Article

Not peer-reviewed version

Evaluating ChatGPT's Semantic Alignment with Community Answers: A Topic-Aware Analysis Using BERTScore and BERTopic

Mashael M. Alsulami [*]

Posted Date: 23 April 2025

doi: 10.20944/preprints202504.2000.v1

Keywords: Large Language Models (LLMs); Open-Domain Question Answering; ChatGPT; Topic Modeling; Semantic Evaluation

Article

Evaluating ChatGPT's Semantic Alignment with Community Answers: A Topic-Aware Analysis Using BERTScore and BERTopic

Mashael M. Alsulami

Department of Information Technology, College of Computers and Information Technology, Taif University, Taif, Saudi Arabia; mashael.s@tu.edu.sa

Abstract: This study evaluates the semantic alignment of ChatGPT's responses with human-selected best answers in an open-domain question answering (QA) setting, using data from the Yahoo! Answers platform. Unlike prior research focused on domain-specific or educational datasets, this work investigates ChatGPT's general-purpose QA capabilities across a diverse topical landscape. We apply BERTopic to extract latent themes from 500 full-question samples and use BERTScore metrics (precision, recall, F1) to quantify semantic similarity between ChatGPT-generated answers and top-rated community responses. Results show that ChatGPT achieves a strong average F1 score of 0.827, indicating high overall alignment with human judgments. Nonetheless, topic level analysis revealed important performance differences: the model was strong when asked factual and encyclopedia type questions, but was less capable of responding to subjective, ambiguous, or advice related questions. In this research, we proposed a topic-sensitive evaluation framework, that can be used to assess large language models in open-domain QA situations, that will add to the understanding of current benchmarking, interpreting performance, and building successful conversational AI systems.

Keywords: Large Language Models (LLMs); open-domain question answering; ChatGPT; topic modeling; semantic evaluation

1. Introduction

Natural Language Processing (NLP) systems are relying more on Large Language Models (LLMs), such as ChatGPT, for generating fluent, contextually aware, and sensible natural language responses across a variety of domains. Community Question Answering (CQA) platforms, such as Yahoo! Answers, provide further rich and difficult environments for evaluating the performance of these models because of the informality, breadth, and diversity of user-generated content. CQA platforms capture the naturalistic information seeking behavior across many domains and scenarios, ranging from technical problem resolution to social dilemmas. Therefore, they are particularly well positioned to evaluate the generalization properties of LLMs in various naturalistic settings [1,2]. While prior research has demonstrated that ChatGPT can perform competitively with human annotators in structured NLP tasks such as sentiment analysis and topic classification [3,4], its performance in replicating or improving upon human-generated responses in open-domain QA—particularly in contexts shaped by community interaction and cultural preferences—remains underexplored. The Yahoo! Answers dataset, with its annotated fields for question titles, question content, and community-voted best answers, offers a grounded benchmark for such comparative evaluation [5]. This study aims to address two central research questions: *How consistent is ChatGPT's semantic alignment with human-voted answers in a large-scale community QA setting?*, and *Can topic modeling reveal patterns in ChatGPT's answer quality across varying types of user-generated questions?* To investigate these questions, we conduct a comprehensive evaluation of ChatGPT's answers using a stratified sample of Yahoo! Answers posts. Semantic alignment is measured using BERTScore [4] which is a contextual similarity metric that accounts for paraphrasing and lexical variation. Additionally, we employ BERTopic to extract latent thematic structures within the dataset and analyze performance differences across topics. This two-tier approach allows for a nuanced understanding of ChatGPT's capabilities and limitations in offering factual or subjective answers. Our results shed light on the situations and contexts in which ChatGPT delivers high semantic fidelity, and more importantly, where responses miss the mark, especially with emotionally tinged or ambiguously stated inquiries. The findings also contribute to the growing discussion of the interpretability and dependability of LLMs for real-life QA scenarios, while providing developers and researchers tangible takeaways when leveraging or evaluating generative AI in open-ended information retrieval environments.

2. Related Work

The surge in studies exploring the capacity of large language models (LLMs) for open-domain question answering (QA) established both their capacities, as well as their associated limits. Most studies in particular have examined ChatGPT and its associated model(s) (e.g., GPT-3.5, GPT-4) in terms of fluency, adaptivity, and semantic understanding in the context of QA.

Earlier studies such as Bahak et al. [6] emphasized that while ChatGPT achieves high performance on general-purpose QA benchmarks (e.g., SQuAD, NewsQA), it tends to underperform compared to task-specific models when contextual grounding is absent. Their comparative evaluation using exact match (EM) and F1 metrics showed that ChatGPT exhibits a strong bias toward simpler factual questions while struggling with inferential or "why" questions, often due to hallucination effects.

In a complementary vein, Tan et al. [7] conducted a large-scale benchmark using over 190,000 KBQA test cases, systematically testing ChatGPT on multi-hop and compositional reasoning tasks. Their results reinforced the notion that while ChatGPT is effective in general scenarios, its accuracy declines in tasks demanding entity disambiguation or precise symbolic operations. The study also introduced a black-box diagnostic inspired by CheckList to identify model weaknesses.

Prompt engineering has emerged as a pivotal technique in improving QA output quality. Nzunda [8] provided a comprehensive taxonomy of prompt engineering strategies, from zero-shot to chain-of-thought prompting. He showed that strategic role-based and instruction-based prompting significantly influences the quality of ChatGPT's responses, particularly in educational and reasoning tasks.

From a generative capability perspective, Chan et al. [9] analyzed ChatGPT's use in question generation (QG) rather than answering. Their work highlighted ChatGPT's versatility in generating well-formed, contextually appropriate questions, and its performance was comparable or superior to transformer-based baselines like T5 and BART. This supports its dual potential in both creating and evaluating QA content.

Meanwhile, Omar et al. [10] compared ChatGPT to symbolic knowledge-based QA (KBQA) systems. Their findings emphasized that ChatGPT's fluency and surface alignment outperform traditional KGQAN systems in human-like delivery, but fall short when answers must be exact or retrieved from structured sources like SPARQL queries. This dichotomy points to the complementarity between neural and symbolic paradigms in QA.

Recent empirical evaluations have also incorporated developer and human feedback. For instance, Kabir et al. [11] conducted a developer-in-the-loop study with 130 StackOverflow questions, comparing ChatGPT with human-voted answers. In over 68% of cases, developers preferred GPT-generated answers based on readability and informativeness, reinforcing the model's perceived value in real-world QA contexts.

In specialized domains, such as medicine, Li et al. [12] explored the integration of domain-specific corpora into ChatGPT pipelines for biomedical question answering (MDQA). They emphasized the need for hybrid unimodal and multimodal methods, especially for high-risk scenarios requiring visual and textual inputs (e.g., X-rays and diagnoses). Their findings outline both the opportunities and constraints of using LLMs in expert-dependent QA domains.

Collectively, these studies illustrate the evolving landscape of ChatGPT in QA tasks: from generative prompting and contextual QA evaluation to domain-specific applications. Our study extends this discourse by using semantic similarity metrics like BERTScore to quantify ChatGPT's alignment with community-voted answers on open-domain datasets. Unlike previous works focused on benchmark datasets or narrow domains, we integrate topic modeling to correlate answer quality with question types, thereby introducing a novel stratification for performance diagnostics.

3. Methodology

This study adopts a multi-stage evaluation pipeline to assess the performance of ChatGPT in open-domain question answering using community-sourced data. As illustrated in Figure 1, the methodology consists of four main components: data preprocessing, ChatGPT response generation, semantic similarity evaluation, and topic-based performance analysis.

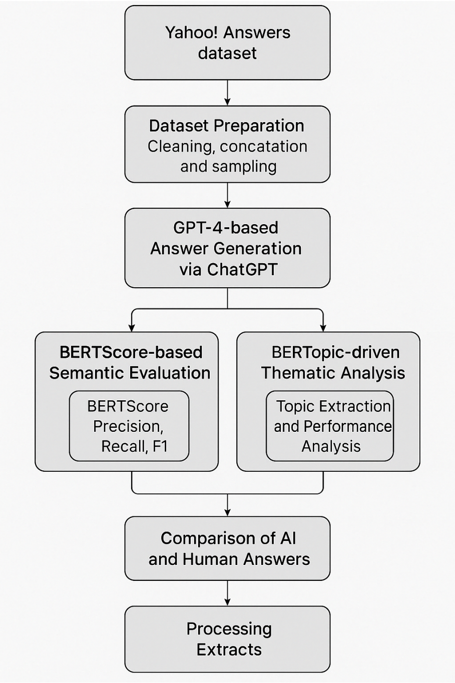
Yahoo! Answers dataset
Dataset Preparation
Cleaning, concatation and sampling
GPT-4-based Answer Generation via ChatGPT
BERTScore-based Semantic Evaluation
BERTScore Precision, Recall, F1
BERTopic-driven Thematic Analysis
Topic Extraction and Performance Analysis
Comparison of AI and Human Answers
Processing Extracts

Figure 1. The workflow of the proposed methodology.

We first preprocessed the Yahoo! Answers dataset by concatenating each question's title and content to form a unified input prompt. ChatGPT (GPT-4) was then prompted with these inputs to generate corresponding answers. To evaluate the semantic quality of these responses, we computed BERTScore precision, recall, and F1 score metrics by comparing the model's output to the human-selected best answers.

To gain insight into thematic variation, we applied BERTopic to the dataset, leveraging transformer-based embeddings and density-based clustering to identify latent topics. Each question was assigned to a topic, and performance metrics were aggregated accordingly. Further analysis involved clustering high- and low-performing responses based on their semantic scores, using TF-IDF and KMeans, followed by PCA for visualization.

This methodological framework enables both quantitative (semantic similarity) and qualitative (topic-based) evaluations of ChatGPT's performance. The following subsections describe each stage in detail.

3.1. Dataset Preparation

We utilized the Yahoo! Answers Topic Classification dataset, a well-known open-domain QA corpus frequently used in benchmarking classification and information retrieval systems [5]. Each sample in the dataset contains a question_title, question_content, and a best_answer field representing

doi:10.20944/preprints202504.2000.v1

the highest-rated response provided by community users. To generate input prompts with sufficient contextual coverage, we concatenated question_title and question_content into a new field labeled full_question. This field was used as the input prompt for ChatGPT. Missing or malformed values were cleaned using string-based imputation, and all text was standardized to UTF-8 encoding. To ensure diversity while managing computational constraints, we extracted two separate samples from the dataset:

- 100-question sample for ChatGPT response generation and semantic evaluation.;
- 500-question sample for topic modeling using BERTopic.

This stratified sampling strategy ensured that both the fine-grained and thematic analyses remained representative of the dataset's topical diversity.

3.2. ChatGPT Response Generation

For generating AI-based answers, we employed GPT-4, one of the most advanced autoregressive language models available at the time of writing [13]. Following best practices in few-shot prompting and controlled evaluation [14], each prompt to the model included:

- A system message guiding behavior: "You are a helpful question-answering assistant. Provide clear, accurate, and reliable responses";
- A user message containing the full_question text.

Responses were generated using OpenAI official SDK with a temperature of 0.7 and a maximum token limit of 300 to balance linguistic creativity and factual consistency [15]. All responses were captured and stored in a new column gpt_answer. Exception handling routines were added to manage rate limits, malformed inputs, or empty completions. This procedure produced a structured dataset of parallel ChatGPT-human answer pairs, suitable for both quantitative evaluation and qualitative inspection.

3.3. Topic Modeling with BERTopic

To gain insights into the types of questions posed within the dataset, and to later assess topic-specific variation in ChatGPT's performance, we employed BERTopic [16], a hybrid topic modeling approach combining transformer embeddings with density-based clustering [17]. Using the all-MiniLM-L6-v2 embedding model [18], we encoded each full_question in the 500-question sample and reduced the embedding dimensionality via UMAP (Uniform Manifold Approximation and Projection) [19]. Clustering was then performed using HDBSCAN [20], an unsupervised density-based algorithm, to group questions into interpretable topics. Each topic was described by a ranked list of top keywords and reviewed manually to assign semantic labels. These topics later served as the basis for stratified performance comparison in the results section.

3.4. Semantic Evaluation with BERTScore

To evaluate the quality of ChatGPT's responses against community-voted best answers, we employed BERTScore [4], a state-of-the-art semantic similarity metric based on contextual embeddings. BERTScore computes similarity by aligning tokens from both candidate and reference responses using cosine similarity in embedding space, rather than relying on exact surface word matches. We computed BERTScore Precision, Recall, and F1 score using the bert-base-uncased model for all 100 question-answer pairs. The metrics are interpreted as follows:

- Precision: measures the proportion of GPT-generated content that is semantically aligned with the reference;
- Recall: captures how much of the original best answer is recovered in the GPT output;
- F1 score: represents the harmonic mean of Precision and Recall, indicating balanced overlap.

doi:10.20944/preprints202504.2000.v1

All metric scores were stored alongside the original dataset to support topic-wise and overall analysis. The use of BERTScore enables a more nuanced understanding of the semantic fidelity of ChatGPT's output, especially in cases of paraphrased or differently worded but conceptually equivalent answers.

3.5. Clustering High and Low Performing Questions

To explore patterns underlying ChatGPT's strengths and weaknesses, we performed a clustering analysis on questions where ChatGPT either succeeded or struggled. This post-evaluation phase builds on prior work in NLP interpretability, where clustering is used to expose latent structure in high-dimensional language tasks [21,22]. Using the BERTScore F1 metric as a proxy for semantic performance, we defined two cohorts:

- High-performance questions: BERTScore F1 ≥ 0.85;
- Low-performance questions: BERTScore F1 ≤ 0.78

The high threshold captures cases where ChatGPT responses closely matched the best human-voted answers, indicating strong semantic fidelity. The low threshold was selected based on distributional analysis of the dataset, ensuring that the filtered subset reflects borderline or semantically deficient responses while retaining enough samples for meaningful analysis. We extracted the full_question field from both groups and applied TF_IDF vectorization (max features = 5000, English stopword removal) to represent each question in vector space [23]. Each set was then clustered separately using KMeans with k=4, allowing us to identify dominant question clusters within high and low performance groups. To visualize semantic separability, we applied PCA (Principal Component Analysis) [24] to project the TF-IDF matrix into 2D space. Cluster assignments were added as labels and stored in new columns within the dataset.

4. Results

4.1. ChatGPT Answer Quality

To assess the semantic quality of ChatGPT-generated responses, we computed BERTScore precision, recall, and F1 score metrics for each question-answer pair across a representative sample of 500 open-domain questions drawn from the Yahoo! Answers dataset. These metrics were selected for their demonstrated robustness in capturing meaning-preserving paraphrasing and lexical variance [4]. As summarized in Table 1, ChatGPT achieved a mean BERTScore F1 of 0.827, with values ranging from 0.7506 to 0.893. This suggests a generally strong semantic alignment between ChatGPT's output and community-voted best answers.

Table 1. Summary statistics for BERTScore precision, recall, and F1 across 500 ChatGPT-generated responses.

Metric	BERTScore Precision	BERTScore Recall	BERTScore F1
Count	500.0	500.0	500.0
Mean	0.8197	0.8286	0.8237
Std	0.0277	0.0296	0.0229
Min	0.7329	0.7072	0.7506
25%	0.8040	0.8105	0.8087
50%	0.8202	0.8300	0.8247
75%	0.8383	0.8483	0.8394
Max	0.9211	0.9141	0.9023

Precision, recall, and F1 are computed using BERTScore to evaluate semantic similarity between ChatGPT responses and community-voted best answers.

Figure 2 illustrates a right-skewed distribution across all three metrics, indicating that most of the responses cluster in the high performance zone.

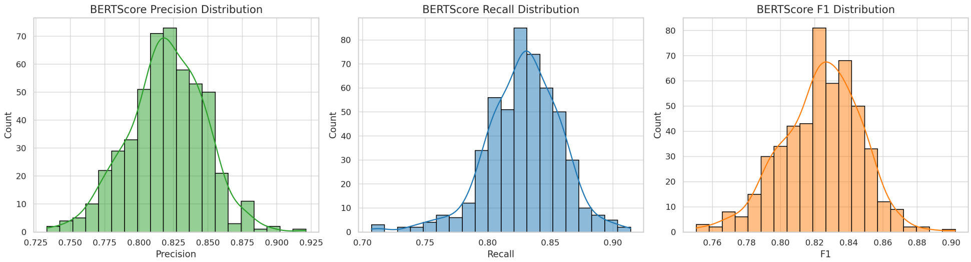

Figure 2. Distribution of BERTScore precision, recall, and F1 scores across 500 ChatGPT responses.

Figure 3 shows that most responses lie along the diagonal between precision and recall, showing that ChatGPT maintains a strong balance between completeness and relevance in its responses.

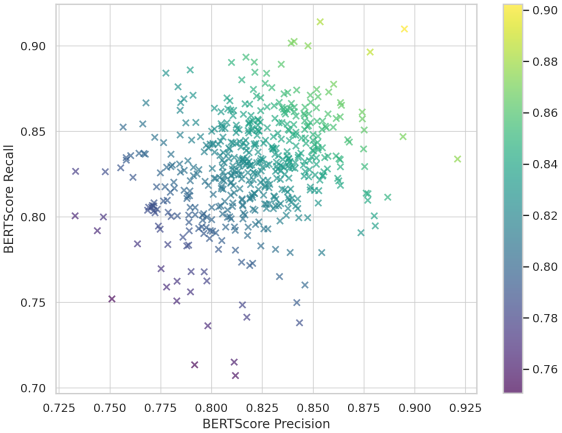

Figure 3. BERTScore precision vs. recall, showing balanced semantic alignment in ChatGPT responses.

These quantitative trends corroborate the findings of previous research [25], demonstrating the capacity of large language models to perform reliably on open-domain queries when provided sufficient context.

4.2. Distribution of High and Low Performance

To explore performance variability, we stratified the dataset using two empirically determined BERTScore F1 thresholds:

- High-performing responses were defined as those scoring F1 ≥ 0.85 (n = 55);
- Low-performing responses were defined as those scoring F1 ≤ 0.78 (n = 18).

These thresholds were selected based on the metric distribution observed in Section III.D and were chosen to focus on confidently successful and failed cases, respectively. The intent was to isolate semantically aligned vs. misaligned responses and to enable targeted downstream analysis.

4.3. Clustering Insights

To better understand patterns in model behavior, we conducted unsupervised clustering on both the high- and low-performing subsets using TF-IDF vectorization followed by KMeans (k=4), as outlined in Section III.E. Principal Component Analysis (PCA) was then applied to reduce dimensionality for 2D visualization, presented in Figure 4.

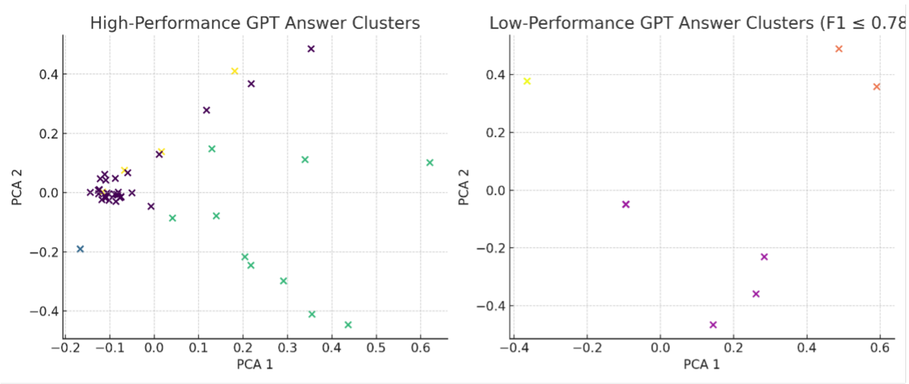

Figure 4. PCA plot of clustered high- and low-performing ChatGPT responses.

Clusters of high-performing responses, generally speaking, had strong topical cohesion–the questions present in these clusters were typically factual in nature involving asking for definitions, asking asking about historical facts, or process explanations. For these categories of queries ChatGPT is well-suited because the training corpus has a very high proportion of such content-to enable it to produce factually grounded responses that hint at little to no ambiguity. Clusters of low-performing responses had more subjective, underspecified requests, or requests with policy constraints. Examples of queries included "How does it feel to...?" (emotionally framed) and requests that triggered ChatGPT's refusal mechanism (i.e., "What are the lyrics?"). These findings indicate that performance breakdowns typically arise when the model a) does not have enough context b) is limited by non-generative safety protocols. The clear separation of topical coherence across clusters provides empirical support for the argument that ChatGPT's performance is heavily influenced by input clarity and task definition—an insight echoed in recent LLM interpretability research [26].

4.4. Topic Modeling Results and Analysis

To better understand the thematic distribution of questions and how topic content correlates with ChatGPT's performance, we applied BERTopic on the full-question text of 500 samples. BERTopic combines transformer-based embeddings (MiniLM-L6-v2), UMAP for dimensionality reduction to uncover coherent latent topics in unstructured text data. The model identified 15 distinct topics, each characterized by ranked keyword descriptors and representative questions. Manual inspection and qualitative labeling [27] revealed the presence of semantically coherent categories such as:

- Technology and Programming Help
- Medical and Health Inquiries
- Relationship and Emotional Advice
- Homework and School Assignments
- Definitions and Encyclopedic Questions

Each question was assigned to its respective topic cluster, and BERTScore F1 values were aggregated per topic to assess performance variability. The following findings were observed:

- High-performing topics (avg. F1 $\geq$ 0.85) included factual domains such as definitions, science-based queries, and technical troubleshooting, where questions had well-defined scopes and ChatGPT's output was aligned with clear reference answers.
- Low-performing topics (avg. F1 $\leq$ 0.78) were concentrated in subjective or underspecified categories like emotional advice, opinions, and non-specific homework help, where ChatGPT's responses often lacked the personal nuance or contextual grounding found in the best community answers.

This performance stratification is visualized in Figure 4, where PCA plots of clustered high- and low-performing responses show distinct thematic boundaries. Clusters of high F1 scores tend to be dense and semantically homogeneous, whereas low-performing clusters are more dispersed and ambiguous in content structure. These results support the hypothesis that ChatGPT's QA performance is topic-sensitive. Its strengths lie in structured, knowledge-oriented queries, likely due to pretraining on factual web corpora. Conversely, its weaknesses emerge in ill-posed or subjective questions, where human answers incorporate personal experience, empathy, or implicit cultural context—dimensions not easily replicated by LLMs. By combining BERTopic and BERTScore, our methodology enables not only evaluation of QA quality but also explainability, offering a topic-level lens through which model strengths and gaps can be identified.

4.5. Representative Examples

To provide concrete illustrations of semantic performance, we extracted the top five and bottom five ChatGPT responses based on BERTScore F1, presented in Tables 2 and 3 respectively.

Table 2. Examples of questions with Top 5 ChatGPT Responses (High BERTScore F1) (best human answers (A) and ChatGPT answers (B)).

Question	Answer A (Human)	Answer B (ChatGPT)
Who hit the longest home run at Jacobs Field?	On July 3rd 1999 Jim Thome hit a 511 foot shot to center field against Kansas City.	The longest home run hit at Jacobs Field (now Progressive Field) was by Jim Thome, who hit a 511-foot home run on July 3, 1999.
What is a procrastinator?	A person who puts off doing something, especially out of habitual carelessness or laziness.	A procrastinator is someone who delays or puts off tasks or responsibilities, often opting to do less urgent or enjoyable activities instead. This tendency can lead to stress, anxiety, and reduced productivity.
Are lab-created gems fake? I see them selling for thousands of dollars...?	Lab created gems are not fake. They are also called synthetic gems. They have the same composition as natural gems but are man-made in a lab.	Lab-created gems are not considered fake, as they have the same chemical and physical properties as natural gems. They are created in a lab under conditions that replicate those in nature, making them chemically identical.
What ballpark's home run balls often land on Waveland Avenue?	Wrigley Field in Chicago, known for fans waiting outside for balls hit out of the park.	Wrigley Field in Chicago is the ballpark where home run balls often land on Waveland Avenue.
Can earned income for one child be taken for the back child support of an adult child?	Depends on state laws—consult a lawyer specializing in family law.	Generally, child support is for minor children, not adults. Laws vary by jurisdiction, so consult a legal professional.

Table 3. Examples of Bottom 5 ChatGPT Responses (Low BERTScore F1): human answers (A) and ChatGPT answers (B).

Question	Answer A (Human)	Answer B (ChatGPT)
ooo00OOO000oooo00OOO00oooo.....la la la..how do I sing?	mmmm lalalalalalalalalaLA	Offers detailed singing advice including breathing techniques, vocal warm-ups, and considering a vocal coach.
What is the number that you divide the divisor into?	Dividend	Explains the concept of dividend with examples and a full breakdown of division terminology.
What are the lyrics? (Two songs: Panic! at the Disco & P. Diddy)	Full lyrics of both songs	Refuses due to copyright but suggests legal alternatives to find the lyrics.
What are Some BG rap links?	Long list of download links	Refuses to share download links due to copyright concerns and encourages legal alternatives.
I need an essay for my research plzz? (on American colonies)	No content (missing)	Provides a detailed multi-section essay outline covering early settlement, growth, conflict, and independence.

Top-performing responses achieved F1 scores approaching 0.89 and were associated with clearly phrased, fact-based questions. For instance, the question "What is a procrastinator?" elicited a precise and textbook-style definition, closely mirroring the best human response. Conversely, low-performing responses suffered from low content alignment, often due to incomplete or generic responses. A notable example is “I need an essay for my research plzz?”, where ChatGPT responded with a vague outline rather than an actual essay. Another illustrative failure involved the question “What are the lyrics?”, for which ChatGPT declined to answer due to copyright constraints—while the best human answer included the actual lyrics. These examples highlight not only content divergence but also structural limitations of LLMs operating under strict output filters.

5. Discussion

This study assessed the performance of ChatGPT in answering open-domain community questions by comparing its responses to human-voted best answers using BERTScore and topic-level cluster analysis. Across 500 QA pairs, ChatGPT-4 showed strong semantic alignment with human responses (mean F1 = 0.827), suggesting high linguistic fluency and contextual understanding. These findings align with prior literature emphasizing ChatGPT’s superiority in generating fluent responses, especially for factual queries or definition-style prompts [6]. Our findings are also consistent with Chan et al.’s observations that effective prompting significantly enhances ChatGPT’s performance in NLP tasks, such as question generation and response evaluation [9]. Through topic modeling and cluster-based review, we further identified that questions with vague context, implicit expectations, or subjective tones contributed disproportionately to lower BERTScore performance. This is in line with limitations reported by Pichappan et al., who observed that ChatGPT’s semantic accuracy deteriorates when handling abstract or evaluative prompts, often resulting in stylistically plausible but semantically imprecise answers [28]. Interestingly, our use of BERTopic uncovered distinct performance patterns tied to content themes. Factual, technical, or procedural topics saw high GPT alignment, while emotionally charged or ambiguous topics tended to cluster within low-performing segments. This echoes Bahak et al.’s findings, where ChatGPT excelled with concrete questions but showed vulnerability to hallucination or content overreach in multi-hop reasoning or "why" questions [6]. Although BERTScore enabled nuanced semantic evaluation beyond surface-level overlap, it does not capture fluency, factual correctness, or pragmatic utility. Prior studies have emphasized the necessity of human evaluations or hybrid scoring frameworks to comprehensively assess LLM responses [9], [28]. As such, the high F1 scores reported in this study may not reflect failure cases involving hallucinated content or missing factual details — a known concern in generative QA systems [6]. This study is not without limitations. First, it was confined to a single model (ChatGPT-4), evaluated only in English, and focused on one dataset (Yahoo! Answers). While this provides focused insights, generalizability to other languages, tasks, or models (e.g., Claude, Gemini) remains uncertain. Secondly, the evaluation relied solely on BERTScore, which, while robust, may benefit from augmentation via human annotation or model explainability techniques. Future work could involve cross-model comparisons, multilingual QA evaluations, and the integration of human preference scores or challenge sets focused on hallucination detection. Additionally, fine-grained error annotation and performance audits by topic and intent type would offer richer interpretability. In summary, our findings reaffirm ChatGPT’s high competence in answering a wide variety of open-domain questions, particularly those grounded in factual knowledge. However, its limitations in nuanced reasoning, subjective interpretation, and question ambiguity reflect persistent challenges in the design and evaluation of general-purpose LLMs.

6. Conclusions

The current study investigated the semantic quality of ChatGPT responses in the context of open-domain question answering, via the Yahoo! Answers dataset. We quantitatively examined how ChatGPT’s answers semantically aligned to the human-selected best answers associated with a broad variety of user-generated questions, while utilizing BERTScore. Further, we provided an overview of latent themes we identified in the dataset with the use of the BERTopic, followed by an analysis of ChatGPT’s performance at a topic-level. We found that ChatGPT demonstrates good semantic alignment in factual, well-structured domains (e.g., definitions, technical explanations, and encyclopedic content, etc.), but performs worse in subjective, underspecified queries, especially those related to emotion, opinion and implicit context. This indicates a limitation in ChatGPT's ability to navigate nuanced human communication. Using semantic similarity metrics with topic modeling provides a complete framework to assess large language models in real-world question and answer situations. In future work, we could expand this work to include human-in-the-loop validation, comparing different models, or examining the influence of prompt design on answer quality.

Author Contributions: The author has read and agreed to the published version of the manuscript.

Data Availability Statement: The data supporting the findings of this study are available upon request from the corresponding author.

Acknowledgments: The author acknowledges the Deanship of Graduate Studies and Scientific Research, Taif University, for funding this work.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* **2018**.
2. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv preprint arXiv:1908.10084* **2019**.
3. Liu, Y.; et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692* **2019**.
4. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.; Artzi, Y. BERTScore: Evaluating Text Generation with BERT. *arXiv preprint arXiv:1904.09675* **2019**.
5. Yang, Y.; et al. Yahoo! Answers Topic Classification Dataset. In Proceedings of the Conference on Natural Language Learning, 2017.
6. Bahak, H.; Taheri, F.; Zojaji, Z.; Kazemi, A. Evaluating ChatGPT as a Question Answering System: A Comprehensive Analysis and Comparison with Existing Models. *arXiv preprint arXiv:2312.07592* **2023**.
7. Tan, R.; Su, Y.; Yu, W.; Tan, X.; Qin, T.; Wang, Y.; Liu, T.Y. A Comprehensive Evaluation of ChatGPT on 190K Knowledge-Based QA Instances. *arXiv preprint arXiv:2306.05685* **2023**.
8. Nzunda, J. Prompt Engineering for ChatGPT: A Taxonomy and Systematic Review. *arXiv preprint arXiv:2306.13676* **2023**.
9. Chan, W.; An, A.; Davoudi, H. A Case Study on ChatGPT Question Generation. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2023.
10. Omar, S.; Gupta, M.; Dutta, S. Symbolic versus Neural QA Systems: A Comparative Analysis of ChatGPT and KGQAN. *Journal of Web Semantics* **2023**, *80*, 100776.
11. Kabir, R.; Dey, T.; Ahmed, N.; Chowdhury, N.H. Evaluating ChatGPT Answers for Stack Overflow Questions: A Developer-Centric Study. *Empirical Software Engineering* **2023**, *29*, 12–30.
12. Li, Y.; He, M.; Zhang, R.; Lin, J.; Xu, W. Bridging Biomedical Multimodal QA with Large Language Models. *arXiv preprint arXiv:2401.02523* **2024**.
13. Brown, T.; et al. Language Models are Few-Shot Learners. *NeurIPS* **2020**.
14. Ouyang, L.; et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155* **2022**.
15. Bergs, A. What, If Anything, Is Linguistic Creativity? *Gestalt Theory* **2019**, *41*, 173–184. https://doi.org/10.2478/gth-2019-0017.
16. Nedungadi, P.; Veena, G.; Tang, K.Y.; Menon, R.R.K.; Raman, R. AI Techniques and Applications for Online Social Networks and Media: Insights From BERTopic Modeling. *IEEE Access* **2025**, *13*, 37389–37402. https://doi.org/10.1109/ACCESS.2025.3543795.
17. Kriegel, H.P.; Kröger, P.; Sander, J.; Zimek, A. Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2011**, *1*, 231–240. https://doi.org/10.1002/widm.30.
18. Chen, H.; Jones, G.J.F.; Brennan, R. An Examination of Embedding Methods for Entity Comparison in Text-Rich Knowledge Graphs. In Proceedings of the Proceedings of the 32nd Irish Conference on Artificial

Intelligence and Cognitive Science (AICS 2024). CEUR Workshop Proceedings, 2024. Available under CC BY 4.0 license.

19. McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint arXiv:1802.03426* **2020**.
20. Rahman, M.F.; Liu, W.; Suhaim, S.B.; Thirumuruganathan, S.; Zhang, N.; Das, G. HDBSCAN: Density based Clustering over Location Based Services. *arXiv preprint arXiv:1602.03730* **2016**. Presented at ACM SIGMOD Workshop.
21. Tenney, I.; Das, D.; Pavlick, E. What do you learn from context? Probing for sentence structure in contextualized word representations. In Proceedings of the Proceedings of ICLR, 2019.
22. Liu, N.F.; Grefenstette, E.; Dyer, C. Analyzing the Structure of Attention in a Transformer Language Model. In Proceedings of the Proceedings of ACL, 2022.
23. Qaiser, S.; Ali, R. Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. *International Journal of Computer Applications* **2018**, *181*, 25–29. https://doi.org/10.5120/ijca2018917395.
24. Jolliffe, I.T. *Principal Component Analysis*; Springer Series in Statistics, Springer, 2002.
25. OpenAI. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* **2023**.
26. Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, M.; et al. On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258* **2021**.
27. Chang, J.; Gerrish, S.; Wang, C.; Boyd-Graber, J.L.; Blei, D.M. Reading tea leaves: How humans interpret topic models. In Proceedings of the Proceedings of the 22nd International Conference on Neural Information Processing Systems (NeurIPS), 2009, pp. 288–296.
28. Pichappan, P.; Krishnamurthy, M.; Vijayakumar, P. Analysis of ChatGPT as a Question-Answering Tool. *Journal of Digital Information Management* **2023**, *21*, 50–61.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.