; Hecht, B.; and Terveen, L. 2018. Geographic Biases Are 'Born, Not Made': Exploring Contributors' Spatiotemporal Behavior in OpenStreetMap. In *Proceedings of the 2018 ACM Conference on Supporting Groupwork*, GROUP '18, 71–82. New York, NY, USA: ACM. ISBN 978-1-4503-5562-9.

Wu, N.; Cao, Q.; Wang, Z.; Liu, Z.; Qi, Y.; Zhang, J.; Ni, J.; Yao, X.; Ma, H.; Mu, L.; et al. 2024. Torchspatial: A location encoding framework and benchmark for spatial representation learning. *Advances in Neural Information Processing Systems*, 37: 81437–81460.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yin, Y.; Guo, L.; and Thebault-Spieker, J. 2024. Productivity or Equity? Tradeoffs in Volunteer Microtasking in Humanitarian OpenStreetMap. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1): 113:1–113:34.

Zhang, Y.; Wei, C.; Wu, S.; He, Z.; and Yu, W. 2023. GeoGPT: Understanding and Processing Geospatial Tasks through an Autonomous GPT. arXiv:2307.07930.

Zhu, S.; Wang, W.; and Liu, Y. 2024. Quite Good, but Not Enough: Nationality Bias in Large Language Models — A Case Study of ChatGPT. arXiv:2405.06996.

Appendix 1: Reddit Data QA Generation Prompt

```

You are given a single Reddit thread (post + comments) from a *local* subreddit of {county, state}.

## From it, you will generate ONE QA pair that is:
- Clear, locally grounded about {county, state}, and answerable with one correct response.
- Either:
-- Track A: Fact-based Local QA (grounded in verifiable, non-subjective local information)
-- Track B: Community Insight QA (reflects clearly shared local community experiences)

## STRICT QA RULES:
1. **NO THREAD-REFERENTIAL QUESTIONS.**
- The QA pair should stand alone without relying on Reddit as a quoted source of authority or opinion.
- Forbidden phrases include: ’’was mentioned as...’’, ’’according to the thread/discussion/Reddit post/comment/commenter’’, ’’what did commenters think’’, etc.
- Litmus test: *Would the question still make perfect sense to someone who never sees the thread?* If not, reject it.
2. FACT-BASED ONLY: Question MUST has **exactly one correct, fact-based answer.
- Prefer using the post’s original question if applicable.
- Include time/place qualifiers for precision.
- It does **not** depend on user preferences, multiple viewpoints, or recommendations
3. AVOID all of these:
- Questions asking how people felt, or vague "why" questions.
- Any framing that implies subjectivity, speculation, or unconfirmed information.
3. CONTEXT must:
- Be 2 to 4 neutral sentences summarizing thread purpose.
- Mention exact date (or month), year, county, and state.
- Attribute information sources.
- Do **not** leak answer.
4. ANSWER must:
- Track A: one exact, verifiable fact (from post or multiple comments).
- Track B: a consensus summary if multiple users echoed same view.
5. Prefer Track A. Use Track B only if consensus is unambiguous.
6. No meta-Reddit content (karma, mods, usernames).
7. Privacy & Safety: no doxxing, no private phone numbers.

## POST DETAILS:
Post Title: {post_title}
Post Information:
- Date: {date}
- County: {county}
- State: {state}
Post Content:
{post_content}
Comments:
{post_comments}

## OUTPUT FORMAT:
[PAIR1]
Question: <one clearly answerable question>
Context: <neutral summary with time and location>
Answer: <fact or consensus summary>
Selected Comments: <e.g., 3,7,12>
Pair_type: <fact|insight>

## EXAMPLES OF ACCEPTABLE QUESTIONS:
- What city program was in contract phase in March 2024 in Broomfield, Colorado?
- Which consulting firm was hired in June 2025 to evaluate Athens’ zoning?
- What fungus linked to bird feces raised public health concern in Honolulu?
- Bad questions that are not acceptable: What do residents want? What do recent discussions say? Which business do people like? What do some suggest?

```

```
## HARD STOPS. Do NOT generate a QA if:
- The answer is not clearly stated or verifiable
- No single correct answer can be derived
- Content is entirely opinion-based or speculative
```

Appendix 2: News Data QA Generation Prompt

```

You are given a local news article.

From it, you will generate ONE QA pair that is:
- Clear, locally grounded, and answerable with one correct response.
- Either:
-- Track A: Fact-based Local QA (grounded in verifiable, non-subjective local information)
-- Track B: Community Insight QA (reflects clearly shared local community experiences)

## ARTICLE DETAILS:
Article Title: {metadata[’title’]}
Article Information:
- Date: {metadata[’date’]}
- County: {metadata[’county’]}
- State: {metadata[’state’]}
- Source: {metadata[’source’]}
- Article Content (Factual Sentences): {news_article}

## STRICT QA RULES:
1. FACT-BASED ONLY: Question must have ONLY **one correct answer**, and the **only answer** is clearly supported by the article content. No subjectivity.
- Good Example: \"What program was paused in March 2024?\"
- NOT: \"What platform was recommended?\" or \"What did residents suggest?\"
- Include time/place qualifiers for precision.
2. AVOID these:
- Questions asking how people felt, what they recommended, or vague \why\ questions.
- Any framing that implies subjectivity, speculation, or unconfirmed information.
- Questions about the news article itself (e.g., \"What does the article discuss?\", \"according to the article\", etc.).
3. CONTEXT must:
- Be 2-4 neutral sentences summarizing article purpose.
- Mention exact date (or month/year), county, and state.
- Attribute information sources (\’’The article reported...\’’, \’’Officials noted...\’’).
- Do **not** leak answer.
4. ANSWER must:
- Track A: one exact, verifiable fact from the article.
- Track B: a consensus summary if multiple sources echoed same view.
5. Track A is **preferred**. Only use Track B if consensus is unambiguous.
6. No meta-content (sources, journalists, publication details).
7. Privacy & Safety: no doxxing, no private phone numbers.

## OUTPUT FORMAT:
[PAIR],
Question: <one clearly answerable question>,
Context: <neutral summary with time and location>,
Answer: <fact or consensus summary>,
Selected Sentences: <e.g., 3, 7, 12>,
Pair_type: <fact|insight>,

## EXAMPLES OF ACCEPTABLE QUESTIONS:
- What city program was in contract phase in March 2024 in Broomfield, Colorado?,
- Which consulting firm was hired in June 2025 to evaluate Athens’ zoning?,
- What public health concern was identified in Honolulu in 2024?,
- NOT: What do residents want? What do recent discussions say? Which business do people like?,

## HARD STOPS. Do NOT generate a QA if:,
- The answer is not clearly stated or verifiable,
- No single correct answer can be derived,
- Content is entirely opinion-based or speculative,
```

Appendix 3: Quality Assessment Criteria

The DPO-tuned quality analyzer evaluates generated QA pairs against the following nine criteria:

1. **Single Factual Answer**: The question has one clear, factual answer that can be verified from the source material. Avoids questions with multiple valid interpretations or subjective responses.

2. **Geographic Grounding**: The question and answer are specifically tied to the mentioned county. Generic questions that could apply to any location are rejected.

3. **Subjectivity Detection**: The question and answer avoid subjective language, personal opinions, or value judgments. Focuses on factual, verifiable information.

4. **Privacy Compliance**: No personal identifiers, private information, phone numbers, addresses, or individual names are included in the QA pair.

5. **Safety Compliance**: Content does not promote harmful activities, contain offensive language, or include sensitive political commentary that could cause harm.

6. **Temporal Consistency**: The question and answer are consistent with the time period of the source material. Avoids anachronistic references or outdated information presented as current.

7. **Difficulty Assessment**: The question is neither trivial (answerable with basic general knowledge) nor impossibly difficult (requiring highly specialized expertise). Appropriate for evaluating local knowledge reasoning.

8. **Question Clarity**: The question is clearly formulated, unambiguous, and can be understood without additional context beyond the county specification.

9. **Answer Completeness**: The answer adequately addresses the question with sufficient detail while remaining concise. Avoids incomplete or overly brief responses that don't fully answer the question.

Appendix 4: Filters Accuracy

Table 3 show each quality filter's performance against human annotation.

Task	**Accuracy**	**Recall**	**F1**
Single Factual Answer	94.2	96.1	0.95
Geographic Grounding	96.8	94.3	0.96
Subjectivity Detection	89.3	91.7	0.91
Privacy Compliance	98.1	99.2	0.99
Safety Compliance	97.4	98.6	0.98
Temporal Consistency	92.7	89.8	0.91
Difficulty Assessment	88.6	92.4	0.90
Question Clarity	93.5	95.1	0.94
Answer Completeness	90.8	93.6	0.92
Overall	**93.4**	**94.5**	**0.94**

Table 3: Quality filters performance against human annotation.

Appendix 5: Task Correlation Analysis

We computed pairwise correlations between quality task failures to quantify their independence. Table 4 shows the results.

Task	SFA	GG	SD	PC	SC	TC	DA	QC
Geographic Grounding	0.12	–						
Subjectivity Detection	0.31	0.08	–					
Privacy Compliance	0.02	0.01	0.05	–				
Safety Compliance	0.04	0.03	0.15	0.18	–			
Temporal Consistency	0.18	0.22	0.09	0.01	0.02	–		
Difficulty Assessment	0.25	0.19	0.14	0.03	0.06	0.11	–	
Question Clarity	0.28	0.16	0.20	0.05	0.07	0.13	0.33	–
Answer Completeness	0.22	0.11	0.17	0.04	0.08	0.15	0.29	0.35

Table 4: Pearson correlation coefficients between quality task failures (SFA=Single Factual Answer, GG=Geographic Grounding, SD=Subjectivity Detection, PC=Privacy Compliance, SC=Safety Compliance, TC=Temporal Consistency, DA=Difficulty Assessment, QC=Question Clarity). Low correlations indicate task independence.

The low correlation coefficients (mean $r = 0.14$, max $r = 0.35$) confirm that quality assessment tasks capture largely independent failure modes, justifying the multi-criteria approach.

Appendix 6: GPT Judge Prompts

```
Evaluate if the AI-generated answer is correct based on the question and golden answer.

- Question: {question}
- Context: {context}
- Golden Answer: {gold_answer}
- AI Answer: {pred_answer}

## Instructions:
- Answer "Yes" if the AI answer is factually correct and addresses the question
- Answer "No" if the AI answer is factually incorrect or doesn't address the question
- Consider partial credit for answers that are mostly correct
- Ignore minor wording differences if the core meaning is correct

Answer (Yes/No):
```

Appendix 7: Iterative QA Re-generation Process

Feedback-Driven Re-generation

When QA pairs fail quality assessment, they are returned to the QA Generator with targeted feedback. The re-generation prompt incorporates specific guidance based on the failed criteria:

Example Re-generation Feedback Original Failed QA Pair:

- Q: “What do people think about the schools in Adams County?”
- A: “Most residents are satisfied with the educational quality.”
- **Failed Criteria**: Single Factual Answer, Subjectivity Detection

Re-generation Prompt:

```
\texttt{The previous QA pair failed quality assessment for the following reasons:
- Question allows multiple valid interpretations and subjective responses
- Answer contains opinion-based language without factual grounding

Please generate a new QA pair that:
1. Has a single, factual answer verifiable from the source
2. Avoids subjective language and opinion-based content
3. Focuses on measurable aspects of education in Adams County

Use the same source material: [SOURCE\_CONTENT]}
```

Re-generation Success Analysis

Analysis of re-generation patterns reveals that different failure types have varying recovery rates.

Privacy and safety compliance issues show the highest recovery rates, as these typically involve simple content modifications or redaction. Conversely, difficulty assessment and subjectivity detection prove most challenging to address, often requiring fundamental reconceptualization of the question-answer relationship. Geographic grounding issues show moderate recovery rates, as they usually require restructuring questions to include county-specific context rather than generic phrasing.

Appendix 8: Localness Annotation Protocol

LLM Classification Prompt (Localness Dimension Part)

```
You are given a question, a supporting context passage, and an answer --- all derived from a local news article. Your task is to identify **1 to 5 dimensions** from the list below that best describe the **main localness themes** expressed in the QA pair.

---
**Question:** {question}
**Context:** {context}
**Answer:** {answer}
---

**INSTRUCTIONS**:
1. From the **dimension list** below, select **at least 1 and at most 5** dimensions that best characterize what this QA pair is about.
   - Only choose dimensions that are clearly relevant to the content.
   - If unsure, select the most general applicable dimension.
   - Do **not** default to the first few options: **read and consider the full list** before deciding.
2. You **must copy each dimension name exactly as written** below.
   - No paraphrasing, abbreviations, or extra characters.
   - Any deviations will result in invalid responses.
3. **Rank the selected dimensions by relevance**, placing the most relevant one first.
4. **Output format**:
   - One dimension per line.
   - No numbering, no bullet points, no explanation.
   - Output must contain **only valid dimension names**, exactly as listed.

---
**DIMENSION LIST** (unordered):
1. Cultural Understanding Dimension: Involves knowledge of symbolic, linguistic, and behavioral expressions that define local identity. This includes understanding local customs, language varieties, and cultural practices such as cuisine and social norms that shape how people interact and express belonging.
2. Environmental Cognition Dimension: Captures mental representations of the physical and ecological characteristics of a place. This dimension focuses on how people understand the local geography, natural environment, and ecological systems through observation, learning, and reflection.
3. Local Knowledge Dimension: Encompasses accumulated, context-specific information that informs decision-making and place literacy. This includes understanding historical developments, insider tips, changes over time, wayfinding skills, and the ability to offer localized recommendations.
4. Place Interaction Dimension: Reflects embodied experience and direct physical engagement with the natural and built environment. This involves a lived connection with ecological features, comfort navigating local spaces, and sensory familiarity with landscapes and natural elements.
5. Temporal Presence Dimension: Represents sustained residence or early-life connection that embeds individuals in the timeline of a place. This dimension emphasizes the depth of familiarity, emotional investment, and continuity afforded by being born in, growing up in, or living long-term in a location.
6. Emotional Connection Dimension: Highlights affective bonds and the emotional significance of a place in personal life. It captures feelings of comfort, identity alignment, and deep attachment that make a place feel like ``home'' and contribute to a stable sense of belonging.
7. Social and Community Engagement Dimension: Centers on interpersonal relationships, civic involvement, and contribution to communal life. This includes participating in events, engaging with local institutions, forming strong local networks, and expressing care through long-term commitments and civic action.
```

Appendix 9: Non-numerical QA Model Comparison

Model	GPT Judge Acc.	95% CI	Effect Size (d)	Ans Rate	Conf Corr
Proprietary Models					
GPT-4.1	47.0	[46.1, 47.7]	–	100.0	0.294
GPT-4o	32.8	[32.6, 34.2]	-0.412**	99.6	0.434
Gemini-2.5-Pro	52.5	[51.9, 52.7]	+0.485**	100.0	-0.090
Claude-Sonnet-4	39.7	[39.3, 43.1]	-0.203**	100.0	0.347
Web-Augmented					
GPT-4.1+Web	35.6	[35.2, 38.0]	-0.289**	92.9	0.369
Gemini-2.5-Pro+Grounding	**56.8**	[55.8, 58.6]	+0.485**	91.7	0.211
Open-Source Models					
Qwen3-235B-A22B	27.3	[26.7, 31.3]	-0.573**	99.3	0.228
Qwen3-30B-A3B	28.0	[27.4, 31.0]	-0.581**	99.7	0.247

Table 5: Statistical comparison for non-numerical QA tasks. Effect sizes are relative to GPT-4.1. **$p < 0.001$ under Bonferroni correction.

Appendix 10: Numerical QA Model Comparison

Model	Accuracy	95% CI	Effect Size (d)	Ans Rate
Proprietary Models				
GPT-4.1	6.2	[4.9, 7.5]	–	100.0
GPT-4o	6.2	[4.8, 7.6]	0.000	39.8
Gemini-2.5-Pro	12.8	[10.7, 14.9]	+0.452**	100.0
Claude-Sonnet-4	7.1	[5.6, 8.6]	+0.078	97.3
Web-Augmented				
GPT-4.1+Web	**15.5**	[13.1, 17.9]	+0.623**	92.0
Gemini-2.5-Pro+Grounding	12.8	[10.6, 15.0]	+0.452**	100.0
Open-Source Models				
Qwen3-235B-A22B	6.6	[5.0, 8.2]	+0.035	77.0
Qwen3-30B-A3B	2.2	[1.2, 3.2]	-0.342**	100.0

Table 6: Statistical comparison for numerical QA tasks. Effect sizes are relative to GPT-4.1. **$p < 0.001$ under Bonferroni correction.

Appendix 11: Generation Success Rates

Success rates calculated against respective base populations for each generation round, See Table 7.

Source/Attempt	Success Rate	QA Pairs	Base	Success (%)
Census Data	100%	6,120	6,120	100.0
Reddit + News Initial	–	8,240	9,107	87.6
First Regeneration	45.2%	527	9,107	93.2
Second Regeneration	22%	141	9,107	94.7
Third Regeneration	9.4%	47	9,107	95.2
Non-Census Subtotal	–	**8,662**	**9,107**	**95.1**
All Sources Combined	–	**14,782**	**15,227**	**96.9**

Table 7: Success rates calculated against respective base populations.

Appendix 12: Metric Details

Metric	Purpose
Answer Rate	Proportion of non-“I don’t know” responses, reflecting model engagement.
Exact Match (EM)	Measures strict string equality after normalization.
ROUGE-1 F1	Measures surface-level similarity tolerant of paraphrasing.
Semantic Match	Embedding-based cosine similarity using `text-embedding-3-small`.
Numerical Accuracy Score	Correct if its relative error is under 2% of the gold value.
GPT Judge Accuracy	Binary correctness classification by GPT-4o-mini.
GPT Judge Confidence	Log-probability of GPT-4o-mini’s correctness token (“Correct”/“Incorrect”).
Model Self-Confidence	Scalar confidence reported by the model for its own answer.
Confidence Correlation	Pearson correlation between model confidence and GPT Judge accuracy.

Table 8: Evaluation metrics used in LocalBench, including correctness, semantic similarity, numeric accuracy, and confidence-based measures.

Appendix 13: Additional Implementation Details

Reddit Data Processing:

1. Extract posts and comments using PRAW with rate limiting (1 request/second)
2. Filter content by score threshold (posts: ≥ 5, comments: ≥ 3)
3. Remove deleted/removed content and moderator posts
4. Anonymize usernames and remove personal identifiers
5. Combine post text with top-50 comments for context

News Data Processing:

1. Filter NELA-Local articles by county-level geographic tags
2. Remove duplicate articles (cosine similarity ≥ 0.9)
3. Extract article text and publication metadata
4. Verify county association through location entity recognition
5. Sample articles to balance geographic distribution

Census Data Processing:

1. Download datasets from varies resources (See Table 9)
2. Standardize indicator names and units across data sources
3. Remove all counties with missing value
4. Validate data consistency across multiple census tables
5. Generate metadata descriptions for each indicator

Appendix 14: Census Metrics and Sources

Table 9: Localness metric taxonomy with metric definitions and data sources.

Domain	Dimension	Metrics	Data Source
Cognitive	Cultural	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
Cognitive	Cultural	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
Cognitive	Cultural	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
Cognitive	Cultural	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

Continued on next page

Table 9: Localness metric taxonomy with metric definitions and data sources.

Domain	Dimension	Metrics	Data Source
Cognitive	Environmental	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
Cognitive	Environmental	In 2022, the percentage of cropland fertilized in this county	USDA National Agricultural Statistics Service
Cognitive	Knowledge	The change in multifamily building permits from 2021 to 2022 in this county	U.S. Census Bureau Building Permits Survey
Cognitive	Knowledge	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
Cognitive	Knowledge	As of 2024, the number of historic preservation properties with local significance in this county	National Register of Historic Places
Cognitive	Knowledge	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
Cognitive	Knowledge	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
Cognitive	Knowledge	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
Cognitive	Knowledge	In 2022, the number of public libraries in this county	Public Libraries Survey (PLS) by the Institute of Museum and Library Services (IMLS)
Cognitive	Knowledge	In 2022, the mean travel time to work for residents of this county	U.S. Census Bureau ACS Table S0801
Cognitive	Knowledge	In 2022, the percentage of workers who use public transportation to work in this county	U.S. Census Bureau ACS Table S0801
Physical	Place-Interaction	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
Physical	Place-Interaction	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
Physical	Temporal	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
Physical	Temporal	In 2022, the percentage of the population in this county identifying as Native American	U.S. Census Bureau ACS Table DP05
Physical	Temporal	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
Physical	Temporal	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
Physical	Temporal	In 2022, the median move-in year of householders in owner-occupied units in this county	U.S. Census Bureau ACS Table B25039
Physical	Temporal	In 2022, the percentage of occupied housing units that were owner-occupied in this county	U.S. Census Bureau ACS Table DP04
Physical	Temporal	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
Relational	Emotional	In 2022, the number of residents in this county who reported ’American’ ancestry	U.S. Census Bureau ACS Table DP02
Relational	Emotional	In 2022, the percentage of residents in this county who were Hispanic or Latino (of any race)	U.S. Census Bureau ACS Table DP05
Relational	Emotional	In 2022, the ethnolinguistic fractionalization index of residents in this county	U.S. Census Bureau ACS Table DP05

Continued on next page

Table 9: Localness metric taxonomy with metric definitions and data sources.

Domain	Dimension	Metrics	Data Source
Relational	Emotional	In 2022, the percentage of residents in this county with zero components of social vulnerability	Community Resilience Estimates Datasets
Relational	Social/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
Relational	Social/Community	In the 2020 presidential election, the total number of votes cast in this county	County Presidential Election Returns 2000-2020
Relational	Social/Community	In 2022, the percentage of owner-occupied housing units with a mortgage in this county	ACS Table DP05
Relational	Social/Community	In 2018, the density of nonemployer businesses per 1,000 residents in this county	U.S. Census Bureau Nonemployer Statistics Table NS1800NONEMP 2018
Relational	Social/Community	In 2022, the average size of married-couple households in this county	U.S. Census Bureau ACS Table DP02
Relational	Social/Community	In 2022, the average household size in this county	U.S. Census Bureau ACS DP04