Assessing LLM Reliability on Temporally Recent Open-Domain Questions

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are increasingly deployed for open-domain question answering, yet their alignment with human perspectives on temporally recent information remains underexplored. We introduce RECOM (Reddit Evaluation for Correspondence of Models), a benchmark dataset of 15,000 recent Reddit questions from September 2025 paired with community-derived reference answers. We investigate how four open-source LLMs (Llama-3.1-8B, Mistral-7B, Gemma-2-9B, and GPT-OSS-20B) respond to these questions, evaluating alignment using lexical metrics (BLEU, ROUGE), semantic similarity (BERTScore, MoverScore, cosine similarity), and logical inference (NLI). Our central finding is a striking semantic-lexical paradox: all models achieve over 99% cosine similarity with references despite less than 8% BLEU-1 overlap - a 90+ percentage point gap indicating that models preserve meaning through extensive paraphrasing rather than lexical reproduction. MoverScore (51–53%) confirms this pattern, occupying an intermediate position that reflects the optimal transport cost of semantic alignment. Furthermore, model scale does not predict performance: Mistral-7B (7B parameters) outperforms GPT-OSS-20B (20B parameters) across all metrics. NLI analysis reveals that contradiction rates remain below 7%, suggesting models rarely generate content that directly conflicts with human consensus. These findings challenge the reliability of lexical metrics for evaluating abstractive generation and argue for multi-dimensional evaluation frameworks that capture semantic fidelity beyond surface-level text matching. The RECOM dataset is publicly available - https://anonymous.4open.science/r/recom-D4B0

1 Introduction

Large Language Models (LLMs) have demonstrated impressive performance across a wide range of natural language understanding and generation tasks. Their growing adoption in conversational agents, search assistants, and question-answering systems has raised questions about how well these models align with human perspectives, particularly when responding to recent or evolving information. This issue becomes particularly critical when models are applied to recent or rapidly evolving content, where the coverage of training data may be limited.

Most prior work on LLM evaluation focuses on benchmark datasets, static knowledge sources, or synthetic prompts. In contrast, real-world user questions, especially those posed on online platforms such as Reddit, often reflect current events, emerging trends, and evolving community knowledge. These questions present a challenging testbed for evaluating LLM reliability, as correct answers may not be well-established or universally agreed upon.

In this work, we study how LLM-generated answers align with community perspectives on recent Reddit questions from September 2025, leveraging aggregated human responses as reference answers. Instead of relying on a single human-annotated gold answer, we summarize multiple human responses to construct a reference answer that captures diverse perspectives. We then evaluate how closely different LLMs align with this reference when answering the same questions.

Our contributions are as follows:

1. We introduce a novel dataset of 11,515 real-world questions paired with community-derived reference responses, where the questions require more than simple factoids.

2. We compare four open-source LLMs using a comprehensive multi-dimensional evaluation framework spanning lexical metrics (BLEU, ROUGE), semantic similarity (BERTScore, cosine similarity), soft semantic alignment via optimal transport (MoverScore), and logical

inference (NLI), revealing a striking semantic-lexical paradox where models achieve 99%+ semantic similarity despite <8% lexical overlap.

3. We demonstrate that model scale does not predict alignment performance, with Mistral-7B (7B parameters) outperforming GPT-OSS-20B (20B parameters) across all metrics.

4. We show that MoverScore occupies an intermediate position (51–53%) between lexical and embedding-based metrics, providing a more granular view of semantic alignment.

2 Related Work

The reliability of neural text generation has been extensively studied and evaluated, particularly in tasks such as summarization, machine translation, and question answering. A key concern is the generation of fluent but unsupported or factually incorrect content (Maynez et al., 2020; Ji et al., 2023). Prior work distinguishes between intrinsic hallucination, where generated content contradicts available evidence, and extrinsic hallucination, where models introduce unverifiable or fabricated information (Dziri et al., 2022).

Recent research has focused on evaluating factual consistency in LLMs using automatic metrics, entailment-based verification, and prompting strategies. Examples include factual consistency evaluation (Kryściński et al., 2020), question-answering–based verification methods (Honovich et al., 2022), and approaches that encourage models to explicitly express uncertainty (Lin et al., 2022). However, most of these studies rely on static benchmarks or curated datasets, which limits their applicability to real-world dynamic information needs.

Open-domain question answering (ODQA) systems are typically evaluated on benchmark datasets such as Natural Questions (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and WebQuestions, which mainly contain well-established factual knowledge. Although these benchmarks allow for standardized evaluation, they do not capture evolving or recently emerging information. Retrieval-based ODQA methods have shown a strong performance on such benchmarks (Karpukhin et al., 2020; Lewis et al., 2020b), but their evaluation remains largely disconnected from community-driven questions in the real-world.

Community forums such as Reddit have been explored as alternative data sources for evaluating question answering systems due to their linguistic diversity and topical breadth (Zhang et al., 2021; Gao et al., 2022). However, these datasets are less commonly used for systematic alignment evaluation. Our work extends this line of research by measuring how closely LLM responses align with community consensus on recent Reddit questions.

Traditional question answering datasets rely on expert annotators or crowd-workers to create gold-standard answers. However, single-reference annotations often fail to capture ambiguity or legitimate disagreement among humans. Previous work has shown that modeling annotator disagreement and using multiple references can improve evaluation robustness (Geva et al., 2019; Pavlick and Kwiatkowski, 2019).

In contrast to previous studies that rely on static benchmarks or synthetic tasks, we evaluated LLMs on recent, real-world questions sourced from online communities. By combining temporally recent data, consensus-based reference construction, and multi-dimensional evaluation across lexical, semantic, and logical inference metrics, our work reveals that models can achieve high semantic alignment while exhibiting minimal lexical overlap with human-generated content.

3 Methodology

We investigate how well open-source LLMs align with community-derived answers on temporally recent questions. We address three research questions:

1. How closely do model responses align with aggregated human perspectives on post-training-cutoff questions?
2. Do lexical and semantic metrics provide consistent alignment signals?
3. Does model scale correlate with alignment performance?

We operationalize alignment through lexical overlap, semantic similarity, and logical inference metrics.

3.1 Dataset

We introduce RECOM (Reddit Evaluation for Correspondence of Models), a benchmark dataset for evaluating LLM alignment with community perspectives on temporally recent questions.

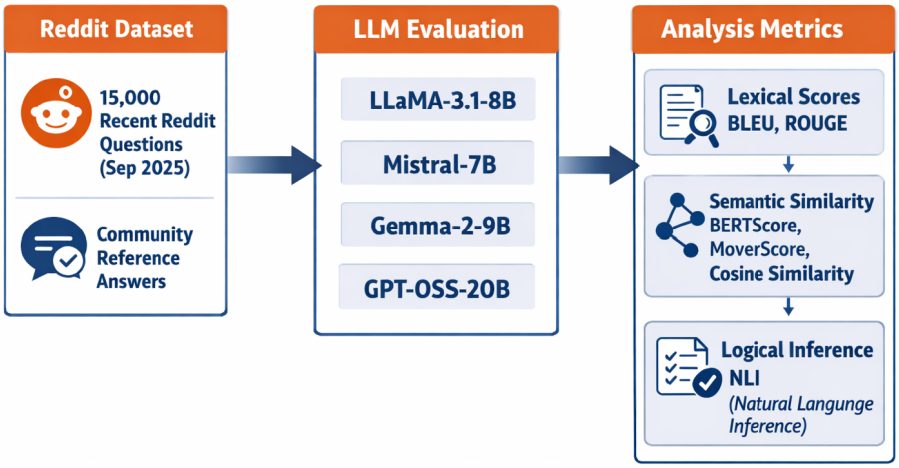

Figure 1: The LLM Alignment Evaluation for RECOM. 1. Extract Reddit questions and summarize human responses to construct reference answers. 2. Four LLMs generate responses to the same questions. 3. Evaluate outputs using lexical, semantic, and NLI metrics.

3.1.1 Data Collection

We collected 132,728 publicly available posts from the community r/AskReddit on the Reddit platform during September 2025. This subreddit provides open-ended questions with multiple human responses across diverse topics, allowing consensus-based reference construction. The temporal window ensures that the questions reflect recent events and information likely absent from the evaluated models' training corpora.

3.1.2 Engagement-Based Filtering and Sampling

To ensure sufficient community input for constructing reliable reference answers, we calculated engagement scores for all 132,728 posts. The engagement score for a post is defined as the total number of comments and nested replies received during the collection period. We retained the top 25,000 highest-engagement posts, reasoning that higher engagement indicates questions that resonated with the community and attracted diverse, substantive responses.

From this filtered set, we randomly sampled 15,000 questions for evaluation. This sampling strategy balances computational tractability with breadth of coverage across question types, while random selection minimizes systematic bias toward particular topics or question styles.

3.1.3 Construction of Reference Answer

For each question, we aggregated human responses and generated a reference answer using Llama-3.1-8B-Instruct via Ollama (default parameters) through summarization prompts. The prompt (Table 1) elicits the central question and a range of perspectives from the responses. This approach approximates community consensus rather than relying on single-annotator judgment. We acknowledge potential self-alignment bias since Llama-3.1-8B is also an evaluated model. We randomly sampled 10% of the answers and had them evaluated by one of the authors.

```
In this task, you are given a Reddit
discussion consisting of one original
post and a collection of replies.
Your goal is to produce a short,
integrated summary that captures the
central question raised in the post
and the main perspectives or arguments
expressed in the replies.
Focus only on the content within the
thread, avoiding any external context.
```

Table 1: Summarization Prompt

3.2 Evaluated Models

We evaluate four open-source LLMs spanning 7B–20B parameters, selected for scale diversity, availability via Ollama, and representation of different model families. All models used default inference parameters.

Methodological Note: Llama-3.1-8B was used for both reference summarization and evaluation, creating potential self-alignment bias. However, the results are agnostic to any self-alignment bias for Llama-3.1-8B as will be observed through the

consistency of the observed outcomes across the different LLMs.

3.3 Response Generation

Each model was prompted with identical instructions (Table 2) designed to elicit direct, concise answers:

```
You are answering this question.
Give a direct answer without
restating or summarizing the question.
Rules:
- Answer ONLY
- Be clear and concise
- Keep it under 50 words
```

Table 2: Answer Generation Prompt

The prompt constraints ("under 50 words," "Answer ONLY") were designed to elicit focused responses comparable across models, avoiding verbose padding that could artificially inflate similarity scores. We set a maximum token limit of 1,024 tokens to accommodate longer responses when necessary.

All inference was conducted on a workstation equipped with an AMD Ryzen Threadripper PRO 3995WX processor (64 cores, 128 threads), 1 TB RAM and four NVIDIA RTX A5000 GPU (24GB VRAM), with models served through Ollama's local inference framework.

3.4 Response Filtering

To ensure that the evaluation focused on substantive model outputs rather than refusal behaviors, we filtered responses containing patterns indicative of the model declining to answer. We implemented regex-based pattern matching to identify responses containing phrases such as:

- Direct AI self-references: "as an AI," "I'm a language model," "as an artificial intelligence"
- Capability disclaimers: "I don't have," "I cannot experience," "I lack emotions"
- Role disclaimers: "I'm not human," "I am not capable of"

After filtering, our final evaluation set comprised 11,515 question-response pairs per model, representing a retention rate of 76.8% of the 15,000 sampled questions. The filtered questions typically involved personal experiences, emotional responses, or physical sensations that the models appropriately declined to simulate.

3.5 Evaluation Framework

We employ a multi-dimensional evaluation framework spanning four complementary perspectives on response alignment. This multi-metric approach enables us to distinguish between surface-level textual similarity and deeper semantic correspondence.

3.5.1 Lexical Overlap Metrics

Lexical metrics measure direct word and phrase overlap between generated responses and reference summaries.

BLEU: We compute BLEU-1 through BLEU-4 scores to measure n-gram precision at multiple granularities. BLEU (Papineni et al., 2002) was originally designed for machine translation evaluation and penalizes outputs that diverge lexically from references, even when semantically equivalent.

ROUGE (Lin, 2004): We report ROUGE-1, ROUGE-2, and ROUGE-L F1 scores to capture recall-oriented lexical similarity. ROUGE-L specifically measures the longest common subsequence, capturing word ordering patterns beyond bag-of-words overlap.

3.5.2 Semantic Similarity Metrics

Semantic metrics assess meaning preservation independent of exact wording.

BERTScore: We compute precision, recall, and F1 using RoBERTa-large embeddings (Liu et al., 2019). BERTScore (Zhang et al., 2019) performs soft token matching using contextual embeddings, rewarding semantic equivalence even when surface forms differ. This metric occupies a middle ground between pure lexical matching and holistic semantic comparison.

MoverScore: We compute MoverScore using contextualized BERT embeddings following (Zhao et al., 2019). Unlike BERTScore's greedy one-to-one token alignment, MoverScore formulates semantic similarity as an optimal transport problem, computing the minimum cost of transforming the token distribution of the generated text into that of the reference using Earth Mover's Distance (EMD). This allows soft many-to-many alignments, capturing cases where meaning is distributed differently across tokens in the two texts.

Cosine Similarity: We compute cosine similarity between mean-pooled RoBERTa-large embeddings of complete generated answers and reference summaries. This metric captures holistic semantic alignment at the document level, abstracting away from token-level correspondences.

3.5.3 Logical Inference Metrics

Natural Language Inference (NLI): (Bowman et al., 2015) We classify each (generated answer, reference summary) pair using facebook/bart-large-mnli (Nie et al., 2020; Lewis et al., 2020a) into three categories: entailment (answer logically follows from summary), contradiction (answer conflicts with summary), or neutral (no clear logical relationship). High contradiction rates may indicate factual inconsistency, while high neutral rates suggest the model generates related but non-derivative content.

3.6 Statistical Analysis

All pairwise model comparisons are evaluated using the Wilcoxon signed-rank test (Casella and Berger, 2024), a non-parametric test appropriate for paired samples that does not assume normal (Gaussian) distributions. This is important given the skewed distribution of metrics like BLEU scores. We report effect sizes using Cohen's d (Cohen, 2013) to distinguish statistically significant differences from practically meaningful ones, interpreting $|d| < 0.2$ as negligible, 0.2–0.5 as small, 0.5–0.8 as medium, and > 0.8 as large effects.

4 Results

We evaluated four open-source LLMs on 11,515 Reddit questions after filtering refusal responses. We report results across four evaluation dimensions: lexical overlap, semantic similarity, embedding-based alignment, and logical inference patterns. All pairwise model differences reported below are statistically significant (Wilcoxon signed-rank test, $p < 0.001$) unless otherwise noted. We additionally report effect sizes (Cohen's d) to contextualize the practical significance of observed differences.

4.1 Lexical Overlap Metrics

Table 3 presents the BLEU and ROUGE scores across the models. All models exhibit low lexical overlap with reference summaries, with BLEU-1 scores ranging from 0.57% to 7.58% and BLEU-4 scores uniformly below 1%.

All values in percentages. Llama-3.1-8B was also used for reference summarization; scores may reflect self-alignment bias.

The scores reveal a consistent model ranking across all lexical metrics: Mistral-7B and Llama-3.1-8B lead, followed by GPT-OSS-20B, with Gemma2-9B trailing substantially. The difference between top and bottom performers is substantial: Mistral-7B achieves 10.8x higher BLEU-1 than Gemma2-9B (6.17% vs. 0.57%), corresponding to a large effect size (Cohen's d = 1.43).

Notably, GPT-OSS-20B's larger parameter count (20B) does not yield higher lexical alignment than the 7-8B models. This finding challenges the assumption that model scale directly translates to improved performance on open-domain questions, at least as measured by lexical overlap metrics.

Threshold Analysis: Only 0.3% of Gemma2-9B responses exceed 10% BLEU-1 overlap, compared to 29.0% for Llama-3.1-8B and 21.7% for Mistral-7B. This distribution indicates that low lexical overlap is not merely a matter of means but reflects a consistent generation strategy across questions.

4.2 Semantic Similarity

Despite low lexical overlap, all models achieve high semantic alignment as measured by BERTScore and cosine similarity (Table 4). This divergence between lexical and semantic metrics constitutes our central finding.

All models achieve cosine similarity above 99%, indicating near-perfect alignment in RoBERTa-large's embedding space. The total range spans only 0.41 percentage points (99.10–99.51%), a dramatic compression compared to the 13-point range observed in ROUGE-1 scores and the 10x range in BLEU-1. BERTScore F1 scores cluster within a narrow 1.54 percentage point range (83.29–84.83%), with Mistral-7B achieving the highest scores. While the BERTScore differences are statistically significant (all pairwise $p < 0.001$), the effect sizes are smaller than those observed for lexical metrics: Mistral-7B versus Gemma2-9B yields d = 0.91 on BERTScore F1 compared to d = 1.50 on ROUGE-1.

MoverScore, which measures semantic similarity via optimal transport over contextualized embeddings, yields scores in the 50.92–53.42% range across models. Unlike cosine similarity and BERTScore which cluster near ceiling values, MoverScore exhibits greater dispersion (2.50 percentage points) while preserving the same model

Model	BLEU-1	BLEU-4	ROUGE-1 F1	ROUGE-L F1
Gemma2-9B	0.57 ± 1.41	0.06 ± 0.24	9.13 ± 6.57	6.89 ± 4.61
GPT-OSS-20B	3.71 ± 4.02	0.29 ± 0.54	14.28 ± 7.80	9.72 ± 4.95
Llama-3.1-8B	7.58 ± 5.27	0.62 ± 1.85	18.81 ± 7.93	11.54 ± 4.97
Mistral-7B	6.17 ± 5.38	0.68 ± 1.18	19.97 ± 7.83	13.16 ± 5.14

Table 3: Lexical Overlap Metrics (mean ± std). BLEU-1 ranges from 0.57% to 7.58%, indicating minimal lexical copying across all models. High standard deviations relative to means reflect heavily skewed distributions, with most responses achieving near-zero n-gram overlap.

Model	BERTScore P	BERTScore R	BERTScore F1	MoverScore	Cosine Sim.
Gemma2-9B	85.55	81.18	83.29	50.92	99.20
GPT-OSS-20B	84.48	82.25	83.33	51.84	99.10
Llama-3.1-8B	84.96	83.14	84.03	52.65	99.47
Mistral-7B	86.20	83.51	84.83	53.42	99.51

Table 4: Semantic Similarity Metrics with 95% confidence intervals. Despite low lexical overlap, all models achieve high semantic alignment: cosine similarity exceeds 99%, BERTScore F1 ranges from 83.3% to 84.8%, and MoverScore ranges from 50.9% to 53.4%. MoverScore's intermediate values reflect the optimal transport cost of token-level semantic alignment. Also used for reference summarization.

ranking: Mistral-7B (53.42%) > Llama-3.1-8B (52.65%) > GPT-OSS-20B (51.84%) > Gemma2-9B (50.92%). The intermediate magnitude of MoverScore substantially higher than lexical metrics yet lower than BERTScore and cosine similarity - reflects its sensitivity to the cost of aligning token-level semantics rather than measuring pure embedding proximity or n-gram overlap. GPT-OSS-20B exhibits the highest MoverScore variance (std = 3.42) with occasional zero-value outliers, consistent with its higher variance observed in cosine similarity (Table 11) and suggesting more erratic response patterns compared to the smaller models.

Precision-Recall Patterns: Gemma2-9B exhibits the largest precision-recall gap (85.55% precision, 81.18% recall, $\Delta = 4.37$ points), suggesting conservative generation that achieves high precision but omits some reference content. Mistral-7B shows a more balanced profile (86.20% precision, 83.51% recall, $\Delta = 2.69$ points), indicating more comprehensive coverage of reference semantics.

Threshold Analysis: 46.2% of Mistral-7B responses exceed 85% BERTScore F1, compared to only 16.4% for Gemma2-9B. For cosine similarity, 59.2% of Mistral-7B responses exceed 99.5%, versus 11.0% for Gemma2-9B. These distributions confirm that Mistral-7B's superior mean scores reflect consistent high performance rather than a few outlier successes.

4.3 Model Ranking and Consistency

Model rankings remain stable across semantic metrics. Mistral-7B achieves highest scores on 51.7% of questions (cosine similarity) and 57.5% (BERTScore F1), with Llama-3.1-8B second. Despite its larger parameter count, GPT-OSS-20B achieves highest alignment on only 8.5% of questions.

Inter-model correlations on BERTScore F1 are moderate (r = 0.21–0.54), indicating questions difficult for one model are not consistently difficult for others. The highest correlation occurs between Llama-3.1-8B and Mistral-7B (r = 0.54). Across all questions, 42.9% show high inter-model agreement (BERTScore std < 0.01), while only 0.3% show substantial disagreement (std > 0.05).

4.4 Logical Inference Patterns

Table 5 presents NLI classification results using BART-large-MNLI.

The majority of response-reference pairs (86–90%) receive neutral classifications, indicating that generated answers are topically related to but neither logically entailed by nor contradictory to the reference summaries. This pattern is consistent with abstractive paraphrasing: models generate semantically related content without strictly deriving it from or contradicting the reference.

Entailment Patterns: Gemma2-9B and Mistral-7B show higher entailment rates (9.04% and 8.71%) despite their different lexical overlap pro-

Model	Entailment	Contradiction	Neutral
Gemma2-9B	9.04%	5.15%	85.81%
GPT-OSS-20B	3.43%	6.94%	89.39%
Llama-3.1-8B	4.06%	5.53%	90.41%
Mistral-7B	8.71%	4.31%	86.98%

Table 5: NLI classification distribution using BART-large-MNLI. Neutral classifications dominate (86–90%), indicating models generate semantically related but non-entailing content. Contradiction rates remain below 7% across all models. Also used for reference summarization.

files. This suggests these models more frequently generate responses that logically follow from the reference content, possibly indicating closer adherence to the factual claims in human responses.

Contradiction Analysis: Overall contradiction rates remain below 7% across all models, suggesting that direct factual conflicts between generated answers and human consensus are relatively rare. However, examining individual responses reveals meaningful variation: (a) 6.1–10.0% of questions elicit at least some contradiction signal (contradiction score > 0), depending on the model (b) GPT-OSS-20B shows the highest rate of strong contradictions (>50% contradiction score) at 5.5% of questions and (c) Mistral-7B shows the lowest rate of strong contradictions at 2.1%. These strong contradiction cases represent instances where model outputs directly conflict with community-expressed perspectives and warrant further investigation.

4.5 The Semantic-Lexical Divergence

Our central finding is a striking divergence between lexical and semantic metrics. Models achieve 99%+ cosine similarity while maintaining less than 8% BLEU-1 overlap—a gap of approximately 90 percentage points.

Taking Mistral-7B as an example: its cosine similarity (99.51%) is 16x higher than its BLEU-1 (6.17%), BERTScore F1 (84.83%) is 14x higher, and MoverScore (53.42%) is 8.7x higher. Effect sizes between Mistral-7B and Gemma2-9B show large effects on ROUGE-1 (d = 1.50) and BERTScore F1 (d = 0.91), but only 0.31% absolute difference on cosine similarity—suggesting models vary in paraphrasing strategies while achieving comparable semantic comprehension.

5 Discussion

Our results reveal fundamental tensions in how we evaluate LLM-generated text and what different metrics actually capture about model capabilities. We examine four key patterns that emerge from this multi-dimensional evaluation and discuss their implications for research and practice.

5.1 The Semantic-Lexical Paradox

Our most striking finding is the dramatic divergence between lexical and semantic evaluation metrics. All four models achieve 99.1–99.5% cosine similarity with reference summaries, yet their lexical overlap remains minimal (BLEU-1: 0.57–7.58%). This 90+ percentage point gap indicates that models capture semantic meaning while using almost entirely different words—effectively performing paraphrase at scale.

This paradox has several important implications. First, it demonstrates that modern LLMs have acquired robust paraphrasing capabilities, transforming reference content into semantically equivalent but lexically distinct expressions. Second, it reveals that high embedding similarity can coexist with near-zero n-gram overlap, challenging the assumption that these metrics measure related constructs.

The semantic metrics themselves form a gradient that illuminates different facets of this paradox. Cosine similarity (99.1–99.5%) captures holistic document-level alignment in embedding space. BERTScore F1 (83.3–84.8%) performs greedy token-level alignment, yielding lower scores that reflect the cost of imperfect one-to-one token matching. MoverScore (50.9–53.4%) sits lower still, as its optimal transport formulation penalizes the cumulative cost of moving semantic mass between token distributions. This gradient—from near-ceiling cosine similarity to moderate MoverScore to minimal BLEU—suggests that the apparent paradox reflects not a binary semantic-lexical divide but a continuum of alignment granularities.

The NLI results add another layer of complexity: despite 99%+ embedding similarity, only 3.4–9.0% of responses show entailment, with 86–90% classified as neutral. This demonstrates that two texts can occupy similar semantic space without one logically following from the other—a key distinction between semantic similarity and logical entailment that has implications for hallucination detection.

5.2 Model Performance Patterns

Model performance does not scale with parameter count. Mistral-7B achieves the strongest results across all metrics, followed closely by Llama-3.1-8B. Despite having 20 billion parameters, GPT-OSS-20B underperforms both smaller models on every metric, while Gemma2-9B lags substantially behind on lexical measures despite competitive semantic alignment.

Critically, all models achieve >99% cosine similarity, indicating they grasp core semantic concepts equally well—the differences emerge in how they express those concepts lexically. This suggests that architectural choices, training data composition, and instruction-tuning strategies matter more than raw parameter count for this task. The finding that 7B models outperform a 20B model challenges simplistic assumptions about scaling laws in abstractive generation contexts.

The inter-model correlation analysis (r = 0.21–0.54 on BERTScore) reveals that questions difficult for one model are not consistently difficult for others. This suggests models have different strengths and weaknesses rather than uniform capability profiles, which has implications for ensemble approaches and model selection.

5.3 Evaluation Metric Implications

Each metric family reveals different aspects of generation quality. BLEU and ROUGE measure surface form correspondence and severely penalize paraphrastic generation, systematically underestimating quality for abstractive tasks. MoverScore (50.9–53.4%) occupies a middle ground via optimal transport, reflecting the cost of redistributing semantic content across different lexical realizations. BERTScore (83–85%) computes token-level semantic similarity, tolerating lexical variation while anchoring to reference structure. Cosine similarity (99%+) captures holistic document-level alignment. NLI metrics capture logical relationships; the 86–90% neutral rates indicate outputs and references are topically related but structurally distinct.

The consistent model ranking across all semantic metrics (Mistral-7B > Llama-3.1-8B > GPT-OSS-20B > Gemma2-9B) suggests they capture a shared underlying construct despite different magnitudes. These findings argue for multi-metric evaluation portfolios rather than single-number summaries.

5.4 Alignment vs. Factual Correctness

Our evaluation measures alignment with community consensus rather than factual correctness. The low contradiction rates (2-6%) suggest models rarely conflict with human perspectives, but high alignment does not guarantee accuracy—models could match popular but incorrect responses. The high neutral NLI rates (86-90%) confirm most outputs discuss related topics differently rather than supporting or contradicting references. Future work should incorporate external knowledge verification to disentangle these factors.

6 Conclusion

We present a large-scale study of LLM alignment with community perspectives on temporally recent questions. By comparing four open-source LLMs against reference answers derived from aggregated human responses, our evaluation reflects realistic information-seeking scenarios using 11,515 Reddit questions from September 2025.

Our central finding is a striking semantic-lexical paradox: models achieve over 99% cosine similarity with references despite less than 8% BLEU-1 overlap, demonstrating that high semantic alignment can coexist with minimal lexical reproduction. BERTScore (83–85%) and MoverScore (51–53%) reveal a gradient of alignment granularities between these extremes. Additionally, model scale does not predict performance—Mistral-7B outperforms the larger GPT-OSS-20B across all metrics, including MoverScore's optimal transport-based evaluation. These results challenge the reliability of lexical metrics for evaluating abstractive generation and argue for multi-dimensional evaluation frameworks combining lexical, semantic, and logical inference measures.

7 Ethical Considerations

This work analyzes publicly available data collected from Reddit, an online platform where users voluntarily post questions and responses. We only use content that is publicly accessible and do not attempt to deanonymize users or infer sensitive personal attributes. To further reduce privacy risks, we do not release raw user identifiers or quote individual posts verbatim in the paper.

The reference answers used in this study are generated by summarizing multiple human responses using a large language model. While this approach

approximates community consensus, it may inadvertently amplify dominant viewpoints or overlook minority perspectives. We acknowledge this risk and emphasize that our reference summaries should not be interpreted as authoritative ground truth.

All evaluated LLMs are used strictly for research purposes. Our analysis focuses on measuring response alignment rather than endorsing or deploying any specific model. The findings should not be interpreted as guarantees of correctness for real-world applications.

Finally, as model-generated content may include inaccurate or misleading information, we avoid presenting example outputs that could cause harm or propagate false claims. Our study is intended to support safer and more reliable use of LLMs by identifying their limitations when answering recent questions.

8 Limitations

Our study has several limitations. First, alignment is measured using automatic similarity metrics (BLEU, ROUGE, BERTScore, and cosine similarity), which capture surface-level and semantic correspondence but not factual correctness. While our multi-metric approach provides robustness, these metrics cannot replace human judgment, especially for ambiguous or opinion-based questions.

Second, reference answers are generated via LLM-based summarization of human responses rather than manual expert annotation. Although this method approximates community consensus, it may introduce summarization bias and propagate errors present in the original responses. Additionally, using Llama-3.1-8B for both summarization and evaluation creates potential self-alignment bias.

Third, our dataset is limited to r/AskReddit posts from September 2025, a subreddit emphasizing open-ended personal questions, opinions, and advice-seeking. Findings may not generalize to technical question-answering, factual retrieval, or domains with objectively verifiable answers. Reddit communities also vary widely in expertise and reliability.

Fourth, we evaluate four open-source models using a single prompting strategy without retrieval augmentation. Performance may differ under alternative prompts, decoding strategies, or retrieval-based settings. The GPT-OSS-20B model has limited documentation, which may affect reproducibility.

Finally, our metrics measure alignment with human consensus rather than factual correctness. Low similarity scores may reflect valid alternative responses rather than errors, while high similarity could mask agreement with incorrect human responses. Future work should incorporate human evaluation and external knowledge verification to disentangle these factors.

References

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 632–642.

George Casella and Roger Berger. 2024. *Statistical inference*. Chapman and Hall/CRC.

Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. routledge.

Nouha Dziri, Ehsan Kamalloo, Hannah Rashkin, Xi Victoria Lin, Yashar Mehdad, and Yejin Choi. 2022. Faithdial: A faithful benchmark for information-seeking dialogue. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2022. Evaluating open-domain question answering in community forums. In *Findings of the Association for Computational Linguistics*.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Towards a model of annotation bias in crowd-sourced text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Or Honovich, Leshem Choshen, Roee Aharoni, and Omri Abend. 2022. Qafacteval: Improved qa-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tianyi Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7871–7880.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, and 1 others. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. In *Transactions of the Association for Computational Linguistics*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4885–4901.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Ellie Pavlick and Tom Kwiatkowski. 2019. Disagreement as a signal for ambiguity in natural language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Sheng Zhang, Yizhou Sun, and Michel Galley. 2021. Redditqa: A large-scale dataset for open-domain question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578.

A Appendix

A.1 Dataset Sample

Table 6 presents representative examples showing the diversity of questions and the nature of community-derived reference answers.

A.2 Illustrative Examples

We present examples spanning the performance spectrum: typical paradox cases (high semantic, low lexical) and edge cases (extractive responses, variance-induced underperformance).

A.2.1 Paradox Examples: High Semantic, Low Lexical

These examples demonstrate our central finding—models achieve near-perfect semantic alignment despite minimal lexical overlap.

Example 1: Unsolved Mystery (Gemma2-9B)

- **Question:** What's the strangest unsolved mystery?
- **Reference Summary:** The replies present various enigmatic cases, including the construction of the pyramids, unexplained disappearances such as Dyatlov Pass, Bermuda Triangle, and US Navy Flight 19 incidents, as well as more mundane yet mystifying occurrences like a missing TV remote.

- **Model Response:** The disappearance of Amelia Earhart in 1937 remains one of history's most perplexing unsolved mysteries.

Metric	Score
BLEU-1	0.27%
ROUGE-L F1	7.92%
BERTScore F1	82.30%
Cosine Similarity	99.52%

Analysis: The model selects one specific mystery (Amelia Earhart) not mentioned in the reference, yet achieves 99.52% cosine similarity. Zero n-gram overlap coexists with near-perfect semantic alignment—the model correctly identifies "unsolved mystery" as the core concept and provides a valid instance.

Example 2: Relationship Advice (Llama-3.1-8B)

- **Question:** What's the most real relationship advice you can give?
- **Reference Summary:** Key themes include: open and honest communication; trusting gut instincts; recognizing that love alone is not enough—compatibility, respect, and shared values are essential; letting go of ego; fostering emotional maturity and self-awareness.
- **Model Response:** Communicate openly and honestly with your partner, but also know when to listen without responding. No one is perfect, so accept each other's flaws and work through issues together with empathy and understanding.

Metric	Score
BLEU-1	0.92%
ROUGE-L F1	10.17%
BERTScore F1	83.36%
Cosine Similarity	99.53%

Analysis: The model distills five bullet-pointed themes into a single coherent sentence capturing the core concepts. The 0.92% BLEU-1 reflects word choice differences ("open communication" $\rightarrow$ "communicate openly"), while 99.53% cosine similarity confirms semantic preservation.

A.2.2 Edge Cases: Extractive Responses

Some questions elicit higher lexical overlap when direct content borrowing is appropriate.

Example 3: Shower Preferences (Mistral-7B)

- **Question:** If you shower once a day, is it in the morning or at night? And how do you feel about that?
- **Reference Summary:** Replies highlighted varied preferences for either morning or evening showers, with some citing benefits such as feeling refreshed, relaxed, or clean, while others mentioned practical considerations like work schedules or exercise routines.
- **Model Response:** Showering once a day can be either in the morning or at night. It's a personal preference and neither is inherently better than the other.

Metric	Score
BLEU-1	22.43%
ROUGE-L F1	30.65%
BERTScore F1	88.70%
Cosine Similarity	99.70%

Analysis: This extractive example shows BLEU-1 at 22.43% (37$\times$ Mistral's mean), indicating substantial lexical overlap. Such extractive cases constitute $<$2% of responses, confirming that paraphrasing is the dominant strategy.

A.2.3 Edge Cases: Variance-Induced Underperformance

Example 4: Workplace Vocabulary (GPT-OSS-20B)

- **Question:** What are the most common words you say at work without naming your job?
- **Reference Summary:** Replies reflect a variety of responses, with some sharing industry-specific jargon, others contributing colloquialisms, and a few including humorous comments about workplace vocabulary.
- **Model Response:** Hello, thanks, please, can you, need, update, schedule, project, report, meeting, deadline, confirm, idea, team, feedback, next, steps, issue, solution, progress, review, discuss, sorry.

Metric	Score
BLEU-1	0.00%
ROUGE-L F1	2.06%
BERTScore F1	79.99%
Cosine Similarity	97.95%

Analysis: GPT-OSS-20B's list format diverges from the reference's narrative structure. Despite 0% BLEU and reduced BERTScore (3.3pp below model mean), cosine similarity remains at 97.95%—demonstrating that even "underperforming" responses maintain strong semantic alignment.

A.3 Complete Metric Statistics

Tables 7–12 report comprehensive statistics for all evaluation metrics computed over 11,515 question-response pairs per model.

A.4 Distribution Visualizations

This section presents visual comparisons of evaluation metrics across all four models. The figures confirm the semantic-lexical paradox quantitatively: lexical metrics (BLEU, ROUGE) show substantial cross-model variation, semantic metrics (BERTScore, Cosine Similarity) cluster near ceiling values, and MoverScore occupies an intermediate position reflecting the optimal transport cost of semantic alignment.

A.4.1 Lexical Overlap Metrics

Figures 2 and 3 show BLEU and ROUGE scores across models.

A.4.2 Semantic Similarity Metrics

Figures 4 and 5 show BERTScore and cosine similarity and MoverScore.

A.4.3 Logical Inference Patterns

Figure 7 shows NLI classification distributions.

A.5 Model Behavior Summary

Our analysis reveals three key patterns:

Abstraction-verbosity spectrum: Models range from extreme abstraction (Gemma2-9B) to extractive-verbose (Llama-3.1-8B). Despite a 13$\times$ difference in lexical overlap, cosine similarity varies by less than 0.5 percentage points.

Consistency: Mistral-7B is 30$\times$ more consistent than GPT-OSS-20B, suggesting architectural or training differences in handling diverse question types.

Scale $\neq$ performance: GPT-OSS-20B (20B parameters) underperforms Mistral-7B (7B parameters) across all metrics.

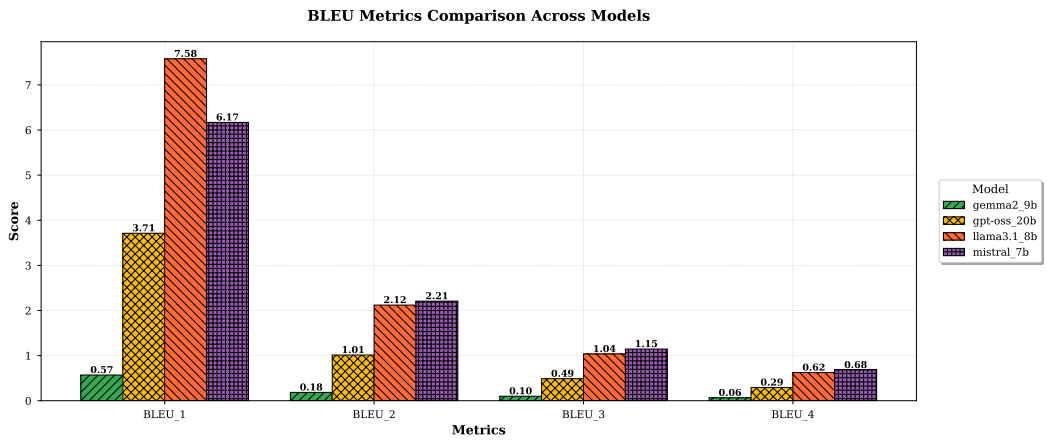

Figure 2: BLEU score comparison across models. BLEU-1 ranges from 0.57% (Gemma2-9B) to 7.58% (Llama-3.1-8B)—a 13× difference—while BLEU-4 scores remain below 1% for all models. Llama-3.1-8B's elevated scores may partially reflect self-alignment bias, as it was also used for reference summarization.

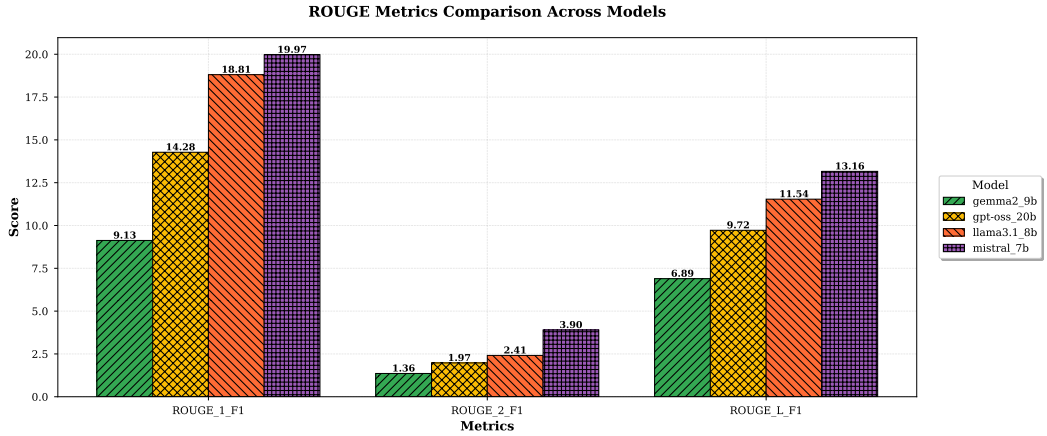

Figure 3: ROUGE F1 score comparison across models. Mistral-7B achieves highest scores across all variants (ROUGE-1: 19.97%, ROUGE-L: 13.16%), followed closely by Llama-3.1-8B. Gemma2-9B shows lowest overlap (ROUGE-1: 9.13%), consistent with its extreme abstraction profile. ROUGE-2 scores remain below 4% for all models, indicating minimal bigram overlap.

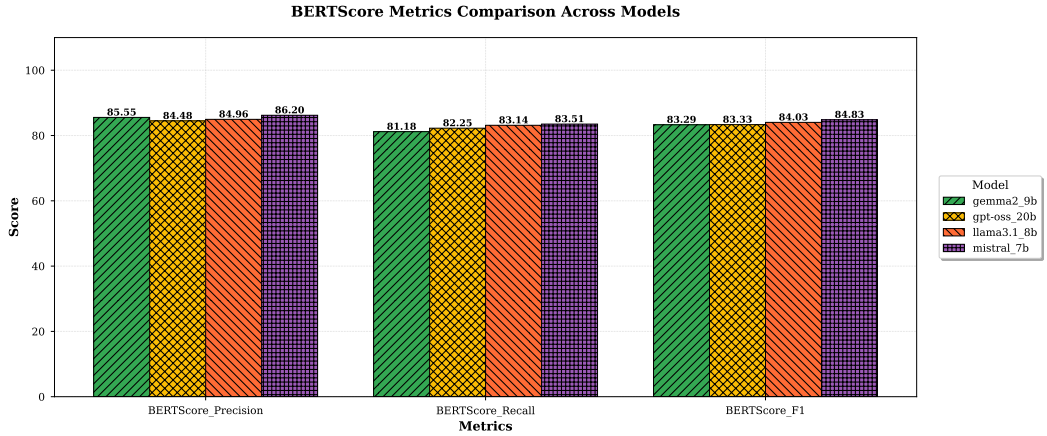

Figure 4: BERTScore comparison across models (RoBERTa-large embeddings). Despite the 13× variation in BLEU-1 scores (Figure 2), BERTScore F1 varies by only 1.54 percentage points (83.29%–84.83%). Precision exceeds recall for all models, indicating generated answers contain semantically relevant content but are more concise than reference summaries.

Reddit Question	Reference Summary	Example Model Response (Mistral-7B)
What made you realize you are not as young as you used to be?	The main question underlying this collection of stories is: "When do you realize you're getting older?" The key discussion themes revolve around the various ways people experience and acknowledge the passage of time, including physical changes, loss of familiarity with popular culture, feelings of irrelevance among younger generations, and a growing sense of responsibility and adulthood. These shared experiences often come as a surprise to individuals who had previously identified themselves as young or youthful, leading to feelings of discomfort, nostalgia, and acceptance of their advancing age.	Experiencing physical changes such as aches, tiredness after less exertion, or needing glasses, among other things, might hint that you're not as young as before.
What's something that instantly makes you lose respect for someone, no matter how successful or attractive they are?	The original post asks what behaviors or actions would instantly lead someone to lose respect for another person, regardless of their success or attractiveness. The replies reveal a range of behaviors that erode respect, including racism, rudeness, cruelty, dishonesty, hypocrisy, and lack of empathy towards others. Many responders emphasize the importance of treating people with kindness and respect, and that actions speak louder than external factors like success or physical appearance.	Consistently demonstrating a lack of empathy, integrity, or kindness can make others lose respect, regardless of success or attractiveness.
What's the best way to learn a new language?	"Here is a concise summary of the Reddit discussion: The original poster asks for advice on learning a new language, sparking a lively debate among commenters. The consensus seems to be that immersion in the target language through various means such as full immersion, watching content in the language, and interacting with native speakers is key to successful language acquisition. However, some commenters also suggest more accessible alternatives, including using apps like Duolingo or joining language exchange groups."	The best way to learn a new language is through immersion, consistent practice, and structured study. Utilize language learning apps, take classes, converse with native speakers, and engage in daily activities in the target language.
Reddit, what is a weird thing that you do?	In this Reddit discussion, users shared unusual habits and practices they engage in when feeling stressed or anxious. The thread revolves around a central question of what constitutes a "weird" behavior, with participants revealing a diverse array of quirky customs, from talking to plants to exposing themselves to various textures on the soles of their feet. The conversation also touches on mental health, with some users acknowledging that these behaviors may be coping mechanisms for anxiety or other issues.	I enjoy alphabetizing my spice rack in reverse order. It's odd, but brings me a strange sense of satisfaction!
What's the first sign you usually notice when your mental health starts declining?	"Summary The main question revolves around identifying early warning signs of mental health decline, with individuals sharing their personal experiences and red flags they notice before things spiral out of control. Common signs include changes in daily routines, unexplained sleep or appetite issues, social withdrawal, loss of interest in activities, and feelings of overwhelm or exhaustion. These subtle signals can serve as a catalyst for self-reflection and intervention to prevent further decline. Key discussion themes: * Early warning signs of mental health decline * Personal experiences with red flags and warning signs * Changes in daily routines and habits * Unexplained sleep or appetite issues * Social withdrawal and isolation * Loss of interest in activities and hobbies * Feelings of overwhelm, exhaustion, and burnout"	The first sign often noticed when mental health starts declining can include feelings of persistent sadness, increased anxiety, loss of interest in activities, changes in sleep patterns, and difficulty concentrating.

Table 6: Sample questions, reference summaries, and model responses from our dataset. References are LLM-generated summaries of multiple human responses; model responses demonstrate the semantic-lexical paradox where meaning is preserved despite different wording.

Model	Metric	Mean	Med.	Std	Min	Max
Gemma2-9B	BLEU-1	0.57	0.05	1.41	0	33.3
	BLEU-4	0.06	0.01	0.24	0	7.1
GPT-OSS-20B	BLEU-1	3.71	2.34	4.02	0	29.8
	BLEU-4	0.29	0.16	0.54	0	10.4
Llama-3.1-8B	BLEU-1	7.58	7.17	5.27	0	78.6
	BLEU-4	0.62	0.34	1.85	0	69.3
Mistral-7B	BLEU-1	6.17	4.86	5.38	0	31.2
	BLEU-4	0.68	0.31	1.18	0	14.1

Table 7: BLEU score statistics (%). Low means with standard deviations exceeding means indicate right-skewed distributions clustered near zero.

Model	Metric	Mean	Med.	Std	Min	Max
Gemma2-9B	ROUGE-1	9.13	8.22	6.57	0	40.5
	ROUGE-L	6.89	6.45	4.61	0	33.3
GPT-OSS-20B	ROUGE-1	14.28	14.14	7.80	0	56.0
	ROUGE-L	9.72	9.71	4.95	0	48.0
Llama-3.1-8B	ROUGE-1	18.81	18.69	7.93	0	86.7
	ROUGE-L	11.54	11.21	4.97	0	86.7
Mistral-7B	ROUGE-1	19.97	19.82	7.83	0	48.4
	ROUGE-L	13.16	12.77	5.14	0	37.2

Table 8: ROUGE F1 score statistics (%). Mistral-7B achieves highest scores; Gemma2-9B lowest.

Model	Metric	Mean	Med.	Std	Min	Max
Gemma2-9B	Precision	85.55	85.45	2.70	70.6	95.5
	Recall	81.18	81.12	1.59	76.3	89.2
	F1	83.29	83.26	1.77	74.7	92.2
GPT-OSS-20B	Precision	84.48	84.65	4.81	0.0	93.5
	Recall	82.25	82.58	4.40	0.0	89.2
	F1	83.33	83.56	4.42	0.0	89.3
Llama-3.1-8B	Precision	84.96	84.86	1.91	74.9	97.8
	Recall	83.14	83.13	1.63	77.8	97.8
	F1	84.03	83.95	1.60	77.5	97.6
Mistral-7B	Precision	86.20	86.22	2.05	73.3	94.0
	Recall	83.51	83.52	1.55	78.2	89.1
	F1	84.83	84.83	1.61	77.6	90.5

Table 9: BERTScore statistics (%, RoBERTa-large). Precision exceeds recall across all models, indicating concise but semantically relevant outputs.

Model	Mean	Med.	Std	Min	Max
Gemma2-9B	99.20	99.25	0.27	97.9	99.8
GPT-OSS-20B	99.10	99.40	4.90	0.0	99.8
Llama-3.1-8B	99.47	99.51	0.18	97.8	99.9
Mistral-7B	99.51	99.54	0.16	98.2	99.8

Table 10: Cosine similarity statistics (%, RoBERTa-large). All models exceed 99% mean similarity. GPT-OSS-20B shows 30× higher variance than Mistral-7B.

Model	Mean	Med.	Std	Min	Max
Gemma2-9B	50.92	50.88	2.27	43.64	60.16
GPT-OSS-20B	51.84	52.04	3.42	0.00	59.38
Llama-3.1-8B	52.65	52.53	2.07	44.37	69.97
Mistral-7B	53.42	53.39	2.08	44.91	61.99

Table 11: MoverScore statistics (%). Mistral-7B achieves the highest mean score. GPT-OSS-20B shows notably higher variance (std = 3.42) with zero-value outliers, consistent with its erratic patterns observed in other metrics. MoverScore's intermediate range (50.9–53.4%) sits between lexical metrics (<8% BLEU-1) and other semantic metrics (>83% BERTScore, >99% cosine similarity).

Model	Label	Mean	Med.	Std	Min	Max
Gemma2-9B	Entailment	9.04	0	27.4	0	100
	Contradiction	5.15	0	21.0	0	100
	Neutral	85.81	100	33.2	0	100
GPT-OSS-20B	Entailment	3.43	0	15.8	0	100
	Contradiction	6.94	0	23.7	0	100
	Neutral	89.39	100	28.0	0	100
Llama-3.1-8B	Entailment	4.06	0	15.1	0	100
	Contradiction	5.53	0	18.1	0	100
	Neutral	90.41	100	22.8	0	100
Mistral-7B	Entailment	8.71	0	23.3	0	100
	Contradiction	4.31	0	16.8	0	100
	Neutral	86.98	100	27.6	0	100

Table 12: NLI distribution statistics (%, BART-large-MNLI). Neutral classifications dominate (86–90%), indicating models generate semantically related but non-entailing content.

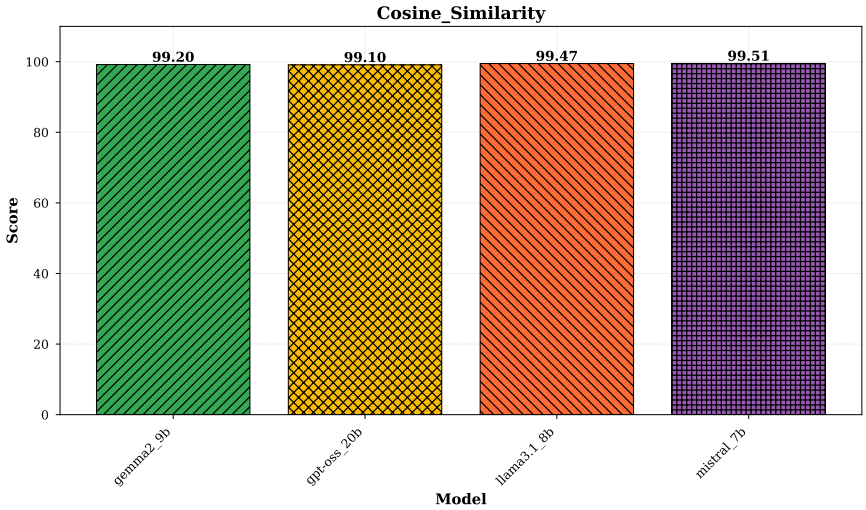

Figure 5: Cosine similarity comparison across models. All models exceed 99% similarity in RoBERTa-large embedding space, with only 0.41 percentage points separating Mistral-7B (99.51%) from GPT-OSS-20B (99.10%). This near-ceiling performance contrasts sharply with the <8% BLEU-1 scores—the 90+ percentage point gap constitutes our central finding.

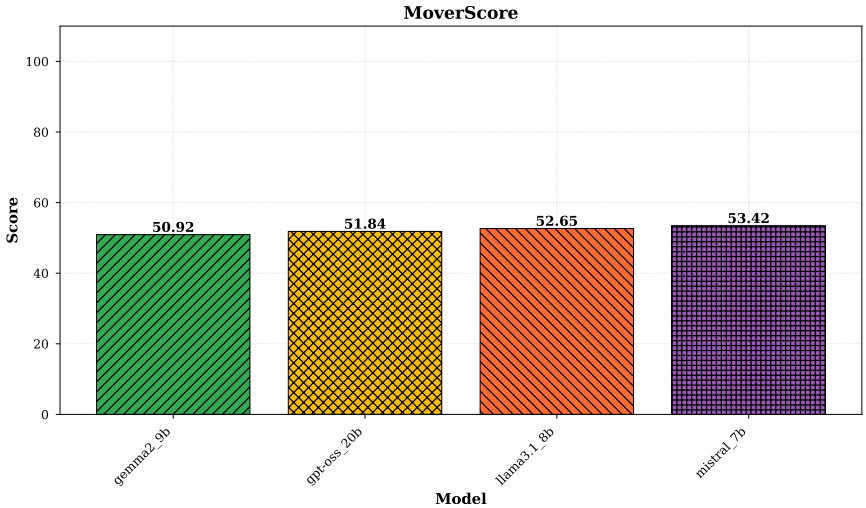

Figure 6: MoverScore comparison across models. Mistral-7B achieves the highest score (53.42%), followed by Llama-3.1-8B (52.65%), GPT-OSS-20B (51.84%), and Gemma2-9B (50.92%).

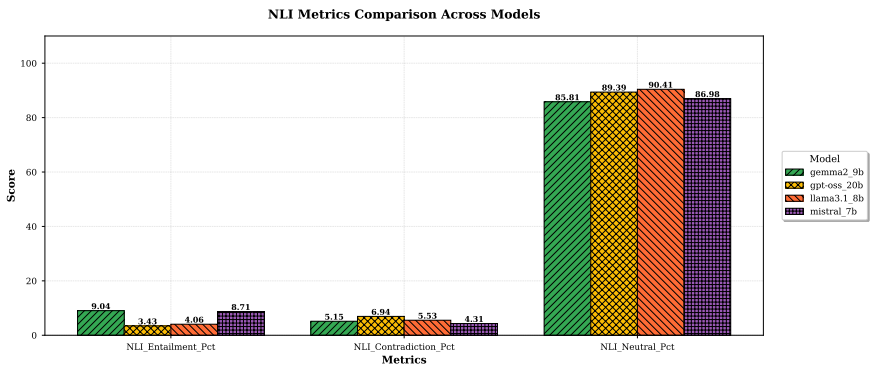

Figure 7: NLI classification distribution across models. Neutral classifications dominate (86–90%), with entailment rates varying from 3.43% (GPT-OSS-20B) to 9.04% (Gemma2-9B). Contradiction rates remain below 7% for all models, suggesting direct factual conflicts with human consensus are rare. See Table 5 in Section 4.4 for exact values.