# Semantic-Eval: A Semantic Comprehension Evaluation Framework for Large Language Models Generation without Training

**Shusheng Li[1,2,3], Jiale Li[1,2,3], Yifei Qu[1,2,3], Xinwei Shi[1,2,3], Yanliang Guo[1,2,3], Ziyi He[1,2,3], Yubo Wang [1,2,3], Wenjun Tan*[1,2,3],**

[1]Key Laboratory of Intelligent Computing of Medical Images, Ministry of Education, Northeastern University, China,
[2]School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China,
[3]National Frontiers Science Center for Industrial Intelligence and Systems Optimization, Northeastern University, China

2310731@stu.neu.edu.cn, tanwenjun@cse.neu.edu.cn

## Abstract

With the increasing prominence of large language models (LLMs), evaluating their text-generation capabilities has become an essential research challenge. Although LLM-based evaluation methods exhibit robust performance, the inherent stochastic nature of the LLM generation process introduces a degree of uncertainty in alignment with human preferences. To address this limitation, we propose Semantic-Eval, the first training-free framework designed to assess LLM-generated text based on semantic understanding. This framework computes semantic similarity between pairwise texts to evaluate the interdependence of semantic units, integrating a graph-based weighting mechanism to account for the differential contributions of individual sentences. A pre-trained natural language inference (NLI) model is also incorporated to mitigate potential semantic relationship biases. We evaluate Semantic-Eval across eight datasets that encompass four common NLP tasks. The experimental results indicate that Semantic-Eval surpasses traditional N-gram and BERT-based evaluation metrics, aligning more closely with human judgments and demonstrating a higher correlation than smaller LLMs. However, it slightly lags behind GPT-4. Finally, we demonstrate the effectiveness of Semantic-Eval in evaluating the generation quality of 13 large language models. The code is publicly available at[1].

## 1 Introduction

Large Language Models, including OpenAI o1[2], LLaMA-3 (Dubey et al., 2024), and DeepSeek-R1 (Guo et al., 2025), have emerged as pivotal drivers of progress in the field of natural language processing. LLMs have demonstrated exceptional capabilities in text generation by learning and capturing intricate semantic relationships from large-scale corpora, utilizing either autoregressive or self-encoding mechanisms. As the adoption of LLMs continues to proliferate, evaluating the text quality they generate has become a critical area of scholarly attention (Polo et al., 2024).

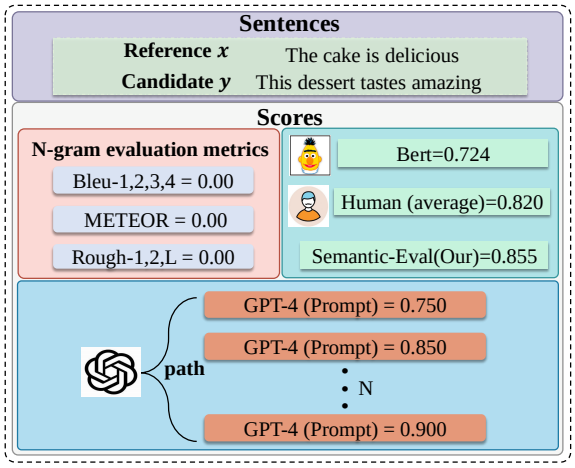



Figure 1: The diagram depicts scores for evaluating sentence similarity using various methods. The prompt is:*"Please assess the similarity of the two sentences and provide a score from 0 to 1."* The path indicates the number of samples. The text-embedding-ada-002 model is utilized as the embedding backbone in the Semantic-Eval framework. The Human (average) score represents the mean score obtained by five human evaluators.

Traditional evaluation metrics employed to assess the quality of text generation are commonly used to evaluate the output of large language models(Li et al., 2023). Metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) primarily assess text similarity by calculating n-gram overlap between the generated and reference texts. However, these metrics are often limited by their emphasis on superficial lexical overlap and fail to effectively capture the deeper semantic relationships between the generated and reference texts. As demonstrated in Figure 1, when evaluating the quality of LLM-generated text, the scores from n-

*Wenjun Tan is the corresponding author.
[1]https://github.com/LssTry/Semantic-Eval
[2]https://openai.com/o1/



*Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9675–9690

gram evaluation metrics are all zero. To overcome the limitations of traditional evaluation metrics in assessing LLM-generated content, recent studies (Zhang et al., 2023a; Zhou et al., 2023; Peng et al., 2024) have employed BERT-based methods. These approaches (Reimers et al., 2019; Zhang et al., 2019) utilize the semantic embeddings inherent in the BERT (Kenton and Toutanova, 2019) model to capture semantic similarity within the text. As shown in Figure 1, while BERT scores are closely aligned with human ratings, a noticeable discrepancy remains, indicating that BERT-based methods have inherent limitations in capturing semantic information.

Recent studies increasingly employ LLM-based evaluation methods to assess the performance of LLMs. These approaches typically leverage the language understanding capabilities of LLMs as the primary criteria for performance assessment (Yu et al., 2024; Ke et al., 2024; Park et al., 2024). Notable methodologies within this domain include instruction-tuning optimization (Wang et al., 2023; Polo et al., 2024; Song et al., 2024; Jiang et al., 2024) and multi-dimensional evaluation frameworks (Lin and Chen, 2023; Liang et al., 2024), both of which heavily rely on LLMs. Among these, GPT-based evaluation methods are often regarded as better aligned with human preferences to a certain extent. As illustrated in Figure 1, the prompt-optimized GPT-4 evaluator demonstrates notable consistency with human preferences. Nevertheless, GPT-4 (Achiam et al., 2023), which utilizes a probabilistic generation strategy, tends to produce outputs with inherent uncertainty. Furthermore, existing benchmarks (Hendrycks et al., 2020; Sarlin et al., 2020; Cobbe et al., 2021), designed for evaluating LLMs, generally yield strong performance results on these specific tasks. However, these benchmarks are meticulously crafted, and the task configurations and evaluation criteria they employ may not directly translate to real-world application scenarios.

To address the aforementioned challenges, this paper introduces Semantic-Eval, the first consistent, training-free automatic evaluation framework designed to assess the quality of text generated by LLMs from the perspective of semantic comprehension. Specifically, the framework computes the semantic similarity between all pairs of sentence embeddings within each text to quantify the degree of interdependence among the semantic units of the text. Recognizing the varying contributions of individual sentence units to the overall semantic content, Semantic-Eval incorporates a graph-based weighting mechanism that evaluates the relative significance of each semantic unit. The weight assigned to each sentence embedding is determined by analyzing the strength of associations between semantic units within an undirected, self-similar weighted graph. Subsequently, the semantic similarity values and their corresponding weights in the weighting matrix are utilized to evaluate sentence-level cross-text similarity between reference and candidate texts. To account for potential semantic dyadic relationships, the framework integrates a pre-trained NLI model, which generates a confidence score that mitigates bias in cross-text similarity measurements. Semantic-Eval was systematically evaluated on seven English datasets and one Dutch dataset, encompassing four tasks in NLP: sentiment analysis, text summarization, natural language Q&A, and sentence similarity. The experimental results demonstrate that, compared to traditional evaluation methods and BERT-based metrics, Semantic-Eval exhibits a significantly higher correlation with human judgments. Moreover, it outperforms smaller parameter large language models in aligning with human preferences, trailing only slightly behind GPT-4. Finally, Semantic-Eval was employed to assess the text generation quality of 13 large language models across six datasets. In summary, the primary contributions of this work are as follows:

1) We propose Semantic-Eval, the first training-free framework for evaluating text generated by large language models based on semantic comprehension.

2) We introduce a graph-based weighting mechanism that computes sentence-level similarity by evaluating the interdependence of semantic units within the text.

3) We demonstrate that Semantic-Eval aligns with human preferences across eight datasets and is used to assess the text generation quality of 13 large language models.

## 2 Related Work

### 2.1 LLM-based evaluation

The evaluation methods based on LLMs capitalize on the models' semantic understanding capabilities to conduct quantitative assessments of language models. Primary categories of LLM-based evaluation include instruction-based evaluation (Wang

et al., 2022a,b; Ajith et al., 2023), comparative evaluation (Chan et al., 2023; Lango and Dušek, 2023; Lambert et al., 2024), and multidimensional evaluation (Liang et al., 2022; Liu et al., 2023; Nguyen et al., 2024). These methods offer dynamic frameworks that provide more comprehensive assessments of model robustness compared to traditional static benchmarks (Zhang et al., 2024). However, the LLM generation process is often based on probabilistic sampling, which introduces variability, such that identical inputs may yield divergent scores (Holtzman et al., 2019). Moreover, the evaluation outcomes are highly sensitive to the precision of the instructions provided (Zeng et al., 2023). These limitations in current evaluation methodologies contribute to the uncertainty in aligning LLM-based evaluations with human preferences.

### 2.2 Static Benchmarks evaluation

The static benchmark evaluation methodology assesses the performance of LLMs using predetermined datasets, which are generally categorized into two primary groups: general language task benchmarks and specific downstream task benchmarks. General Language Task Benchmarks aim to evaluate the overall language comprehension abilities of LLMs across a wide array of tasks, with prominent benchmarks including GLUE (Wang et al., 2018), SuperGLUE (Sarlin et al., 2020), MMLU (Hendrycks et al., 2020), BIG-bench (Srivastava et al., 2022), and C-EVAL (Huang et al., 2024). In contrast, Specific Downstream Task Benchmarks are designed to assess LLM performance within specific application domains, with notable examples such as MultiMedQA (Singhal et al., 2023), MathBench (Liu et al., 2024), GAOKAO-Bench (Zhang et al., 2023b), and LawBench (Fei et al., 2023). Nevertheless, relying on static datasets in benchmark evaluations constrains the generalization capabilities of LLMs in real-world contexts (Mousavi et al., 2024). Furthermore, developing static benchmark datasets necessitates significant resource investment (Li et al., 2024).

### 2.3 Human evaluation

HumanEval-based methods primarily rely on manual scoring, user feedback, and expert reviews to evaluate LLMs across various dimensions, including accuracy, fluency, relevance, and diversity (Elangovan et al., 2024; Feng et al., 2024; Watts et al., 2024). However, these approaches are inherently prone to subjectivity and inconsistency, which may undermine the reliability of the assessment results (Chang et al., 2024; Li et al., 2025). Moreover, human evaluation methods face challenges related to inefficiency and scalability when assessing LLMs (Chiang and Lee, 2023; Chen et al., 2024a).

## 3 Semantic-Eval Framework

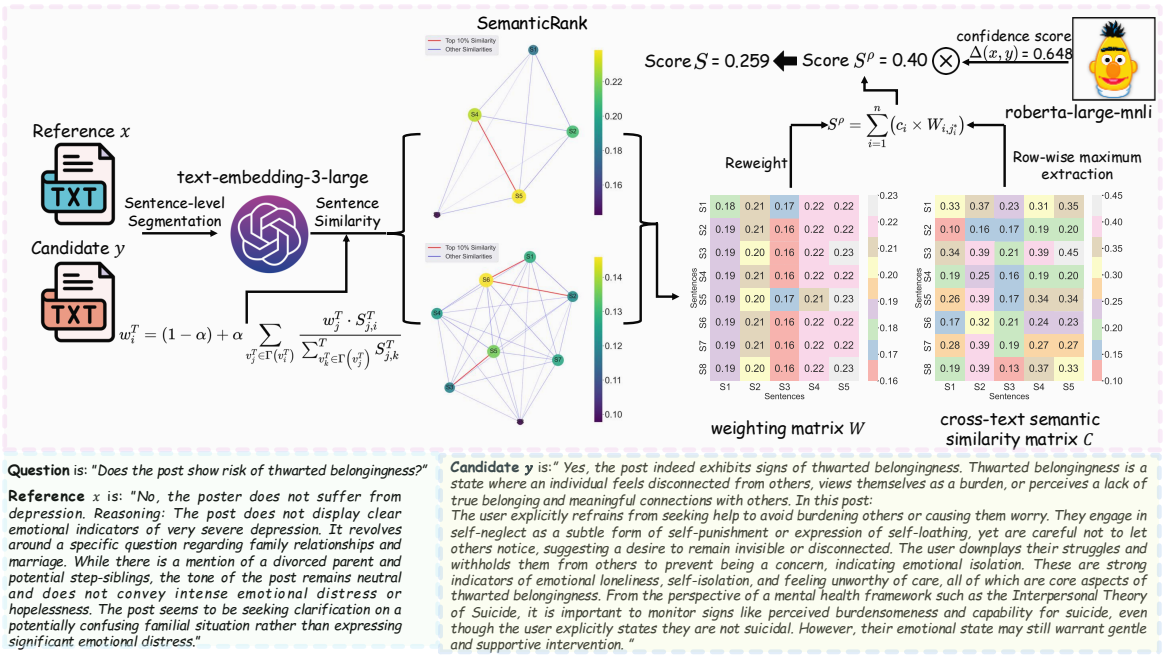
The Semantic-Eval framework aims to evaluate the quality of generated text by LLMs from the perspective of semantic understanding without additional training. Additionally, Semantic-Eval focuses on various NLP tasks, demonstrating its generalizability. The computation illustration of the Semantic-Eval framework is shown in Figure 2. An overview of the details of the semantic evaluation framework is provided below.

Initially, the reference text $x$ and the candidate text $y$ generated by LLM are parsed into a sequence of sentences using regular expressions, producing $x = \{x_1, \ldots, x_n\}$ and $y = \{y_1, \ldots, y_n\}$, respectively. Next, the same pre-trained vector embedding model, specifically text-embedding-3-large[3], projects each sentence from $x$ and $y$ into a $d$-dimensional vector space, yielding vector embedding $t_i \in \mathbb{R}^d$. To unify the notation, let $T$ represent a text which can be $x$ or $y$, where $|T|$ denotes the number of sentences. The cosine similarity is then used to measure the semantic correlation between all pairs of sentence embeddings within $T$, producing a self-similarity matrix $S^T$:

$$S^T_{i,j} = \frac{t_i \cdot t_j}{\|t_i\| \, \|t_j\|}, \quad 1 \leq i, j \leq |T|.$$

Given that the contribution of individual sentences to the overall semantics of a text is often disproportionate, it is critical to assign appropriate weights to sentences according to their semantic relevance to highlight key information. LLMs employed for sentence scoring (Zhang et al., 2023c) have demonstrated remarkable performance, illustrating their capacity for nuanced language comprehension. However, the inherent sampling strategy employed by these models introduces a degree of stochasticity, which leads to variability in the generated results. To ensure the consistency of evaluation outcomes, an unsupervised method, SemanticRank, is introduced to assign sentence

[3] https://platform.openai.com/docs/guides/embeddings




**Question** is: *"Does the post show risk of thwarted belongingness?"*

**Reference** $x$ is: *"No, the poster does not suffer from depression. Reasoning: The post does not display clear emotional indicators of very severe depression. It revolves around a specific question regarding family relationships and marriage. While there is a mention of a divorced parent and potential step-siblings, the tone of the post remains neutral and does not convey intense emotional distress or hopelessness. The post seems to be seeking clarification on a potentially confusing familial situation rather than expressing significant emotional distress."*

**Candidate** $y$ is:*" Yes, the post indeed exhibits signs of thwarted belongingness. Thwarted belongingness is a state where an individual feels disconnected from others, views themselves as a burden, or perceives a lack of true belonging and meaningful connections with others. In this post: The user explicitly refrains from seeking help to avoid burdening others or causing them worry. They engage in self-neglect as a subtle form of self-punishment or expression of self-loathing, yet are careful not to let others notice, suggesting a desire to remain invisible or disconnected. The user downplays their struggles and withholds them from others to prevent being a concern, indicating emotional isolation. These are strong indicators of emotional loneliness, self-isolation, and feeling unworthy of care, all of which are core aspects of thwarted belongingness. From the perspective of a mental health framework such as the Interpersonal Theory of Suicide, it is important to monitor signs like perceived burdensomeness and capability for suicide, even though the user explicitly states they are not suicidal. However, their emotional state may still warrant gentle and supportive intervention."*

Figure 2: Illustration of the computation of the Semantic-Eval. The candidate $y$ is the text generated by GPT-4o in response to a corresponding query. The reference text $x$ and candidate $y$ represent semantically opposing responses to the same question.

weights based on semantic significance without requiring training. Initially, an undirected graph is constructed to quantify the semantic relevance of each sentence vector embedding in relation to other sentences within the same text, using cosine similarity between sentence embeddings. Subsequently, TextRank (Mihalcea and Tarau, 2004) is employed to evaluate the relative importance of each sentence vector embedding derived from the connectivity and centrality of sentence embeddings within the graph. Further details are provided below.

The undirected weighted graph $G^T = (V^T, E^T)$ is constructed using a self-similarity matrix $S^T$, where $V^T = \{v_1^T, v_2^T, \ldots, v_{|T|}^T\}$ represents the set of nodes corresponding to the embeddings of individual sentences, and $E^T = \{e_1^T, e_2^T, \ldots, e_{|T|}^T\}$ represents the edges connecting the nodes based on the similarity between their vector embeddings. To filter out sentences with weaker semantic relationships, an edge is created between two nodes if the cosine similarity $S_{i,j}^T$ between their embeddings exceeds a threshold of 0.1, the weight of the edge is then set to the similarity value. Then, TextRank is employed to calculate the score $w_i^T$ for each node $v_i^T$:

$$w_i^T = (1-\alpha) + \alpha \sum_{v_j^T \in \Gamma(v_i^T)} \frac{w_j^T \cdot S_{j,i}^T}{\sum_{v_k^T \in \Gamma(v_j^T)} S_{j,k}^T},$$

where $\Gamma(v_i^T)$ denotes the set of neighbors of $v_i^T$, and $\alpha$ is the damping factor, which is set to 0.85. After several iterations, the steady-state vector scores denoted as $w^T$, are obtained and subsequently normalized:

$$\widetilde{w}_i^T = \frac{w_i^T}{\sum_{i=1}^{|T|} w_i^T},$$

where, $\sum_{i=1}^{|T|} \widetilde{w}_i^T = 1$. Then, the normalized weight vectors $\widetilde{w}^x = (\widetilde{w}_1^x, \ldots, \widetilde{w}_n^x)$ and $\widetilde{w}^y = (\widetilde{w}_1^y, \ldots, \widetilde{w}_m^y)$ of sentences in the reference $x$ and candidate $y$ texts are obtained. Here, $n$ and $m$ represent the number of sentences for $x$ and $y$, respectively.

With the above operations, we have completed the sentence slicing of the reference text $x$ and candidate text $y$ and re-weighted the sliced sentences according to their semantic importance degree. Next, we will assess the semantic similarity between the reference text $x$ and the candidate text $y$.

Firstly, the weighting matrix $W \in \mathbb{R}^{n \times m}$ is constructed, where each element is the average of the normalized weight vectors $\widetilde{w}^x$ and $\widetilde{w}^y$:

$$W_{i,j} = \frac{\widetilde{w}_i^x + \widetilde{w}_j^y}{2}, \quad 1 \le i \le n, 1 \le j \le m.$$

| Class | Metrics | SemRel2024(eng) | | | STS-B | | |
|---|---|---|---|---|---|---|---|
| | | $r$ | $\rho$ | $\tau$ | $r$ | $\rho$ | $\tau$ |
| Traditional-Eval | BLEU-1 | 0.542 | 0.533 | 0.375 | 0.445 | 0.432 | 0.304 |
| | BLEU-2 | 0.514 | 0.513 | 0.357 | 0.460 | 0.444 | 0.311 |
| | ROUGE-1 | 0.619 | 0.614 | 0.436 | 0.615 | 0.590 | 0.425 |
| | ROUGE-2 | 0.536 | 0.544 | 0.411 | 0.474 | 0.459 | 0.326 |
| | ROUGE-L | 0.574 | 0.567 | 0.401 | 0.571 | 0.543 | 0.387 |
| | METEOR | 0.415 | 0.412 | 0.285 | 0.311 | 0.304 | 0.210 |
| BERT-based | BERT | 0.582 | 0.578 | 0.413 | 0.531 | 0.491 | 0.345 |
| | BERTScore_$P$ | 0.671 | 0.668 | 0.485 | 0.565 | 0.520 | 0.367 |
| | BERTScore_$R$ | 0.657 | 0.653 | 0.472 | 0.561 | 0.520 | 0.367 |
| | BertScore_$F1$ | 0.678 | 0.675 | 0.491 | 0.578 | 0.536 | 0.380 |
| LLM-based | Qwen2-7B | 0.763 | 0.751 | 0.586 | 0.759 | 0.746 | 0.608 |
| | GPT-4 | <u>0.803</u> | <u>0.798</u> | **0.634** | **0.861** | **0.851** | **0.698** |
| OUR | Semantic-Eval | **0.806** | **0.808** | <u>0.620</u> | <u>0.787</u> | <u>0.807</u> | <u>0.635</u> |

Table 1: The correlations between automatic metrics and annotated sentence similarity scores on Natural Language Sentence Similarity Task. $r$, $\rho$, and $\tau$ denote the Pearson correlation coefficient, Spearman's rank correlation coefficient, and Kendall's tau coefficient, respectively. In each column, the top score is displayed in **bold** while the second highest is <u>underlined</u>.

Secondly, cosine similarity is employed to measure the semantic correlation between the sentences of the reference $x$ and the candidate $y$, resulting in the cross-text semantic similarity matrix $C \in \mathbb{R}^{n \times m}$. Then, each row of the $C$ is normalized. Following this, the sentences embedding in candidate text $y$ most semantically related to reference text $x$ are selected:

$$c_i = \arg \max_{1 \le j \le m} \frac{x_i^e \cdot y_j^e}{\|x_i^e\| \|y_j^e\|} \Big/ \sum_{j=1}^{m} \frac{x_i^e \cdot y_j^e}{\|x_i^e\| \|y_j^e\|}.$$

where, $x_i^e$ and $y_j^e$ represent the vector embeddings of $x_i$ and $y_j$, respectively. $c_i$ denotes the similarity value of the semantic most relevant embedding.

Finally, the sentence-level cross-text similarity between the reference text $x$ and the candidate text $y$ is quantified:

$$S^\rho = \sum_{i=1}^{n} \Big( c_i \times W_{i,j_i^*} \Big),$$

The $W_{i,j_i^*}$ represents the element of the weighting matrix $W$ corresponding to $c_i$, where $j_i^*$ denotes the index of the element in $W$ that corresponds to $c_i$. However, cross-text similarity fails to fully account for potential semantic opposition or ambiguity between the reference text $x$ and the candidate text $y$. To address this, the confidence score from natural language inference is introduced to evaluate the reliability of $S^\rho$. Specifically, the confidence score $\Delta(x, y)$, output by the pre-trained roberta-large-mnli model(Liu et al., 2019), is computed as the dot product with $S^\rho$, resulting in the final semantic score $S = \Delta(x, y) \cdot S^\rho$.

## 4 Experiments

### 4.1 Datasets

This study utilizes seven English-language datasets and one Dutch-language dataset to investigate four distinct tasks in NLP: text summarization, sentiment analysis, natural language Q&A, and natural language sentence similarity.

**Natural Language Sentence Similarity Task**

**SemRel2024**(Ousidhoum et al., 2024) is a semantic text relevance dataset encompassing 13 languages. Each data sample consists of a pair of sentences, each assigned a relevance score ranging from 0 to 1. For this study, only the English subset of the dataset is utilized.

**STS-B**(Cer et al., 2017) is a subtask of SemEval-2017 Task 1, designed to evaluate the semantic similarity between text pairs. The dataset comprises 8,628 sentence pairs from diverse corpora across domains and scenarios. Each sentence pair is manually labeled with a similarity score ranging from 0 to 5, where 0 indicates complete semantic irrelevance, and 5 indicates perfect semantic equivalence.

**Text Summarization Task**

| Class | Metrics | CNN/Daily Mail | | | XSUM | | |
|---|---|---|---|---|---|---|---|
| | | $r$ | $\rho$ | $\tau$ | $r$ | $\rho$ | $\tau$ |
| Traditional-Eval | BLEU-1 | 0.322 | 0.316 | 0.217 | 0.029 | 0.142 | 0.098 |
| | BLEU-2 | 0.367 | 0.376 | 0.260 | 0.082 | 0.195 | 0.133 |
| | ROUGE-1 | 0.383 | 0.358 | 0.246 | 0.465 | 0.412 | 0.288 |
| | ROUGE-2 | 0.350 | 0.359 | 0.246 | 0.316 | 0.305 | 0.215 |
| | ROUGE-L | 0.335 | 0.332 | 0.227 | 0.398 | 0.352 | 0.244 |
| | METEOR | 0.357 | 0.330 | 0.227 | 0.160 | 0.219 | 0.150 |
| BERT-based | BERT | 0.479 | 0.441 | 0.307 | 0.272 | 0.202 | 0.138 |
| | BERTScore_$P$ | 0.558 | 0.531 | 0.373 | 0.291 | 0.226 | 0.155 |
| | BERTScore_$R$ | 0.518 | 0.504 | 0.354 | 0.304 | 0.241 | 0.165 |
| | BERTScore_$F1$ | 0.601 | 0.580 | 0.414 | 0.304 | 0.241 | 0.165 |
| LLM-based | Qwen2-7B | 0.427 | 0.454 | 0.320 | 0.607 | 0.577 | 0.426 |
| | GPT-4 | **0.669** | 0.572 | **0.450** | **0.842** | **0.820** | **0.662** |
| OUR | Semantic-Eval | 0.554 | **0.583** | 0.429 | 0.699 | 0.661 | 0.494 |

Table 2: The correlations between automatic metrics and human judgments on Text Summarization Task. $r$, $\rho$, and $\tau$ denote the Pearson correlation coefficient, Spearman's rank correlation coefficient, and Kendall's tau coefficient, respectively. In each column, the top score is displayed in **bold** while the second highest is underlined.

**CNN/Daily Mail**(Nallapati et al., 2016)(Nallapati et al., 2016) is a dataset predominantly employed for abstractive text summarization in NLP. It contains approximately 30K data pairs, each consisting of a news article sourced from CNN or the Daily Mail, alongside a manually generated summary.

**XSUM**(Narayan et al., 2018) is a dataset designed to assess abstractive single-document summarization systems, focusing on producing concise single-sentence summaries from lengthy news articles. The dataset contains 226,711 BBC news articles from a variety of domains. Each data entry comprises a news article and its corresponding manually labeled single-sentence summary. In this study, the Dutch version of the XSum dataset is employed.

**Sentiment Analysis Task**

**IMDB**(Maas et al., 2011) is a dataset widely used in sentiment analysis research containing many movie reviews. The dataset consists of 50K reviews, each labeled as positive (1) or negative (0).

**Yelp Polarity**(Zhang et al., 2015) is a dataset designed to support sentiment analysis research. It contains 560,000 labelled samples, with ratings ranging from 1 to 5 stars. Reviews with ratings of 1 or 2 are classified as negative, while reviews of 3 or 4 are classified as positive.

**Natural Language Q&A Task**

**IMHI**(Yang et al., 2024) is a dataset tailored for interpretable mental health analysis on social media. It contains 105K samples covering eight distinct mental health-related tasks. Each sample consists of expert-written instructions and labels.

**Medical-o1**(Chen et al., 2024b) is a dataset designed for complex medical reasoning tasks, containing 40K samples that span a wide range of clinical scenarios. Each entry includes a prompt, an open-ended question, and a corresponding authentic answer based on a challenging medical examination.

### 4.2 Implementation Details

In all experiments, four RTX 4090 GPUs, each equipped with 24GB of memory, were utilized. Due to constraints in computational resources and the associated costs of the OpenAI API, we randomly selected 1,000 samples from each dataset to form the evaluation subset for each task. These evaluation subsets were then input into 13 LLMs to generate output results. The performance of Semantic-Eval in validating the tasks exhibited slight variation across different task types. Specifically, for sentence-level tasks, Semantic-Eval functions as an evaluator, assessing the outputs of the evaluation subsets, after which the dataset's evaluation scores are used to compute correlations. In contrast, for the text-level task, to the best of our knowledge, no existing dataset provides labeled scores to evaluate the similarity between pairs of texts. Consequently, three natural language processing annotators were employed to manually score the similarity of each text pair in the text-level evaluation subset. Furthermore, the self-labeled texts from the natural language Q&A and text summarization

| Class | Metrics | IMDB | | | Yelp Polarity | | |
|---|---|---|---|---|---|---|---|
| | | $r$ | $\rho$ | $\tau$ | $r$ | $\rho$ | $\tau$ |
| Traditional-Eval | BLEU-1 | 0.098 | 0.016 | 0.011 | 0.109 | 0.163 | 0.115 |
| | BLEU-2 | 0.121 | 0.040 | 0.027 | 0.144 | 0.238 | 0.170 |
| | ROUGE-1 | 0.409 | 0.168 | 0.114 | 0.447 | 0.515 | 0.383 |
| | ROUGE-2 | 0.310 | 0.227 | 0.155 | 0.417 | 0.523 | 0.389 |
| | ROUGE-L | 0.387 | 0.190 | 0.128 | 0.419 | 0.503 | 0.372 |
| | METEOR | 0.540 | 0.211 | 0.145 | 0.463 | 0.509 | 0.377 |
| BERT-based | BERT | 0.656 | 0.275 | 0.252 | 0.459 | 0.554 | 0.415 |
| | BERTScore_$P$ | 0.644 | 0.320 | 0.220 | 0.529 | 0.555 | 0.416 |
| | BERTScore_$R$ | 0.550 | 0.157 | 0.108 | 0.517 | <u>0.592</u> | <u>0.445</u> |
| | BERTScore_$F1$ | 0.542 | 0.232 | 0.172 | <u>0.534</u> | 0.589 | 0.444 |
| LLM-based | Qwen2-7B | 0.614 | 0.418 | 0.393 | 0.516 | 0.552 | 0.412 |
| | GPT-4 | **0.864** | **0.626** | **0.480** | **0.649** | **0.679** | **0.546** |
| OUR | Semantic-Eval | <u>0.705</u> | <u>0.523</u> | <u>0.461</u> | 0.531 | 0.571 | 0.432 |

Table 3: The correlations between automatic metrics and human judgments on Sentiment Analysis Task. $r$, $\rho$, and $\tau$ denote the Pearson correlation coefficient, Spearman's rank correlation coefficient, and Kendall's tau coefficient, respectively. In each column, the top score is displayed in **bold** while the second highest is <u>underlined</u>.

tasks were used as reference texts, while texts generated by GPT-4o mini were used as candidate texts. For the two datasets in the sentiment analysis task, the original datasets include sentiment scores for individual texts. Since one objective of this experiment is to evaluate the effectiveness of Semantic-Eval in comparing reference and candidate texts in the sentiment analysis task, the focus was on verifying the alignment between Semantic-Eval's semantic understanding of the two texts and human judgment. Thus, we only used Llama-3-8B-Instruct as the candidate text and GPT-4o mini-generated text as the reference text.

### 4.3 Meta-evaluation

We evaluated the alignment between Semantic-Eval and human assessments across four NLP tasks involving textual data: natural language sentence similarity, text summarization, sentiment analysis, and natural language Q&A. To quantify this alignment, we employed three statistical coefficients: Pearson's correlation coefficient ($r$), Spearman's rank correlation coefficient ($\rho$), and Kendall's tau coefficient ($\tau$), which served as meta-evaluation metrics for comparing the assessment outputs with human judgments.

We validated the performance of Semantic-Eval on a sentence-level task involving annotated sentence similarity scores, with the experimental results presented in Table 1. The findings indicate that Semantic-Eval significantly outperforms traditional metrics and BERT-based approaches, aligning with human relevance judgments on the SemRel2024(eng) and STS-B datasets. Specifically, Semantic-Eval yields superior relevance scores for $r$, $\rho$, and $\tau$ compared to Qwen2-7B. Notably, on the SemRel2024(eng) dataset, Semantic-Eval exhibits a slight advantage over GPT-4 regarding $r$ and $\tau$. While Semantic-Eval's correlation scores on the STS-B dataset (i.e., $r$ = 0.787, $\rho$ = 0.807, $\tau$ = 0.635) are slightly lower than those of GPT-4, they nonetheless demonstrate a high degree of alignment with human evaluations, confirming the validity of Semantic-Eval in sentence-level tasks.

Additionally, we evaluated the effectiveness of Semantic-Eval on text-level tasks, which include sentiment analysis, text summarization, and natural language Q&A, with the corresponding. Tables 2, 3, and 4 detail the corresponding experimental results. Across all text-level tasks, GPT-4 consistently achieved the highest evaluation scores, most closely mirroring human judgments, while Semantic-Eval ranked second. Except for the Medical-o1 dataset, Semantic-Eval's $r$ scores exceeded 0.5 on five other datasets, reaching a peak value 0.705. Of particular note, on the CNN/Daily Mail dataset, Semantic-Eval's $\rho$ score surpassed that of GPT-4 by 1.1%. These results collectively reinforce the robustness of Semantic-Eval in accurately reflecting human judgments across a range of multi-task, text-level evaluations.

| Class | Metrics | IMHI | | | Medical-o1 | | |
|---|---|---|---|---|---|---|---|
| | | $r$ | $\rho$ | $\tau$ | $r$ | $\rho$ | $\tau$ |
| Traditional-Eval | BLEU-1 | 0.352 | 0.342 | 0.254 | 0.272 | 0.258 | 0.175 |
| | BLEU-2 | 0.448 | 0.429 | 0.321 | 0.291 | 0.289 | 0.190 |
| | ROUGE-1 | 0.383 | 0.343 | 0.256 | 0.292 | 0.287 | 0.196 |
| | ROUGE-2 | 0.415 | 0.404 | 0.310 | 0.282 | 0.295 | 0.202 |
| | ROUGE-L | 0.348 | 0.309 | 0.229 | 0.284 | 0.293 | 0.201 |
| | METEOR | 0.371 | 0.342 | 0.258 | 0.291 | 0.296 | 0.203 |
| BERT-based | BERT | 0.398 | 0.471 | 0.355 | 0.243 | 0.246 | 0.169 |
| | BERTScore_$P$ | 0.515 | 0.480 | 0.366 | 0.311 | 0.330 | 0.227 |
| | BERTScore_$R$ | 0.465 | 0.482 | 0.364 | 0.275 | 0.256 | 0.175 |
| | BERTScore_$F1$ | 0.503 | 0.499 | 0.379 | 0.309 | 0.311 | 0.213 |
| LLM-based | Qwen2-7B | 0.562 | 0.595 | 0.476 | 0.403 | 0.425 | 0.305 |
| | GPT-4 | **0.806** | **0.770** | **0.658** | **0.618** | **0.677** | **0.515** |
| OUR | Semantic-Eval | <u>0.637</u> | <u>0.636</u> | <u>0.507</u> | <u>0.473</u> | <u>0.445</u> | <u>0.320</u> |

Table 4: The correlations between automatic metrics and human judgments on Natural Language Q&A Task. $r$, $\rho$, and $\tau$ denote the Pearson correlation coefficient, Spearman's rank correlation coefficient, and Kendall's tau coefficient, respectively. In each column, the top score is displayed in **bold** while the second highest is <u>underlined</u>.

| Datasets | Settings | $r$ | $\rho$ |
|---|---|---|---|
| IMDB | w/o SemanticRank | 0.646 | 0.415 |
| | w/o roberta-large-mnli | 0.681 | 0.458 |
| | Semantic-Eval | 0.705 | 0.523 |
| IMHI | w/o SemanticRank | 0.595 | 0.583 |
| | w/o roberta-large-mnli | 0.612 | 0.576 |
| | Semantic-Eval | 0.637 | 0.636 |
| XSUM | w/o SemanticRank | 0.630 | 0.601 |
| | w/o roberta-large-mnli | 0.583 | 0.554 |
| | Semantic-Eval | 0.699 | 0.661 |

Table 5: Text-level Pearson $r$ and Spearman's rank $\rho$ correlations of ablation models in settings of Datasets.

### 4.4 Ablation Study

To further investigate the influence of each component of Semantic-Eval on its alignment with human preferences, we conducted ablation studies on the system. The results, as presented in Table 5, demonstrate that Semantic-Eval, when incorporating the SemanticRank module, exhibits a closer alignment with human ratings across three text-level task datasets compared to the version of Semantic-Eval that omits this module. This underscores the importance of assigning semantic weight distributions to individual sentences within a text. Furthermore, we visualized the outputs generated by the SemanticRank module for the three text-level tasks, with detailed illustrations provided in Appendix B. Additionally, our analysis reveals that Semantic-Eval aligns more effectively with human ratings when equipped with the NIL inference model, particularly on the XSUM dataset, where the discrepancy between the human ratings and the model's outputs is notably smaller. These findings emphasize the significance of incorporating semantic relationships within the text when evaluating text similarity.

| Settings | $r$ | $\rho$ | $\tau$ |
|---|---|---|---|
| T(0.1) | 0.554 | 0.583 | 0.429 |
| T(0.2) | 0.542 | 0.571 | 0.420 |
| T(0.3) | 0.554 | 0.583 | 0.429 |
| T(0.4) | 0.588 | 0.622 | 0.459 |
| T(0.5) | 0.606 | **0.641** | 0.473 |
| T(0.6) | **0.610** | **0.641** | **0.474** |
| T(0.7) | 0.554 | 0.584 | 0.432 |
| T(0.8) | 0.544 | 0.574 | 0.423 |
| T(0.9) | 0.542 | 0.572 | 0.422 |

Table 6: The correlations between automatic metrics and human judgments on CNN/Daily Mail dataset. T(number) denotes the threshold for filtering the cosine similarity of the embedding between two nodes.$r$, $\rho$, and $\tau$ denote the Pearson correlation coefficient, Spearman's rank correlation coefficient, and Kendall's tau coefficient, respectively. In each column, the top score is displayed in **bold**.

Additionally, we investigated the impact of varying the threshold parameter from 0.1 to 0.9 (in increments of 0.1) on the performance of the Semantic-Eval framework using the CNN/Daily

Mail dataset, as presented in Table 6. As the threshold increases from 0.1 to 0.6, the Pearson correlation coefficient $r$, Spearman's rank correlation coefficient $\rho$, and Kendall's tau $\tau$ generally improve. However, once the threshold surpasses 0.6 (e.g., 0.7 or higher), these correlation metrics begin to decline. This pattern suggests that, for the CNN/Daily Mail summarization task, a low threshold introduces excessive noise by retaining spurious edges, while a high threshold excessively prunes the graph, eliminating potentially informative edges. As shown in Table 6, a threshold of 0.6 yields the highest, or jointly highest, values for all three correlation measures. Notably, the Spearman correlation coefficient $\rho$ remains unchanged at thresholds 0.5 and 0.6. The observed differences between the maximum and minimum values for $r$, $\rho$, and $\tau$ are 0.068, 0.070, and 0.054, respectively. Finally, to demonstrate the construction process of SemanticRank more intuitively under different threshold conditions, we visualized the candidate text Y results in Figure 2, as shown in Appendix Figure 4.

### 4.5 Evaluation LLMs using Semantic-Eval

Semantic-Eval is employed to assess the quality of text generation from 13 general-purpose LLMs on a text-level task dataset, focusing on semantic understanding. To provide a more intuitive representation of the performance differences across LLMs, we visualized the evaluation results for these models on the XSUM dataset, as depicted in Figure 3. Additionally, the evaluation results of

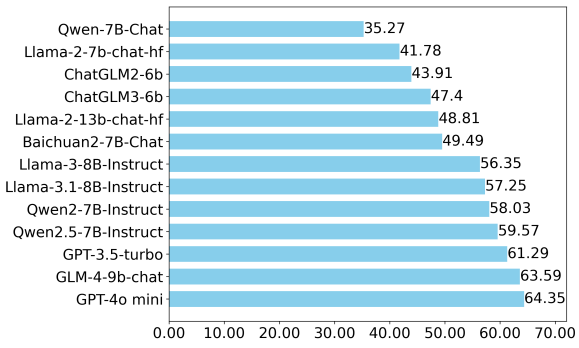



Figure 3: The visualization results of LLMs on the XSUM dataset assessed by Semantic-Eval.

Semantic-Eval for the LLMs on five other text-level task datasets are presented in Appendix B. Figure 3 illustrates that, for the text summarization task, GPT-4o mini achieved the highest score of 64.35, followed by GLM-4-9b-chat in second place, while the Qwen-7B-Chat model received the lowest score. Notably, half of the LLMs scored above 55 points.

## 5 Conclusion

In this paper, we introduce Semantic-Eval, the first framework designed for automatically evaluating text produced by LLMs without requiring training. The framework emphasizes assessing generated text quality from the perspective of semantic understanding. Semantic-Eval utilizes a graph-based weighting mechanism to evaluate the interdependence of semantic units within a text. Sentence-level similarities are computed using semantic embeddings, and a NLI model is integrated to address potential pairwise relationships. When tested across various NLP tasks and datasets, including sentiment analysis, summarization, question answering, and sentence similarity, Semantic-Eval outperforms existing evaluation metrics and demonstrates a closer alignment with human judgments. Additionally, Semantic-Eval is employed to evaluate the text quality generated by thirteen distinct large language models.

## 6 Limitations

Semantic-Eval is designed to evaluate the quality of text generated by LLMs from the perspective of semantic understanding, without the need for additional training. However, its performance is inherently limited by the constraints of the pre-trained sentence embeddings and NLI models upon which it relies. While Semantic-Eval demonstrates a higher alignment with human preferences compared to smaller parameter LLMs, its performance still lags behind state-of-the-art models, such as GPT-4. This discrepancy suggests that there is significant potential for improvement, particularly in refining the underlying models used for semantic analysis. In future work, we will focus on enhancing the precision of Semantic-Eval in aligning with human preferences, while preserving its training-free nature.

## 7 Acknowledgements

This work was supported by the National Key Research and Development Program of China(with 2022YFB4703100) and the National Natural Science Foundation of China (62471122), Fundamental Research Funds for the Central Universities(N25GFZ017, N25BJD005).

## References


Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anirudh Ajith, Chris Pan, Mengzhou Xia, Ameet Deshpande, and Karthik Narasimhan. 2023. Instructeval: Systematic evaluation of instruction selection methods. *arXiv preprint arXiv:2307.00259*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2024a. Dual-reflect: Enhancing large language models for reflective translation through dual learning feedback mechanisms. *arXiv preprint arXiv:2406.07232*.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024b. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Aparna Elangovan, Ling Liu, Lei Xu, Sravan Bodapati, and Dan Roth. 2024. Considers-the-human evaluation framework: Rethinking human evaluation for generative large language models. *arXiv preprint arXiv:2405.18638*.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023. Lawbench: Benchmarking legal knowledge of large language models. *arXiv preprint arXiv:2309.16289*.

Xueyang Feng, Zhi-Yuan Chen, Yujia Qin, Yankai Lin, Xu Chen, Zhiyuan Liu, and Ji-Rong Wen. 2024. Large language model-based human-agent collaboration for complex task solving. *arXiv preprint arXiv:2402.12914*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. 2024. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36.

Zhengping Jiang, Yining Lu, Hanjie Chen, Daniel Khashabi, Benjamin Van Durme, and Anqi Liu. 2024. Rora: Robust free-text rationale evaluation. *arXiv preprint arXiv:2402.18678*.

Pei Ke, Bosi Wen, Andrew Feng, Xiao Liu, Xuanyu Lei, Jiale Cheng, Shengyuan Wang, Aohan Zeng, Yuxiao Dong, Hongning Wang, et al. 2024. Critiquellm: Towards an informative critique generation model for evaluation of large language model generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13034–13054.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1. Minneapolis, Minnesota.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. 2024. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*.

Mateusz Lango and Ondřej Dušek. 2023. Critic-driven decoding for mitigating hallucinations in data-to-text generation. *arXiv preprint arXiv:2310.16964*.

Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. 2023. Huatuo-26m, a large-scale chinese medical qa dataset. *arXiv preprint arXiv:2305.01526*.

Qintong Li, Leyang Cui, Lingpeng Kong, and Wei Bi. 2025. Exploring the reliability of large language models as customized evaluators for diverse nlp tasks. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10325–10344.

Yang Li, Jie Ma, Miguel Ballesteros, Yassine Benajiba, and Graham Horwood. 2024. Active evaluation acquisition for efficient llm benchmarking. *arXiv preprint arXiv:2410.05952*.

Jingcong Liang, Rong Ye, Meng Han, Ruofei Lai, Xinyu Zhang, Xuanjing Huang, and Zhongyu Wei. 2024. Debatrix: Multi-dimensinal debate judge with iterative chronological analysis based on llm. *arXiv preprint arXiv:2403.08010*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

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

Yen-Ting Lin and Yun-Nung Chen. 2023. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. *arXiv preprint arXiv:2305.13711*.

Chuang Liu, Renren Jin, Yuqi Ren, Linhao Yu, Tianyu Dong, Xiaohan Peng, Shuting Zhang, Jianxiang Peng, Peiyi Zhang, Qingqing Lyu, et al. 2023. M3ke: A massive multi-level multi-subject knowledge evaluation benchmark for chinese large language models. *arXiv preprint arXiv:2305.10263*.

Hongwei Liu, Zilong Zheng, Yuxuan Qiao, Haodong Duan, Zhiwei Fei, Fengzhe Zhou, Wenwei Zhang, Songyang Zhang, Dahua Lin, and Kai Chen. 2024. Mathbench: Evaluating the theory and application proficiency of llms with a hierarchical mathematics benchmark. *arXiv preprint arXiv:2405.12209*.

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*.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Seyed Mahed Mousavi, Simone Alghisi, and Giuseppe Riccardi. 2024. Dyknow: dynamically verifying time-sensitive factual knowledge in llms. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8014–8029.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*.

Bang Nguyen, Mengxia Yu, Yun Huang, and Meng Jiang. 2024. Reference-based metrics disprove themselves in question generation. *arXiv preprint arXiv:2403.12242*.

Nedjma Ousidhoum, Shamsuddeen Hassan Muhammad, Mohamed Abdalla, Idris Abdulmumin, Ibrahim Said Ahmad, Sanchit Ahuja, Alham Fikri Aji, Vladimir Araujo, Abinew Ali Ayele, Pavan Baswani, et al. 2024. Semrel2024: A collection of semantic textual relatedness datasets for 14 languages. *arXiv preprint arXiv:2402.08638*.

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.

ChaeHun Park, Minseok Choi, Dohyun Lee, and Jaegul Choo. 2024. Paireval: Open-domain dialogue evaluation with pairwise comparison. *arXiv preprint arXiv:2404.01015*.

Bo Peng, Xinyi Ling, Ziru Chen, Huan Sun, and Xia Ning. 2024. ecellm: Generalizing large language models for e-commerce from large-scale, high-quality instruction data. *arXiv preprint arXiv:2402.08831*.

Felipe Maia Polo, Ronald Xu, Lucas Weber, Mírian Silva, Onkar Bhardwaj, Leshem Choshen, Allysson Flavio Melo de Oliveira, Yuekai Sun, and Mikhail Yurochkin. 2024. Efficient multi-prompt evaluation of llms. *arXiv preprint arXiv:2405.17202*.

Nils Reimers et al. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2020. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Hwanjun Song, Hang Su, Igor Shalyminov, Jason Cai, and Saab Mansour. 2024. Finesure: Fine-grained summarization evaluation using llms. *arXiv preprint arXiv:2407.00908*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, et al. 2023. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. *arXiv preprint arXiv:2306.05087*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022a. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022b. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *arXiv preprint arXiv:2204.07705*.

Ishaan Watts, Varun Gumma, Aditya Yadavalli, Vivek Seshadri, Manohar Swaminathan, and Sunayana Sitaram. 2024. Pariksha: A large-scale investigation of human-llm evaluator agreement on multilingual and multi-cultural data. *arXiv preprint arXiv:2406.15053*.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Mentallama: interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 4489–4500.

Zhuohao Yu, Chang Gao, Wenjin Yao, Yidong Wang, Wei Ye, Jindong Wang, Xing Xie, Yue Zhang, and Shikun Zhang. 2024. Kieval: A knowledge-grounded interactive evaluation framework for large language models. *arXiv preprint arXiv:2402.15043*.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2023. Evaluating large language models at evaluating instruction following. *arXiv preprint arXiv:2310.07641*.

Jing Zhang, Xiaokang Zhang, Daniel Zhang-Li, Jifan Yu, Zijun Yao, Zeyao Ma, Yiqi Xu, Haohua Wang, Xiaohan Zhang, Nianyi Lin, et al. 2023a. Glm-dialog: Noise-tolerant pre-training for knowledge-grounded dialogue generation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5564–5575.

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

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. 2023b. Evaluating the performance of large language models on gaokao benchmark. *arXiv preprint arXiv:2305.12474*.

Zhihan Zhang, Yixin Cao, Chenchen Ye, Yunshan Ma, Lizi Liao, and Tat-Seng Chua. 2024. Analyzing temporal complex events with large language models? a benchmark towards temporal, long context understanding. *arXiv preprint arXiv:2406.02472*.

Ziyin Zhang, Yikang Liu, Weifang Huang, Junyu Mao, Rui Wang, and Hai Hu. 2023c. Mela: Multilingual evaluation of linguistic acceptability. *arXiv preprint arXiv:2311.09033*.

Shuyan Zhou, Uri Alon, Sumit Agarwal, and Graham Neubig. 2023. Codebertscore: Evaluating code generation with pretrained models of code. *arXiv preprint arXiv:2302.05527*.

# A Experimental Setting

## A.1 Metrics Baselines

### A.1.1 Traditional-Eval

**BLEU** (Papineni et al., 2002) is a scoring mechanism primarily based on the n-gram overlap between the generated text and the reference text, which is used to evaluate the quality of the generated text. BLEU-1,2 is used as one of the metric baselines.

**ROUGE** (Lin, 2004) is one of the improved methods based on BLEU, which evaluates the accuracy of the generated text by calculating the overlap between the automatically generated text and the reference text.ROUGE-1,2 and ROUGE-L are used as metric baselines.

**METEOR**(Banerjee and Lavie, 2005) also improves BLEU and assesses quality mainly by calculating the degree of match between the machine translation result and the reference translation. It considers lexical variations and allows synonym matching.

### A.1.2 BERT-based

**BERT**(Devlin et al., 2018) is a pre-trained language model based on the transformer(Vaswani et al., 2017) model, which captures the semantic relations between sentences by introducing a bidirectional encoding mechanism.

**BERTScore**(Zhang et al., 2019) provides an overall picture of the quality of text generation by calculating Precision, Recall, and F1 scores between candidate and reference sentences. In particular, Precision measures the proportion of semantic information for each word in the generated text that occurs in the reference text, Recall measures the proportion of semantic information for each word in the reference text that occurs in the generated text, and F1 score is the reconciled average of Precision and Recall, which is used to combine Precision and Recall.

### A.1.3 LLM-based

**Qwen2-7B**[4] and **GPT-4**[5] (gpt-4-turbo-preview) generalized LLMs are employed as one of the metrics baselines. When LLM-based methods are used as raters, we design the simplest prompts to guide these LLMs to score paired texts. The prompt is:*"Please score the following two texts on a scale of 0-1."*

## A.2 Large Language Models

We compare the quality of generated text for 13 generalized large language models for LLMs: Llama-3.1-8B-Instruct[6], Llama-2-13b-chat-hf[7], Llama-2-7b-chat-hf[8], Llama-3-8B-Instruct[9], Qwen-7B-Chat[10], Qwen2-7B-Instruct[11], Qwen2.5-7B-Instruct[12], chatglm3-6b-32k[13], chatglm3-6b[14], chatglm2-6b[15], glm-4-9b-chat[16], Baichuan2-7B-Chat[17], GPT-3.5-turbo[18], GPT-4o mini[19].

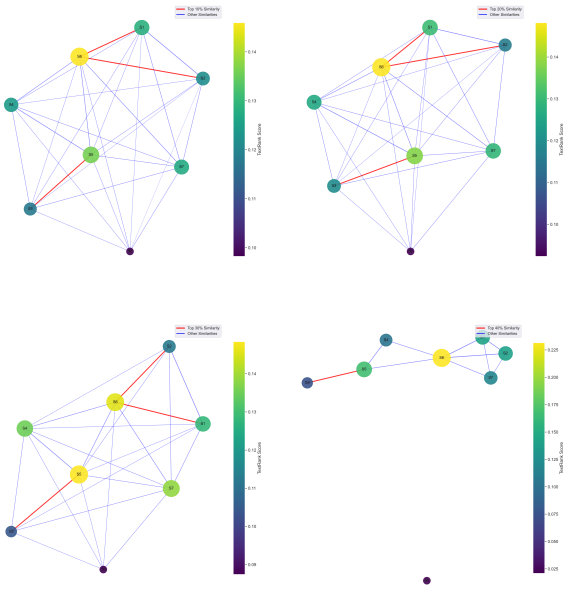


Figure 4: The visualization shows the construction of SemanticRank under different threshold conditions.

[4] https://huggingface.co/Qwen/Qwen2-7B-Instruct

[5] https://platform.openai.com/docs/models#gpt-4-turbo-and-gpt-4

[6] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[7] https://huggingface.co/meta-llama/Llama-2-13b-chat-hf

[8] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

[9] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

[10] https://huggingface.co/Qwen/Qwen-7B-Chat

[11] https://huggingface.co/Qwen/Qwen2-7B-Instruct

[12] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

[13] https://huggingface.co/THUDM/chatglm3-6b-32k

[14] https://huggingface.co/THUDM/chatglm3-6b

[15] https://huggingface.co/THUDM/chatglm2-6b

[16] https://huggingface.co/THUDM/glm-4-9b-chat

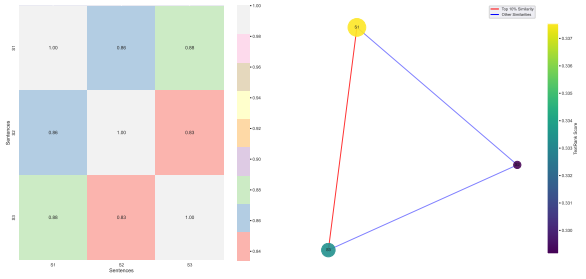
[17] https://huggingface.co/baichuan-inc/Baichuan2-7B-Chat

[18] https://platform.openai.com/docs/models#gpt-3.5-turbo

[19] https://platform.openai.com/docs/model#gpt-4o-mini

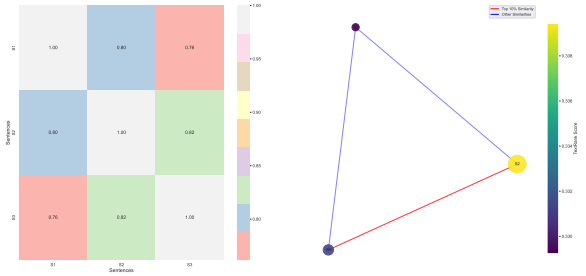
## B Semantic-Eval Evaluates LLMs

One data sample was selected from each of the five datasets in the three text-level tasks, and the SemanticRank module visualized the construction process for each of the different data samples, as shown in the six figures below.



(a) Visualization of self-similarity matrix of reference text.

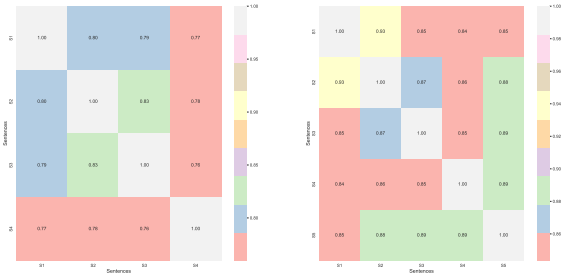
(b) Visualization of self-similarity matrix for candidate text.



(c) Visualization of self-similar TextRank plots for reference text.

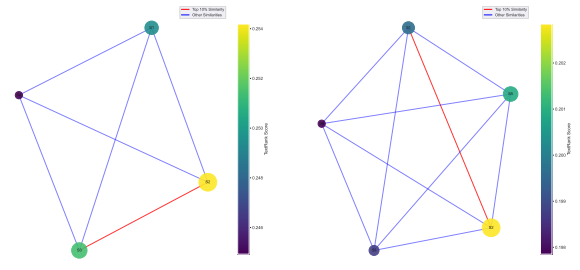
(d) Visualization of self-similar TextRank graphs for candidate texts.

Figure 6: The visualization shows the construction process of the SemanticRank on the sample on the Medical-o1 dataset.



(a) Visualization of self-similarity matrix of reference text.

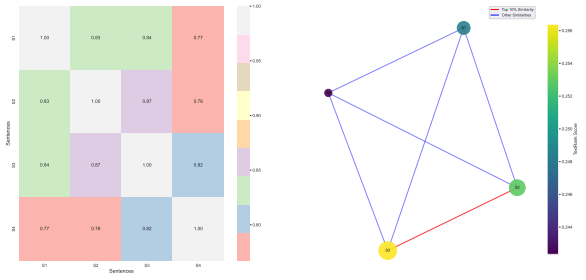
(b) Visualization of self-similarity matrix for candidate text.



(c) Visualization of self-similar TextRank plots for reference text.

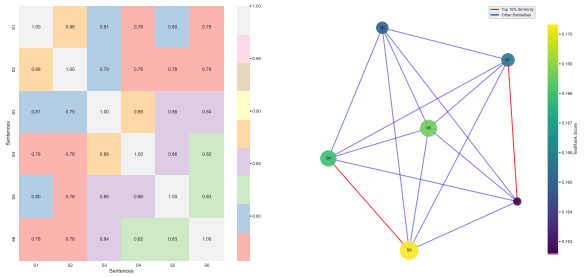
(d) Visualization of self-similar TextRank graphs for candidate texts.

Figure 5: The visualization shows the construction process of the SemanticRank on the sample on the IMHI dataset.



(a) Visualization of self-similarity matrix of reference text.

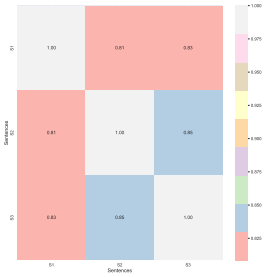
(b) Visualization of self-similarity matrix for candidate text.



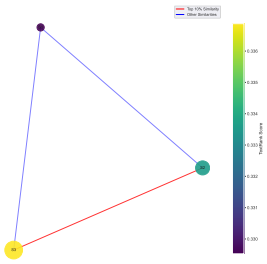
(c) Visualization of self-similar TextRank plots for reference text.

(d) Visualization of self-similar TextRank graphs for candidate texts.

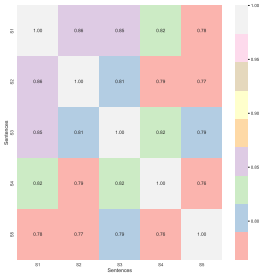
Figure 7: The visualization shows the construction process of the SemanticRank on the sample on the IMDB dataset.



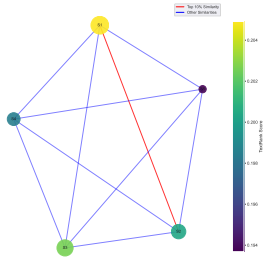
(a) Visualization of self-similarity matrix of reference text.



(b) Visualization of self-similarity matrix for candidate text.

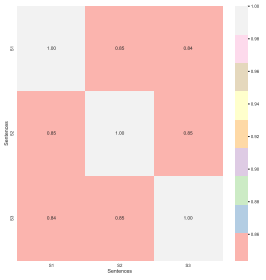


(c) Visualization of self-similar TextRank plots for reference text.

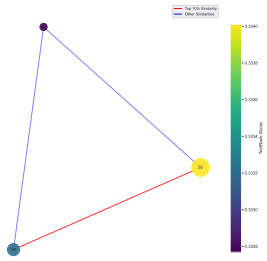


(d) Visualization of self-similar TextRank graphs for candidate texts.

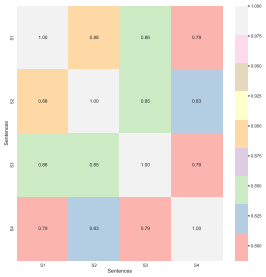
Figure 8: The visualization shows the construction process of the SemanticRank on the sample on the Yelp Polarity dataset.



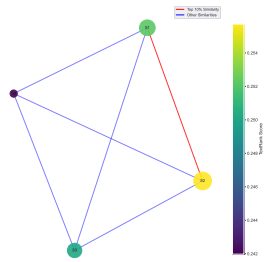
(a) Visualization of self-similarity matrix of reference text.



(b) Visualization of self-similarity matrix for candidate text.



(c) Visualization of self-similar TextRank plots for reference text.



(d) Visualization of self-similar TextRank graphs for candidate texts.

Figure 9: The visualization shows the construction process of the SemanticRank on the sample on the CNN/Daily Mail dataset.

## C Semantic-Eval evaluates LLMs

The following five figures show the results of Semantic-Eval's evaluation of LLMs on each of the five text-level task datasets.

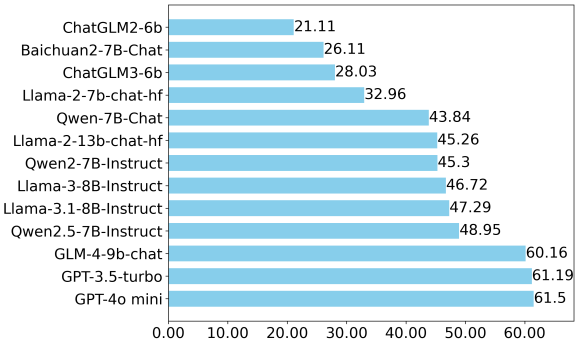



Figure 10: The visualization results of LLMs on the CNN/Daily Mail dataset assessed by Semantic-Eval.

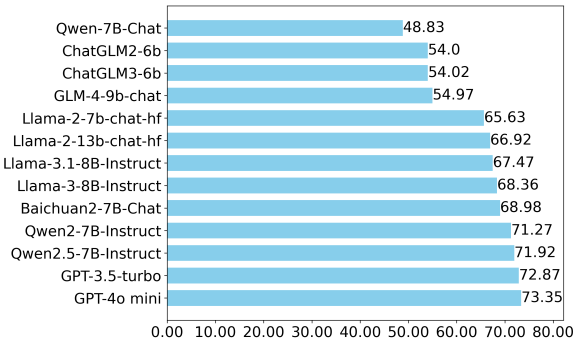



Figure 11: The visualization results of LLMs on the IMDB dataset assessed by Semantic-Eval.

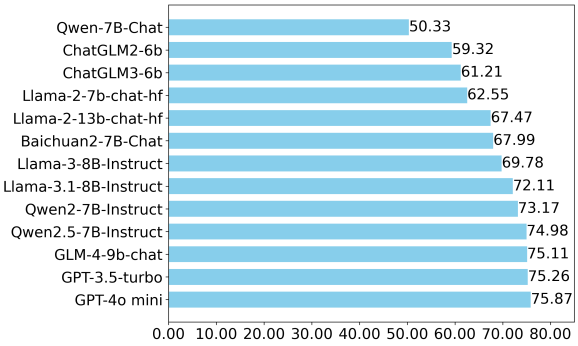



Figure 12: The visualization results of LLMs on the Yelp Polarity dataset assessed by Semantic-Eval.

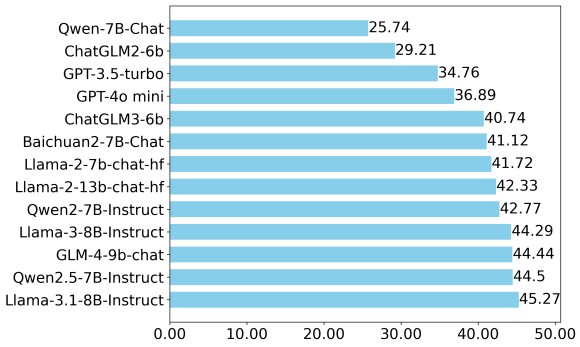



Figure 13: The visualization results of LLMs on the IMHI dataset assessed by Semantic-Eval.

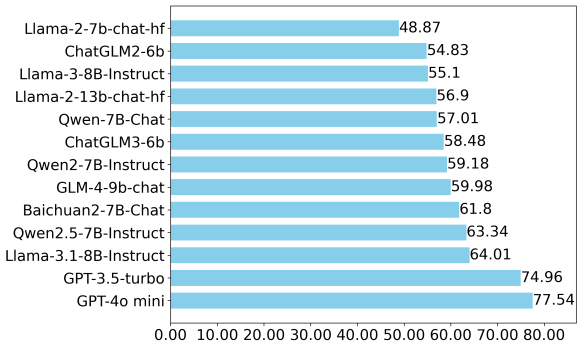



Figure 14: The visualization results of LLMs on the Medical-o1 dataset assessed by Semantic-Eval.