Bridging the LLM Accessibility Divide? Performance, Fairness, and Cost of Closed versus Open LLMs for Automated Essay Scoring

Kezia Oketch,[1] **John P. Lalor,**[1] **Yi Yang,**[2] **Ahmed Abbasi**[1]
[1]Human-centered Analytics Lab, University of Notre Dame
[2]Department of Information Systems, Business Statistics and Operations Management, HKUST
koketch@nd.edu, john.lalor@nd.edu, imyiyang@ust.hk, aabbasi@nd.edu

Abstract

Closed large language models (LLMs) such as GPT-4 have set state-of-the-art results across a number of NLP tasks and have become central to NLP and machine learning (ML)-driven solutions. Closed LLMs' performance and wide adoption has sparked considerable debate about their accessibility in terms of availability, cost, and transparency. In this study, we perform a rigorous comparative analysis of eleven leading LLMs, spanning closed, open, and open-source LLM ecosystems, across text assessment and generation within automated essay scoring, as well as a separate evaluation on abstractive text summarization to examine generalization. Our findings reveal that for few-shot learning-based assessment of human generated essays, open LLMs such as Llama 3 and Qwen 2.5 perform comparably to GPT-4 in terms of predictive performance, with no significant differences in disparate impact scores when considering age- or race-related fairness. For summarization, we find that open models also match GPT-4 in ROUGE and METEOR scores on the CNN/DailyMail benchmark, both in zero- and few-shot settings. Moreover, Llama 3 offers a substantial cost advantage, being up to 37 times more cost-efficient than GPT-4. For generative tasks, we find that essays generated by top open LLMs are comparable to closed LLMs in terms of their semantic composition/embeddings and ML assessed scores. Our findings challenge the dominance of closed LLMs and highlight the democratizing potential of open LLMs, suggesting they can effectively bridge accessibility divides while maintaining competitive performance and fairness.

1 Introduction

The rapid development of machine learning (ML) technologies, particularly large language models (LLMs), has led to major advancements in natural language processing (NLP, Abbasi et al., 2023). While much of this advancement happened under the umbrella of the common task framework which espouses transparency and openness (Abbasi et al., 2023), in recent years, closed LLMs such as GPT-3 and GPT-4 have set new performance standards in tasks ranging from text generation to question answering, demonstrating unprecedented capabilities in zero-shot and few-shot learning scenarios (Brown et al., 2020; OpenAI, 2023). Given the strong performance of closed LLMs such as GPT-4, many studies within the LLM-as-a-judge paradigm rely on their scores as ground truth benchmarks for evaluating both open and closed LLMs (Chiang and Lee, 2023), further entrenching the dominance of SOTA closed LLMs (Vergho et al., 2024). Along with closed LLMs, there are also LLMs where the pre-trained models (i.e., training weights) and inference code are publicly available ("open LLMs") such as Llama (Touvron et al., 2023; Dubey et al., 2024) as well as LLMs where the full training data and training code are also available ("open-source LLMs") such as OLMo (Groeneveld et al., 2024) and Prometheus (Kim et al., 2024). Open and open-source LLMs provide varying levels of transparency for developers and researchers (Liu et al., 2023).

Access to model weights, training data, and inference code enables several benefits for the user-developer-researcher community, including lower costs per input/output token through third-party API services, support for local/offline pre-training and fine-tuning, and deeper analysis of model biases and debiasing strategies. However, the dominance of closed LLMs raises a number of concerns, including accessibility and fairness (Strubell et al., 2020; Bender, 2021; Irugalbandara et al., 2024). The accessibility divide in this context can be understood in three dimensions: uneven availability due to geographic and economic barriers, prohibitive costs that limit adoption, and a lack of transparency that hinders research and innovation.

In the LLM space, corporate-driven commod-

Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM² 2025), pages 655–669
July 31 - August 1, 2025

ification through monopolized APIs and exclusive licensing is exacerbating the accessibility divide (Luitse and Denkena, 2021; Abbasi et al., 2024). These challenges are both technical and ethical, impacting who can access and benefit from the opportunities afforded by SOTA LLMs; those affected include researchers and practitioners residing in less affluent regions and/or complex sociopolitical environments. Open and open-source LLMs such as Llama 3, Qwen 2.5, and OLMo 2 provide greater transparency and customization potential (Touvron et al., 2023; Dubey et al., 2024; Bai et al., 2023; Groeneveld et al., 2024). As these models improve in general benchmarking tasks, there is a need to systematically compare open and open-source LLMs with their closed SOTA counterparts on different assessment/scoring and generation tasks across various dimensions including performance and fairness. We aim to address this gap by conducting a comprehensive comparative analysis of eleven LLMs, encompassing closed, open, and open-source LLMs, across multiple text generation and evaluation tasks. The Research Questions (RQs) guiding this study are: **RQ1**: How do different generations of open, open-source and closed LLMs compare in their assessment capabilities? **RQ2**: When performing assessments/scoring, to what extent do closed and open LLMs exhibit biases? **RQ3**: How comparable are open and open-source LLMs to their closed counterparts in terms of text generation capabilities?

To answer these questions, we use automated essay scoring (AES) as our focal context. AES is well-suited for our research questions; it has been studied extensively by the NLP community (Ke and Ng, 2019), entails prompt-guided text generation, has readily available large-scale human testbeds with demographic information, and includes well-defined evaluation rubrics.

Our contributions are three-fold: (1) we provide empirical evidence of the trade-offs between accuracy, cost, and fairness for LLMs when performing assessment/scoring tasks; (2) we statistically and visually demonstrate the text generation capabilities of leading open, open-source, and closed LLMs; (3) we highlight the growing viability of open and open-source LLMs as cost-effective alternatives to closed LLMs. To the best of our knowledge, this is the first study to compare the three LLM ecosystems, closed, open, and open-source, across multiple assessment and text generation tasks.[1]

2 Related Work

2.1 LLMs and Accessibility

Accessibility concerns can manifest in many ways, including the ability to serve those with physical impairments or cognitive impediments. Here, following prior work, we focus on accessibility as it relates to availability, cost, and transparency (Luitse and Denkena, 2021; Abbasi et al., 2024). Until recently, much of the progress in NLP representation learning and language modeling over the past 20 years occurred under the common task framework and transpired via publicly available, open and open-source LLMs, methods, algorithms, architectures, and systems (Abbasi et al., 2024, 2023). New proprietary LLMs such as GPT-4 are less available in lower- and middle-income countries due to inadequate internet penetration, underdeveloped infrastructure, and/or strict censorship policies (Wang et al., 2023).

Moreover, cost-efficiency is a critical factor influencing the adoption of LLMs for various NLP tasks. Strubell et al. (2020) examined the environmental and financial costs associated with training LLMs like GPT-3. Their findings suggest that the high costs are not only a barrier to accessibility but also raise concerns about the sustainability of such models. Furthermore, proprietary models like GPT-4, despite their strong performance, limit researchers' ability to scrutinize and mitigate biases due to their closed nature (Raji et al., 2020; Bommasani et al., 2021; Liao and Vaughan, 2023). In contrast, open and open-source LLMs, with their publicly available model weights and training data/code, offer greater traceability and scrutiny (Eiras et al., 2024).

2.2 The Performance of Open, Open-source, and Closed LLMs

The strong performance of closed LLMs such as GPT-3.5 and GPT-4 has led to their adoption as stand-in proxies for human assessors for ground-truth evaluation (Chiang and Lee, 2023). Such models have been used as judges in various studies related to the evaluation of open-ended tasks (An et al., 2024). For instance, Zheng et al. (2023a) found models such as GPT-4 can yield agreement rates of up to 80% with human experts. However,

[1] Our code is available on GitHub: https://github.com/nd-hal/llm-accessibility-divide.

the growing capabilities of open and open-source LLMs warrant a systematic comparison.

Prior work highlights that while closed LLMs often lead in terms of raw performance, open and open-source LLMs offer substantial cost advantages, making them more accessible to a wider range of users (Irugalbandara et al., 2024; Kukreja et al., 2024). Recently, Wolfe et al. (2024) examined the impact of fine-tuning smaller open LLMs versus employing few-shot learning for larger closed LLMs. Their results were mixed; for certain text classification problems, fine-tuning two open LLMs, Llama-2-7b and Mistral-7b, led to performance comparable to few-shot learning with GPT-4. For some other tasks, the fine-tuned closed LLMs attained markedly better classification performance. We build on this emergent literature by comparing open, open-source, and closed LLMs in terms of their generation, few-shot assessment/scoring, and fairness capabilities.

2.3 Automated Essay Scoring and LLMs

Automated Essay Scoring (AES) entails rule-based or ML model-based assessment of human-generated essays in response to different genres of prompts. Essays are scored against a defined evaluation rubric focusing on overall essay quality and/or aspect-oriented quality (Ke and Ng, 2019; Attali and Burstein, 2006). NLP models for AES have evolved from feature-based ML to RNN/CNN-based deep learning to the use of fine-tuned or few-shot-learned language models (Ke and Ng, 2019; Taghipour and Ng, 2016; Bevilacqua et al., 2023).

While AES models have improved, concerns about fairness and bias in AES have persisted. Ke and Ng (2019) highlighted that AES models could inadvertently reinforce biases present in training data, including those related to socioeconomic background or language proficiency. Schaller et al. (2024) explored strategies for mitigating such biases to ensure that AES systems produce fair and equitable scores. Bevilacqua et al. (2023) examined the behavior of ML assessment models scoring human- versus LLM-generated essays and found that assessors such as BERT and RoBERTa may exhibit a familiarity bias when scoring LLM-generated essays. As noted in the introduction, we use AES as our focal context to compare open and closed LLMs because of the familiarity of the problem to the NLP community, availability of large-human-generated text corpora, presence of different genres of text with clear prompts, and well-defined instructions and evaluation rubrics.

Data	Essay Type	N	Avg. Length	Score
ASAP				
1	A	1784	350	1 - 6
2	A	1800	350	1 - 6
3	R	1726	150	0 - 3
4	R	1772	150	0 - 3
5	R	1805	150	0 - 4
6	R	1800	150	0 - 4
7	N	1569	300	0 - 30
8	N	723	650	0 - 60
FCE				
1	L	1237	200-400	0 - 40
2	A,C,N,S	362	200-400	0 - 40
3	A,C,L,N	340	200-400	0 - 5
4	A,C,L,N	498	200-400	0 - 5
5a	A,C,L,S	15	200-400	0 - 5
5b	A,C,L	14	200-400	0 - 5

Table 1: Description of the data used in this study. *Avg. Length* gives the average essay length in number of words. *Score* lists the scoring range of the various essays. Essay types: argumentative (A), commentary (C), letter (L), suggestion (S), narrative (N), response (R).

3 Data, Models, and Experiments

To answer our three research questions, we developed a robust analysis framework (Figure 1). In the remainder of the section, we describe the data, models, and experiments in detail.

3.1 Human Text Data and Prompts

We use two human-generated essay datasets the Automated Student Assessment Prize (ASAP, Mathias and Bhattacharyya, 2018) and the Cambridge Learner Corpus-First Certificate in English exam (FCE, Yannakoudakis et al., 2011). The ASAP dataset is widely used as a benchmark dataset in the AES field (Taghipour and Ng, 2016; Jin et al., 2018), and consists of 12,979 essays across 8 prompts (Table 1). For all essays, we use the overall quality score. FCE is a large collection of texts produced by English language learners from around the world. Like ASAP, FCE is a widely recognized resource in NLP that has been used in previous benchmarking studies (Ramesh and Sanampudi, 2022; Ke and Ng, 2019). FCE assesses English at an upper-intermediate level. Test-takers were prompted to complete two writing tasks: a letter, a report, an article, a composition, or a short story. For each test-taker a composite score was given across the two tasks. FCE is comprised of 2,466 essays spanning 5 genres.

As depicted in Figure 1, we use these testbeds,

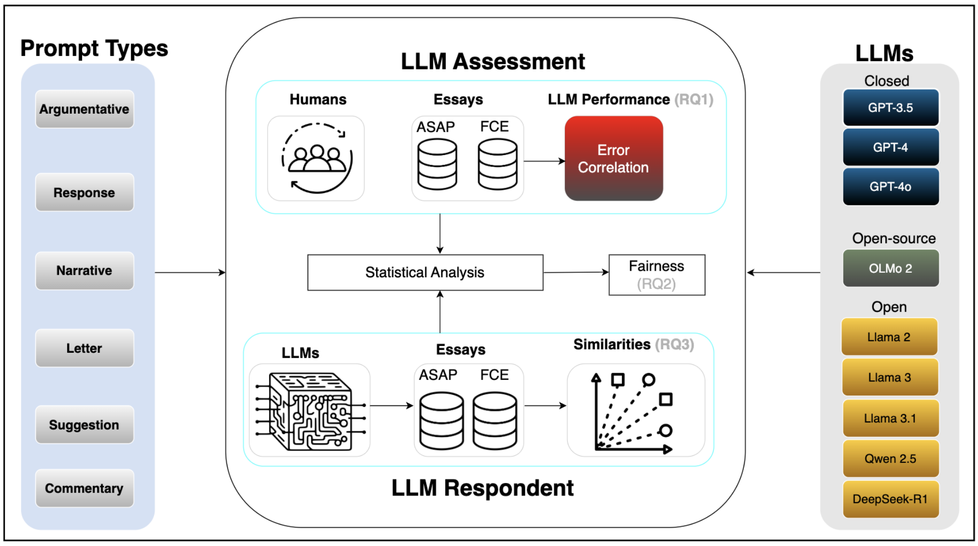

Figure 1: Human vs. LLM Essay Workflow by Prompt Type and Model Access

including the evaluation rubrics, directly as the input data for zero/few-shot-based LLM assessment (RQ1 and RQ2). We also use the six prompt types and associated instructions to generate essays with LLM respondents (RQ3).

3.2 Using LLMs for Assessment

Following prior work on zero and few-shot in-context learning (Chiang and Lee, 2023; Chen et al., 2023; Duan et al., 2024), and based on our first research question (RQ1), we evaluate the quality of text written by humans using LLMs for assessment/scoring. We present the LLM with the task instruction, description of the rating task, rating criteria, the sample to be rated, and a sentence that prompts the LLM to give the rating. The instructions, description, and rating criteria are presented exactly as they appear in our corpora. The rating sentence at the end of the prompt asks the LLM to rate the overall sample quality using a specified scale based on the original scoring range (Table 1). We tested two settings: zero-shot, where no example essays were provided, and few-shot, where in addition to the rubric and task instructions, three randomly selected human essays were provided along with their human expert ratings.[2] We intentionally selected one random sample per tertile from the human scoring range. LLM scores were normalized to a 0-1 range.

Consistent with RQ1, we compare the performance of LLMs for assessing human-generated text. Following prior research (Bevilacqua et al., 2023; Ramesh and Sanampudi, 2022; Ke and Ng, 2019), two categories of metrics were utilized. The first category comprised of two error metrics: mean squared error (MSE) and mean absolute error (MAE). The second category comprised of agreement and correlational metrics, specifically Quadratic Weighted Kappa (QWK), Pearson correlation coefficient (PCC), and Spearman's rank correlation (SRC).

3.3 LLMs Generating Textual Data

We followed prior work when designing our prompts for LLM essay generation (Bevilacqua et al., 2023; Zheng et al., 2023b). Specifically, we used the superset of prompts seen by human respondents across the ASAP and FCE. This resulted in nearly 150 prompts associated with 68 prompt IDs. To better align with a human text generation process, we used a zero-shot setting where the LLMs were provided the exact same instructions as humans, and did not see example essays as part of the prompts. For the GPT models, we provided essay prompts via the OpenAI API. For the Llama models, we used the Replicate API for Llama 2 and Llama 3, and the Llama API for Llama 3.1. For Qwen 2.5 and DeepSeek-R1, we used DeepInfra API. OLMo 2 was run locally. Each prompt was provided to the LLM 10 times resulting in 1,537 total essays for each model.[3] The LLM-generated essays are depicted in the bottom part of Figure

[2]We did not include OLMo 2 in the few-shot assessment task, as its smaller context window (4k) meant a large number of few-shot cases would have been excluded.

[3]GPT-4 and GPT-4o failed to respond to two/one of the 68 prompts resulting in 1,486 and 1,527 essays, respectively.

1 under "LLM Respondent" and inform our third research question (RQ3).

3.4 Statistical Analysis

For both RQ2 and RQ3, as noted in Figure 1, we used statistical models to allow us to parsimoniously examine the fairness and generation capabilities of open and closed LLMs while controlling for the types of prompts, specific prompt IDs, and assessment models.

3.4.1 Statistical Analysis for Fairness

For RQ2, we wanted to examine the fairness of the LLM assessors while controlling for prompt types/IDs, and the various assessment models. To achieve this, we ran a three-way ANOVA (split-plot design). We focused solely on human-generated essays appearing in the FCE corpus due to the availability of demographic information about the human authors. Following prior work, we define bias as representational harm from model error attributed to protect attributes such as demographics (Lalor et al., 2024). We used the available demographics in FCE, age (a) and race (r), as independent variables in separate ANOVA models. We also include prompt type (p) as an independent variable, as well as the assessment LLM employed (s); we also control for the specific prompt ID (d). The dependent variable (Δ_{R}) is the difference between the actual ground truth quality score for the essay (z), and the LLM score ($\hat{z}$). Hence, the statistical fairness ANOVA model is as follows:

$$\begin{aligned}
\Delta_{\mathrm{R}_{ijk}} = \frac{p_i}{d} + p_i + a_j + s_k + (pa)_{ij} + \quad & (ps)_{ik} + \\
(as)_{jk} + (pas)_{ijk} + \epsilon_{ijk} \quad & \textit{age} \\
\Delta_{\mathrm{R}_{ijk}} = \frac{p_i}{d} + p_i + r_j + s_k + (pr)_{ij} + \quad & (ps)_{ik} + \\
(rs)_{jk} + (prs)_{ijk} + \epsilon_{ijk} \quad & \textit{race}
\end{aligned}$$

Where $\Delta_{\mathrm{R}} = z - \hat{z}$, a is binarized into two groups: Young (25 and below) and Old (26 and above), r is binarized based on racial groups (Asian and Non-Asian), i,j,k refer to the factor category levels for p,a,s, respectively, and ϵ is the random error term.

3.4.2 Statistical Analysis for Generation

For RQ3, we wanted to examine the response generation commonalities and differences of various open and closed LLMs relative to one another and humans. Similar to the fairness statistical model, here, we controlled for prompt types/IDs, and the various assessment models. To achieve this, we ran another three-way ANOVA (split-plot design) setup. We used the full set of essays generated by humans (ASAP and FCE) and the six LLMs (across all ASAP/FCE prompts). The dependent variable is the assessment LLM score ($\hat{z}$). Instead of demographics, we use t to indicate the respondent type with seven possible values: one of the six LLMs or human. Once again, we include prompt type (p) as an independent variable, as well as the assessment LLM employed (s), and control for the prompt ID (d). Hence, the statistical response generation model is as follows:

$$\begin{aligned}
\hat{z} = \frac{p_i}{d} + p_i + t_j + s_k + (pt)_{ij} + (ps)_{ik} + \\
(ts)_{jk} + (pts)_{ijk} + \epsilon_{ijk}
\end{aligned}$$

Where i,j,k refer to the factor category levels for p,t,s, respectively, and ϵ is the random error term.

4 Results

4.1 Performance of LLMs for Assessment

Related to RQ1, we evaluated the assessment/scoring performance of LLMs when evaluating human-generated text with expert ground-truth labels. We present our benchmarking results in Table 2. Each of the eleven LLMs was presented with both human-generated and LLM-generated text. As noted, the dependent variable was normalized to a continuous scale ranging from 0 to 1. We applied two error metrics, MSE and MAE, along with three agreement and correlation measures, QWK, PCC, and SRC (Bevilacqua et al., 2023; Ramesh and Sanampudi, 2022; Ke and Ng, 2019). We also report macro-QWK (mQWK) which represents the arithmetic mean of QWK scores computed separately for each prompt to account for different score ranges, thus mitigating the effects of prompt imbalance and over-representation (Voskoboinik et al., 2025). For closed LLMs, GPT-4o demonstrated the best performance in both zero-shot and few-shot settings on the ASAP dataset, followed by GPT-4 and GPT-3.5, respectively. On the FCE dataset, however, GPT-4 achieved the highest performance, slightly outperforming GPT-4o, while GPT-3.5 remained the lowest among the closed models.

For open LLMs, Llama 3-70B achieved the highest overall performance on both ASAP and

FCE datasets, followed by Qwen 2.5, Llama 3.1, DeepSeek-R1, and Llama 2, in both zero-shot and few-shot conditions. Notably, the performance gap between zero-shot and few-shot settings is narrower for open LLMs compared to closed LLMs, suggesting that open models may be more stable across inference settings or benefit less from few-shot learning.

In particular, Qwen 2.5 (FS) and Llama 3 (FS) are highly competitive with GPT-4 (FS). Qwen 2.5 outperformed GPT-4 on MSE (0.185 vs. 0.296) and MAE (0.349 vs. 0.442), Llama 3 outperformed GPT-4 on QWK (0.357 vs. 0.246) while achieving comparable results on PCC and SRC when evaluated on the ASAP dataset. This highlights that certain open models are closing the performance gap with state-of-the-art closed models in structured evaluation tasks.

For the open-source LLM, OLMo 2 was evaluated in a zero-shot setting only. While its performance lags behind closed and open models, particularly in QWK (0.105 and 0.081), it remains competitive in correlation metrics (PCC: 0.201 and 0.214, SRC: 0.164 and 0.296), outperforming some open and closed models in their zero-shot settings. This suggests that, although open-source models may currently trail behind leading LLMs, they offer a viable alternative for users prioritizing transparency, cost-efficiency, and local deployment.

In regards to the performance of GPT-4 and Qwen 2.5, Figure 2 shows the MAE (left chart) and QWK (right chart) for the two LLMs across each of the six prompt types. In terms of MAE, Qwen 2.5's assessment score errors are comparable to those attained by GPT-4 for most prompt types, including response (RESP), commentary (COMM), letter (LETT), and suggestion (SUGG) essays. GPT-4 had slightly higher error rates for narrative (NARR), and markedly higher error when scoring argumentative (ARG) texts. For QWK, once again, GPT-4 and Qwen 2.5 were comparable, with GPT-4 attaining slightly better scores on letters, commentary and suggestions, while Qwen 2.5 scored higher on narratives and response. Overall, the results shed light on the assessment performance of top closed and open LLMs for different types of prompts and further underscore the closing performance gap between such models in the context of essay scoring.

4.2 Fairness Results

The results in Figure 3 depict the scoring error (y-axis) for each LLM (x-axis) on a given prompt type (the five charts). Differences between the two lines (e.g., non-Asian and Asian or older and younger authors) indicate biases. The results reveal that all 8 LLMs excluding OLMo 2 and Prometheus, exhibited relatively little bias. The relative error rates for Young/Old (bottom charts) and Asian/non-Asian (top charts) are comparable; that is, the two subgroup lines overlay one another. This is especially true for argument (ARG) and letter (LETT) essays. The two exceptions are commentaries (COMM) and suggestions (SUGG), where various LLMs do exhibit biases of up to 5% disparate impact (i.e., differences in scoring error rates attributable to race or age). These differences, although important to note, are relatively mild in terms of legal, practical, and policy implications (Lalor et al., 2022, 2024). Interestingly, GPT-4 and Llama 3 exhibit similar sub-group error profiles across prompt types. In the context of essay scoring, the results suggest that leading open LLMs may be comparable to SOTA closed LLMs in terms of their sub-group-level bias profiles across an array of prompt types.

4.3 Performance of LLMs for Generation

Regarding RQ3, we first present a t-SNE (t-Distributed Stochastic Neighbor Embedding) visualization (Van der Maaten and Hinton, 2008) of LLM-generated and human-written essays based on their BERT embeddings (Figure 4). This visualization supports the notion that while open and open-source LLMs like Qwen 2.5 and OLMo 2 respectively, are closing the gap with closed LLMs such as GPT-4, there remains a distinguishable difference between machine-generated and human-written texts. The relative proximity of LLM clusters to one another suggests that while some variability remains based on the specific model, overall these models produce essays with similar attributes.

To examine the assessment-generation interplay (RQ3), using the ANOVA model described in Section 3.4.2, analysis results depicting statistical significance for the main-effects, two-way, and three-way interactions are shown in Table 3. All the factors were significant ($p < 0.05$), suggesting that prompt-type, LLM/human respondent, and LLM assessor all significantly impact essay assessment scores (in terms of main effects, two-way, and three-way interactions). Figure 5 depicts the two-way interactions between assessment-respondent (left chart) and prompt-type-respondent (right chart). The assessment-respondent interactions show that LLMs tend to rate other LLM text higher than hu-

Model	Size	Release	ASAP Cost	ASAP MSE	ASAP MAE	ASAP mQWK*	ASAP QWK	ASAP PCC	ASAP SRC	FCE Cost	FCE MSE	FCE MAE	FCE mQWK*	FCE QWK	FCE PCC	FCE SRC
Closed LLMs																
GPT-3.5	175B	11/2022	$116.06	.233	.396	.206	.127	.178	.134	$27.12	.200	.617	.018	.039	.168	.161
				.244	.377	.894	.228	.412	.369		.843	.211	.367	.352	.227	.448
GPT-4	1T+	03/2023	$2815.19	.308	.452	.889	.269	.496	.444	$449.21	**.189**	.187	.460	**.541**	**.359**	.571
				.296	.442	.868	.246	.506	.464		.347	.171	.443	.378	.247	**.584**
GPT-4o	$\approx 200B$	11/2023	$577.49	.254	.423	.192	.143	.241	.209	$109.72	3.38	.677	.016	.031	.178	.145
				.143	**.299**	.908	.316	.557	.517		.545	**.168**	**.469**	.407	.233	.576
Open LLMs																
Llama 2	70B	07/2023	$77.03	1.232	.956	.175	.005	.034	.024	$14.64	.646	.268	.164	.137	.221	.349
				.232	.371	.878	.172	.106	.076		.644	.205	.219	.182	.193	.349
Llama 3	8B	04/2023	$6.32	.309	.397	.253	.205	.346	.337	$2.37	.648	.263	.002	-.036	.152	.198
				.898	.535	.516	.137	.069	.099		.439	.231	-.013	-.121	.126	.099
Llama 3	70B	04/2024	$75.21	.250	.421	.883	.214	.443	.403	$14.29	.601	.261	.148	.147	.199	.347
				.153	.303	**.947**	**.357**	.564	**.552**		.462	.186	.355	.326	.231	.484
Llama 3.1	405B	07/2024	$177.69	.288	.447	.854	.184	.438	.382	$43.26	.481	.235	.162	.255	.215	.409
				.239	.390	.924	.179	.441	.377		.513	.197	.307	.289	.225	.454
DeepSeek-R1	671B	01/2025	$75.52	.283	.442	.828	.179	.375	.327	$23.15	.536	.298	.035	.015	.177	.185
				.203	.353	.885	.203	.345	.310		.407	.239	.004	-.007	.145	.111
Qwen 2.5	72B	09/2024	$29.71	.254	.432	.873	.185	.442	.403	$12.33	.648	.283	.031	.053	.158	.167
				.185	.349	.924	.304	**.569**	.539		.484	.223	.023	.003	.146	.138
Open-Source LLMs																
Prometheus	13B	10/2023	$9.11	.342	.439	.549	.059	.105	.096	$4.27	1.310	.499	-.009	-.064	.154	.088
				.779	.661	.491	.026	.028	.028		.598	.286	.000	-.032	.104	.053
*OLMo 2	13B	11/2024	-	.283	.459	.235	.105	.201	.164	-	1.251	.436	.076	.081	.214	.296

Table 2: Performance metrics for benchmark models on ASAP and FCE under zero-shot (shaded) and few-shot (unshaded) settings. mQWK* = macro QWK averaged over prompts.

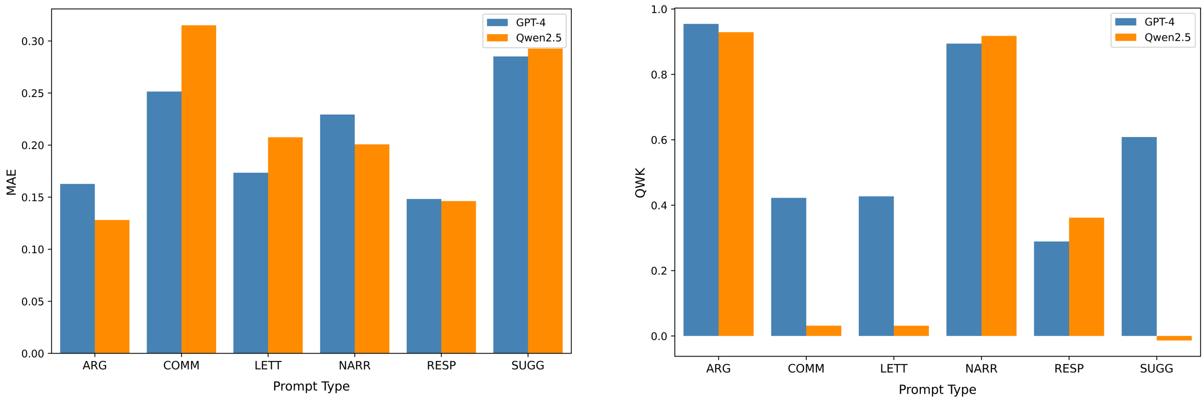

Figure 2: Few-shot results comparing GPT-4 and Qwen 2.5 across prompt types.

Term	DF	SS	MS	F-statistic
A (Prompt Type)	5	4.58e6	916900	62074.90^{***}
B (Respondent)	9	2.59e6	288144	19507.59^{***}
C (Assessor)	8	1.73e5	21674	1467.32^{***}
A × B	45	3.68e6	81787	5537.07^{***}
A × C	40	1.74e5	4355.04	294.84^{***}
B × C	71	2.22e3	31.26	2.12^{***}
A × B × C	355	6.54e3	18.42	1.25^{**}

***: $p < 0.001$

Table 3: Few-Shot ANOVA Results with Nine LLMs & Human Text.

man content (left chart). Moreover, when looking at the assessment LLMs with the lowest prediction error on humans, namely GPT-4, GPT-4o, Qwen 2.5, and Llama 3, they tend to rate GPT-4, Qwen 2.5, and Llama 3 generated essays the highest (left chart). These results are consistent across prompt types, with response essays (RESP) having the greatest variability (right chart). A detailed breakdown of assessment scores is provided in Appendix A.3 (8), illustrating these scoring trends.

4.4 Cost Analysis

To compare and contrast the cost-benefit trade-offs of open vs. closed LLMs, we computed the input and output token utilization cost of the LLMs across the assessment and generation tasks. In order to allow a fair comparison of cost, we compared the open and closed models when running both via APIs (i.e., we used the OpenAI, Replicate, Llama, and DeepInfra APIs). Figure 6 shows the eight LLMs and the cost in thousands (in USD) associated with input and output tokens per LLM. GPT-4 exhibits the highest input and output costs,

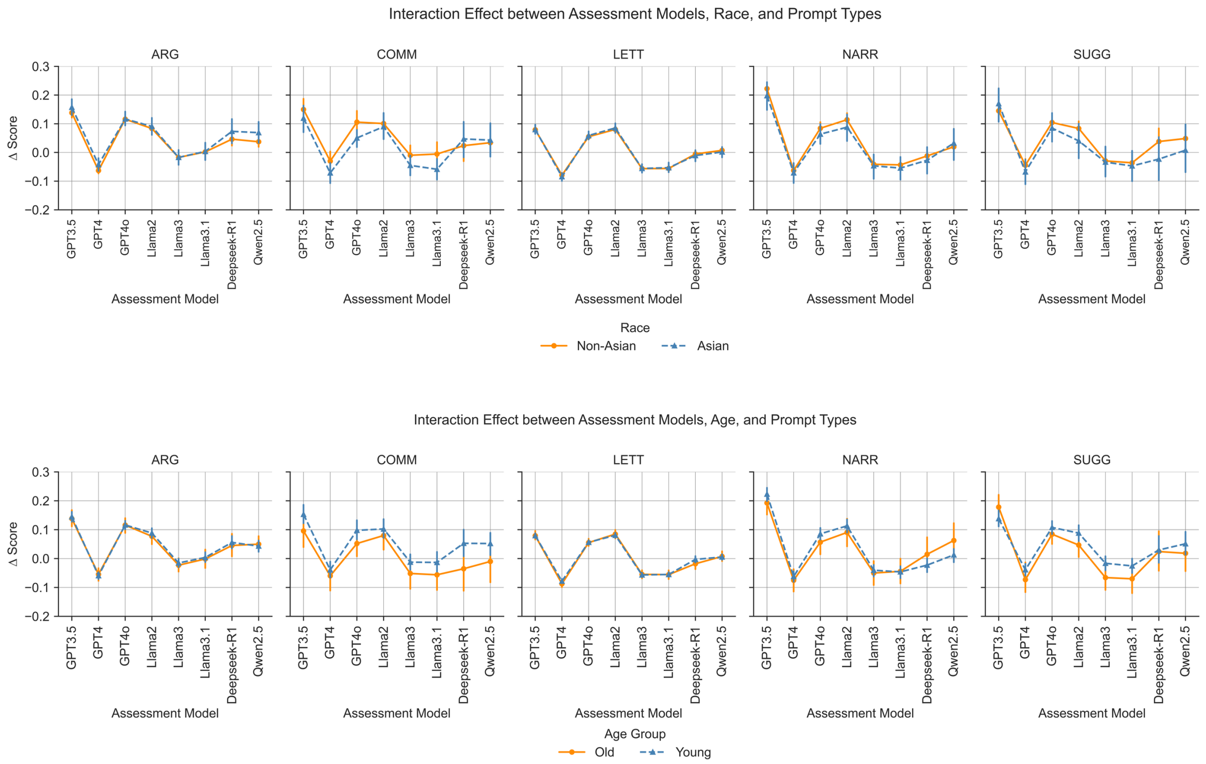

Figure 3: Few-shot Results Comparing Δ Scores (Human - LLM prediction) Across Assessment Models and Prompt Types. (left) Differences by Race, (right) Differences by Age

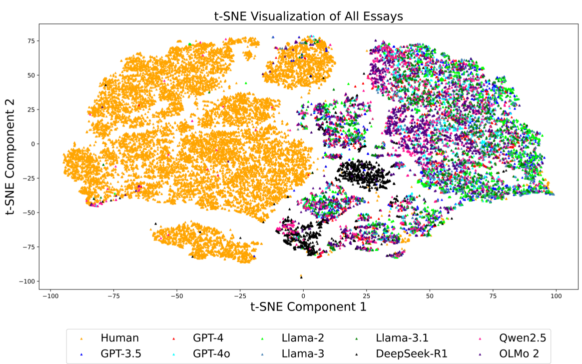

Figure 4: t-SNE plot of Human and LLM Generated Essays

reflecting its substantial computational resource requirements. In contrast, open LLMs such as Llama 3, DeepSeek-R1, and Qwen 2.5 demonstrate significantly lower costs (15-17 times lower than GPT-4), emphasizing their cost-efficiency for comparable performance relative to closed alternatives.

4.5 Further Analysis: Abstractive Summarization

To further assess the generalization and applicability of open versus closed LLMs beyond essay scoring, we extend our evaluation to the domain of abstractive text summarization (See et al., 2017) as described in Appendix B. We benchmark model performance on the CNN/DailyMail dataset (Hermann et al., 2015; Nallapati et al., 2016), a widely-used corpus for summarization tasks, using standard evaluation metrics including ROUGE-1, ROUGE-2, ROUGE-L, and METEOR. This additional task allows us to test whether the trends observed in AES hold in a more general-purpose generation setting. Results in Table 4 show that open models such as Llama 3.1 and Qwen 2.5 perform competitively with GPT-4 across both zero-shot and few-shot settings. GPT-4 achieved the highest ROUGE scores while Llama 3.1-405B attained the highest METEOR score. Open models approached GPT-4 within 1-2 points across all metrics, reinforcing our findings on the growing utility of open LLMs in a broader range of language tasks.

5 Discussion and Conclusion

This study contributes to the growing body of research exploring LLM accessibility divides. While the emerging literature has made some strides in evaluating the performance, bias, and costs asso-

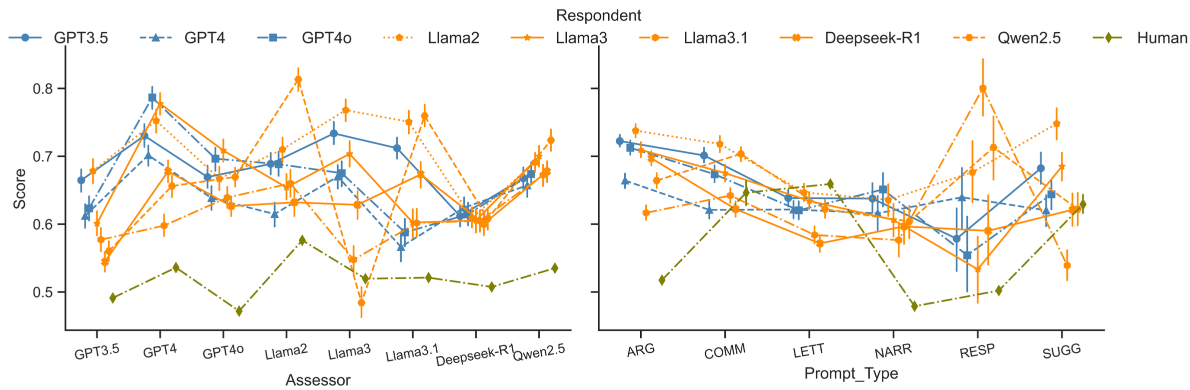

Figure 5: (left) Comparing Scores of Different LLM Assessors for LLMs/Human Generated Text, (right) Interaction Effect Between Respondent and Prompt. Blue Lines Denote Closed LLMs, Orange Denote Open LLMs

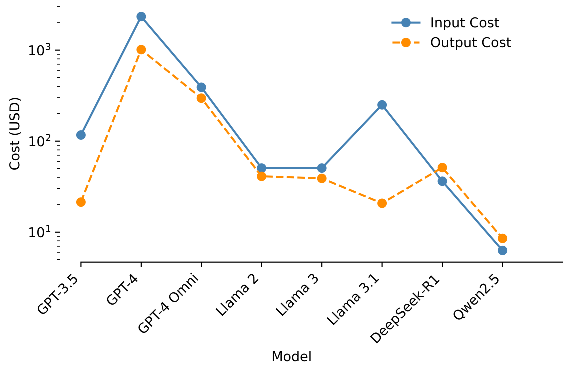

Figure 6: Input and Output Token Cost of Various LLMs across ASAP and FCE. The y-axis is log-scaled for readability. Costs calculated as of January 2025

Model	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
Closed LLMs				
GPT-3.5	0.116	0.043	0.078	0.089
	0.361	0.132	0.236	0.272
GPT-4	0.367	0.145	0.244	0.286
	0.371	**0.146**	**0.248**	0.283
GPT-4o	0.339	0.119	0.216	0.275
	0.354	0.125	0.227	0.268
Open LLMs				
Llama 2 70B	0.334	0.125	0.217	0.286
	0.342	0.129	0.225	0.278
Llama 3 8B	0.351	0.133	0.228	0.291
	0.352	0.134	0.231	0.286
Llama 3 70B	0.351	0.132	0.225	0.293
	0.361	0.138	0.235	0.293
Llama 3.1 405B	0.342	0.129	0.219	**0.296**
	0.233	0.064	0.154	0.189
Qwen2.5 72B	0.346	0.124	0.221	0.276
	0.363	0.133	0.235	0.269
Open-Source LLM				
Prometheus 13B	0.335	0.121	0.217	0.273
	0.345	0.127	0.227	0.269

Table 4: Summarization performance of LLMs on CNN/DailyMail (n=2000) in zero-shot (shaded) vs. few-shot (unshaded) conditions.

ciated with LLMs (Brown et al., 2020; OpenAI, 2023; Touvron et al., 2023; Bolukbasi et al., 2016; Buolamwini and Gebru, 2018; Raji et al., 2020; Strubell et al., 2020), our study offers an extensive, statistically robust multi-dimensional comparison that focuses strongly on the practical and ethical implications of model choice. The performance analyses demonstrate that while closed LLMs, particularly GPT-4, lead in raw performance metrics, the margin is small. Open LLMs like Qwen 2.5 and Llama 3 closely match GPT-4's performance. Additionally, the analysis of fairness of the models showed that top models maintained consistent Δ scores across race and age, indicating a low propensity for demographic bias when provided with context (i.e., few-shot learning).

Open LLMs such as Llama 3 offer substantial cost savings, being up to 37 times more cost-efficient than GPT-4. This cost advantage, combined with relatively comparable performance and fairness, positions newer open LLMs as attractive options, particularly for those operating with limited resources and/or in environments where greater transparency is important.

These findings have significant implications for the NLP community. The increasing viability of open LLMs more closely aligns with the principles of the common-task framework. The NLP community may continue to find greater value in adopting and contributing to open-source ecosystems, which promote innovation while ensuring equitable access to advanced AI technologies. To conclude, this study provides empirical evidence that challenges the dominance of closed LLMs in recent years by demonstrating the comparative performance, fairness, and cost-efficiency of open alternatives. Our findings underscore the democratizing potential of SOTA open LLMs.

Limitations

Our work is not without limitations. Recent research on LLM security suggests that open models may be more susceptible to security issues and attacks relative to their closed counterparts. Furthermore, although open LLMs are objectively more transparent – the inference code and tuned weights are not readily available for closed models – the massive size of open LLMs does raise questions about how explainable, interpretable, transparent, and scrutable multi-billion parameter LLMs can really be (Bender et al., 2021). Nevertheless, if existing in an LLM-powered world, we believe that relative to closed models, viable open LLM alternatives capable of alleviating availability,

Moreover, we chose to focus on three generations of closed and open GPT and Llama and one generation of Qwen and DeepSeek LLMs. Other viable alternatives such as Mistral, Falcon, and so forth could also have been included. We did so for financial/cost reasons, and to make the ANOVA plot results more manageable and readable. Limitations notwithstanding, our work contributes to the nascent emerging literature on LLM accessibility divides. Our hope is that future research can build upon our work. We intend to make all generated text, assessment data, statistical models, and analyses scripts publicly available as a resource for future evaluation research.

Lastly, we note that many open models (e.g., Llama 2, Llama 3) can also be downloaded and run locally. To ensure a fair cost comparison, we intentionally relied on API-based services for the closed (GPT) and open (Llama, Qwen, DeepSeek-R1) models, rather than running them on local or cloud-based servers, as done in some prior studies (Wolfe et al., 2024). However, we ran the OLMo 2 open-source model locally due to their full availability. This distinction highlights key trade-offs in accessibility: API-based models offer ease of use but involve ongoing costs, while locally run models—whether open or open-source—require technical setup and computational resources but eliminate API-related expenses in the long run.

Ethics Statement

This study adheres to the ACL Code of Ethics. All data used in this research is publicly available and has been previously collected and released for research purposes. No personally identifiable information is included. No human subjects were recruited for this study, and IRB approval was not required. We have released all code and data used in our evaluations to support reproducibility. We discuss the limitations in the previous section.

References

Ahmed Abbasi, Roger HL Chiang, and Jennifer J Xu. 2023. Data science for social good. *Journal of the AIS*.

Ahmed Abbasi, Jeffrey Parsons, Gautam Pant, Olivia R Liu Sheng, and Suprateek Sarker. 2024. Pathways for design research on artificial intelligence. *Information Systems Research*.

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2024. L-eval: Instituting standardized evaluation for long context language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14388–14411, Bangkok, Thailand. Association for Computational Linguistics.

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, 4(3).

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenhang Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, K. Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Yu Bowen, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xing Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *ArXiv*, abs/2309.16609.

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.

Lochan Basyal and Mihir Sanghvi. 2023. Text summarization using large language models: a comparative study of mpt-7b-instruct, falcon-7b-instruct, and openai chat-gpt models. *arXiv preprint arXiv:2310.10449*.

EM Bender. 2021. Bender, emily m., timnit gebru, angelina mcmillan-major, and shmargaret shmitchell. *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big*, pages 610–623.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Marialena Bevilacqua, Kezia Oketch, Ruiyang Qin, Will Stamey, Xinyuan Zhang, Yi Gan, Kai Yang, and Ahmed Abbasi. 2023. When automated assessment meets automated content generation: Examining text quality in the era of gpts. *arXiv preprint arXiv:2309.14488*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: An empirical study. *Preprint*, arXiv:2304.00723.

Cheng-Han Chiang and Hung-Yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631.

Hanyu Duan, Yixuan Tang, Yi Yang, Ahmed Abbasi, and Kar Yan Tam. 2024. Exploring the relationship between in-context learning and instruction tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

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*.

Francisco Eiras, Aleksander Petrov, Bertie Vidgen, Christian Schroeder, Fabio Pizzati, Katherine Elkins, Supratik Mukhopadhyay, Adel Bibi, Aaron Purewal, Csaba Botos, et al. 2024. Risks and opportunities of open-source generative ai. *arXiv e-prints*, pages arXiv–2405.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. 2024. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Chandra Irugalbandara, Ashish Mahendra, Roland Daynauth, Tharuka Kasthuri Arachchige, Jayanaka Dantanarayana, Krisztian Flautner, Lingjia Tang, Yiping Kang, and Jason Mars. 2024. Scaling down to scale up: A cost-benefit analysis of replacing openai's llm with open source slms in production. In *2024 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, pages 280–291. IEEE.

Cancan Jin, Ben He, Kai Hui, and Le Sun. 2018. Tdnn: a two-stage deep neural network for prompt-independent automated essay scoring. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1088–1097.

Zixuan Ke and Vincent Ng. 2019. Automated essay scoring: A survey of the state of the art. In *IJCAI*, volume 19, pages 6300–6308.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. *arXiv preprint arXiv:2405.01535*.

Sanjay Kukreja, Tarun Kumar, Amit Purohit, Abhijit Dasgupta, and Debashis Guha. 2024. A literature survey on open source large language models. In *Proceedings of the 2024 7th International Conference on Computers in Management and Business*, pages 133–143.

John P Lalor, Ahmed Abbasi, Kezia Oketch, Yi Yang, and Nicole Forsgren. 2024. Should fairness be a metric or a model? a model-based framework for assessing bias in machine learning pipelines. *ACM Transactions on Information Systems*, 42(4):1–41.

John P Lalor, Yi Yang, Kendall Smith, Nicole Forsgren, and Ahmed Abbasi. 2022. Benchmarking intersectional biases in nlp. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 3598–3609.

Q Vera Liao and Jennifer Wortman Vaughan. 2023. Ai transparency in the age of llms: A human-centered research roadmap. *arXiv preprint arXiv:2306.01941*, pages 5368–5393.

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

Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, et al. 2023. Llm360: Towards fully transparent open-source llms. *arXiv preprint arXiv:2312.06550*.

Dieuwertje Luitse and Wiebke Denkena. 2021. The great transformer: Examining the role of large language models in the political economy of ai. *Big Data & Society*, 8(2):20539517211047734.

Sandeep Mathias and Pushpak Bhattacharyya. 2018. Asap++: Enriching the asap automated essay grading dataset with essay attribute scores. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

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*.

Abdurrahman Odabaşı and Göksel Biricik. 2025. Unraveling the capabilities of language models in news summarization. *arXiv preprint arXiv:2501.18128*.

R OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5).

Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 33–44.

Dadi Ramesh and Suresh Kumar Sanampudi. 2022. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55(3):2495–2527.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

Nils-Jonathan Schaller, Yuning Ding, Andrea Horbach, Jennifer Meyer, and Thorben Jansen. 2024. Fairness in automated essay scoring: A comparative analysis of algorithms on german learner essays from secondary education. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 210–221.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2020. Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13693–13696.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1882–1891.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Tyler Vergho, Jean-Francois Godbout, Reihaneh Rabbany, and Kellin Pelrine. 2024. Comparing gpt-4 and open-source language models in misinformation mitigation. *arXiv preprint arXiv:2401.06920*.

Ekaterina Voskoboinik, Nhan Phan, Tamás Grósz, and Mikko Kurimo. 2025. Leveraging uncertainty for finnish l2 speech scoring with llms. In *The Workshop on Automatic Assessment of Atypical Speech*. University of Tartu Library.

Xiaofei Wang, Hayley M Sanders, Yuchen Liu, Kennarey Seang, Bach Xuan Tran, Atanas G Atanasov, Yue Qiu, Shenglan Tang, Josip Car, Ya Xing Wang, et al. 2023. Chatgpt: promise and challenges for deployment in low-and middle-income countries. *The Lancet Regional Health–Western Pacific*, 41.

Robert Wolfe, Isaac Slaughter, Bin Han, Bingbing Wen, Yiwei Yang, Lucas Rosenblatt, Bernease Herman, Eva Brown, Zening Qu, Nic Weber, et al. 2024. Laboratory-scale ai: Open-weight models are competitive with chatgpt even in low-resource settings. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1199–1210.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 180–189.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023a. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Mingkai Zheng, Xiu Su, Shan You, Fei Wang, Chen Qian, Chang Xu, and Samuel Albanie. 2023b. Can gpt-4 perform neural architecture search? *arXiv preprint arXiv:2304.10970*.

A Further Evaluations

A.1 Additional Few Shot Evaluation

Figure 7 presents an extension of our few-shot evaluation, comparing GPT-4 and Llama 3 across different prompt types. Consistent with our findings earlier, where Qwen 2.5 demonstrated strong performance relative to GPT-4, Llama 3 exhibits comparable effectiveness across multiple prompt types, further reinforcing the capability of open models. While GPT-4 maintains a slight advantage in COMM and SUGG, Llama 3 closely matches or outperforms GPT-4 in NARR, RESP, and ARG when measured by QWK. These results provide additional evidence that open LLMs are increasingly competitive with closed SOTA models.

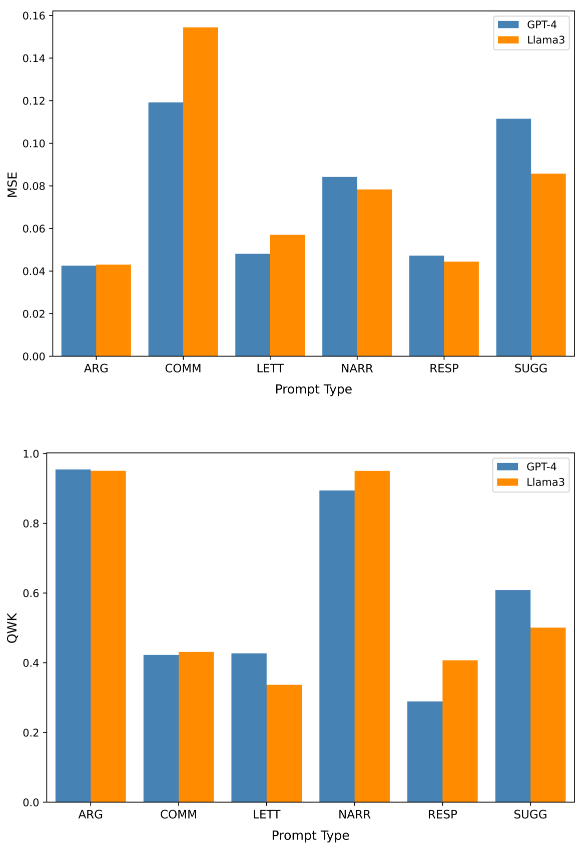

Figure 7: Few-shot Results Comparing GPT-4 and Llama 3 Across Prompt Types

A.2 LLM Assessment Scores Breakdown

Figure 8 presents average assessment scores assigned by different LLMs to essays generated by LLMs and human respondents. The red-to-green color scale highlights score variations, where green represents higher ratings and red represents lower ratings. This visualization further supports the trends observed in Figure 5, showing that LLM assessors tend to rate other LLM-generated text higher than human-written responses.

Generating	Assessment LLM				
LLM/Human	GPT4	GPT4o	Llama 3 70B	Qwen2.5-72B	DeepSeek-R1
GPT4	0.702	0.638	0.670	0.657	0.624
GPT-4o	0.787	0.696	0.675	0.674	0.614
Llama 3 70B	0.776	0.708	0.702	0.700	0.605
Qwen2.5-72B	0.656	0.669	0.484	0.724	0.608
DeepSeek-R1	0.674	0.627	0.626	0.676	0.597
Human	0.536	0.472	0.520	0.535	0.507

Figure 8: Average Assessment Scores of LLMs/Human-Generated Text by Different LLMs

A.3 QWK Scores per Prompt

To further understand model-level variability, we report prompt-level QWK scores across the ASAP and FCE datasets in Tables 5 and 6. These results reveal that performance varies across prompt types, consistent with prior findings that essay genre and rubric complexity can influence model agreement with human raters (Taghipour and Ng, 2016; Ke and Ng, 2019). For instance on ASAP, Llama 3-70B and GPT-4 achieve highest agreement on argumentative (prompt 1) and narrative (prompt 8) respectively in few-shot settings. In FCE, models tend to show lower agreement on commentary types (e.g., 26 and 44). This variation reflects known genre effects in AES and reinforces the value of prompt-level evaluation (Ke and Ng, 2019; Bevilacqua et al., 2023).

B Text Summarization

To extend our evaluation beyond essay scoring, we assessed the performance of open, open-source, and closed LLMs on the task of abstractive summarization using the CNN/DailyMail dataset (Hermann et al., 2015; Nallapati et al., 2016). Abstractive summarization involves generating a concise, paraphrased summary that captures the salient points of a source document, rather than simply extracting sentences verbatim (See et al., 2017; Rush et al., 2015).

B.1 Experimental Setup

We sampled 2,000 examples from the test set of CNN/DailyMail to evaluate model performance.

Model	Prompts							
	1	2	3	4	5	6	7	8
Closed LLMs								
GPT-3.5	.096	.174	.054	.127	.282	.257	.008	.019
	.329	.144	.191	.266	.287	.263	.169	.172
GPT-4	.261	.174	.218	.256	.252	.176	.305	**.517**
	.393	.244	.202	.247	.252	.198	.222	.207
GPT-4o	.084	.149	.186	.231	.242	.216	.024	.013
	.304	**.342**	.267	.336	.391	**.309**	.414	.165
Open LLMs								
Llama 2-70B	.034	-.003	-.001	-.002	.001	-.002	.003	.008
	.371	.007	.099	.088	.157	.258	.386	.011
Llama 3-70B	.320	.160	.221	.247	.185	.155	.230	.196
	.522	.235	.329	**.389**	.272	.284	**.437**	.389
Llama 3.1-405B	.300	.119	.223	.217	.171	.157	.188	.099
	.084	.151	.274	.336	.185	.254	.136	.017
DeepSeek-R1	.326	.114	.178	.195	.159	.161	.287	.018
	.456	.121	.202	.233	.242	.234	.042	.096
Qwen 2.5-72B	.230	.126	.222	.216	.203	.176	.211	092
	.493	.212	**.282**	.331	.289	.261	.405	.155
Llama 3-8B	.199	.185	.263	.244	**.413**	.276	.054	.003
	.367	.128	.039	.049	.113	.096	.297	.004
Open-Source LLMs								
Prometheus-13B	.065	.035	.049	.031	.142	.058	.099	-.002
	.204	-.011	.011	-.009	.004	.000	.000	.009

Table 5: Prompt-level QWK scores on ASAP under zero-shot (shaded) and few-shot (unshaded) settings.

This is a significantly larger evaluation set than is typical in the literature where many studies sample 25-100 examples for benchmark comparison (Basyal and Sanghvi, 2023). Notably, (Odabaşı and Biricik, 2025) used 1,000 test instances and acknowledged this trend toward limited sample sizes. Our expanded test sample allows for more stable comparisons across model families and inference conditions. Each model was evaluated under zero-shot and few-shot configurations. In the few-shot setting, we included three examples randomly sampled from the CNN/DailyMail validation set, chosen to fit within the context window for all models and to represent varied content domains. This design is consistent with prior work (Odabaşı and Biricik, 2025) balancing context diversity and token constraints.

All generations were produced with a temperature of 0.3 and maximum output length of 100 tokens, consistent with prior evaluations in summarization (See et al., 2017). Summaries were evaluated using standard metrics: ROUGE-1, ROUGE-2, and ROUGE-L (Lin, 2004), which measure lexical overlap with human-written references, and METEOR (Banerjee and Lavie, 2005), which accounts for several linguistic phenomena such as synonymy, stemming, and word order.

B.2 Prompt Design

We designed task-oriented prompts that simulate and editorial summarization context.

Zero-shot Prompt

The zero-shot prompt included task instructions only:

```
As a news editor, your task
is to provide a concise, clear,
and informative summary of the
provided news article. The
summary should capture the main
events, important details, and
context presented in the original
article.

To accomplish this task:
- Carefully read and analyze the
news article provided.
- Identify the most important
events, key people, and essential
details.
- Write a summary in 2-3 concise
sentences that clearly convey the
primary content and significance
of the article.

Instructions:
- Ensure clarity, coherence, and
factual accuracy.
- Avoid redundancy or irrelevant
information.

Article Text: {ARTICLE TEXT}
Concise Summary (2-3 sentences):
{Model Output}
```

Few-shot Prompt

In the few-shot condition, the prompt included three article-summary examples in the same format as the target instance:

```
Article: {Example Article 1}
Summary: {Example Summary 1}

Article: {Example Article 2}
Summary: {Example Summary 2}

Article: {Example Article 3}
Summary: {Example Summary 3}

Now, summarize the following
article in 2-3 concise sentences:
Article Text: {Target ARTICLE
TEXT}
Summary: {Model Output}
```

(a) Prompts 9–24

Model	9	10	12	13	14	15	16	17	18	19	20	21	22	23	24
Closed LLMs															
GPT-3.5	-.011	.043	.039	.071	.182	.009	.006	.044	-.040	-.667	-.036	.046	.048	.044	.585
	.305	.531	.371	.444	**.830**	.310	.370	.436	.434	.600	.629	.224	.659	.538	-.105
GPT-4	.380	.484	.329	.225	.625	**.547**	**.795**	**.730**	**.702**	.600	**.909**	**.756**	**.713**	**.686**	.526
	.395	**.644**	.443	.571	.727	.340	.395	.474	.563	**.667**	.750	.504	.574	.592	-.378
GPT-4o	-.016	-.031	.052	.201	-.339	-.079	.010	.123	.015	**.667**	.343	.039	.008	-.015	-.065
	.437	.596	.498	.201	.727	.392	.403	.532	.596	.625	.498	.709	.695	-.246	-.233
Open LLMs															
Llama 2	.184	.229	.284	.225	.133	.164	.086	.197	.207	.600	.500	.259	.177	.155	.250
	.156	.309	.262	.296	.727	.159	.349	.247	.315	-.600	.313	.249	.427	.250	**.632**
Llama 3	.331	.217	.125	.079	.065	.376	.034	.249	.174	.600	.444	-.002	-.018	.161	.063
	.247	.508	.354	.370	.727	.248	.429	.353	.462	-.600	.444	.430	.512	.595	.063
Llama 3.1	.326	.257	.265	.119	.065	.388	.230	.255	.258	.600	.500	-.006	.331	.241	.375
	.238	.518	.295	.531	**.830**	.316	.357	.269	.377	-.600	.489	.382	.475	.535	-.125
DeepSeek-R1	-.017	.148	.032	.648	-.727	.085	.029	.142	-.102	.600	-.434	.006	.007	.038	-.667
	-.017	-.132	.009	.029	.133	.093	-.201	.099	-.081	-.600	-.063	-.339	.096	.074	-.522
Qwen 2.5	.013	-.134	.009	**.720**	.276	.089	-.171	.269	-.098	.600	.850	.018	.130	.009	-.500
	-.021	-.161	.078	.178	-.421	.047	-.151	.195	.157	.600	-.154	.032	.190	.094	-.500
Llama 3-8B	.009	.012	.019	-.014	.008	-.049	.059	.007	-.421	-.006	.057	.016	.108	-.387	.000
	-.026	-.065	-.023	.295	-.842	.063	-.161	.298	.093	-.813	.048	.036	-.089	-.500	-.291
Open-Source LLMs															
Prometheus-13B	-.019	.026	-.017	.052	-.065	-.079	.017	-.006	.117	-.600	.008	.001	-.030	.068	-.727
	.076	.215	-.024	-.129	.038	-.079	-.114	.072	.098	**.667**	-.275	-.066	.121	-.050	-.981

(b) Prompts 26–48

Model	26	27	29	30	39	40	41	42	43	44	45	46	47	48
Closed LLMs														
GPT-3.5	.276	.109	.100	.065	.023	.032	.033	.047	.028	-.111	.066	-.018	.056	.063
	.081	-.165	-.065	.401	.307	.299	.447	**.889**	-.111	.299	.238	.123	.525	.345
GPT-4	**.729**	.812	.427	.500	.402	.429	.415	.415	.645	-.111	.483	.469	.591	.477
	-.812	.293	-.539	.182	**.424**	.409	.477	.494	**.868**	-.111	.423	**.514**	.467	**.624**
GPT-4o	-.246	.100	.248	-.105	.009	-.003	-.018	-.005	-.029	-.111	.005	.014	.128	-.021
	.348	.071	.000	.496	.417	.352	.572	.693	-.111	**.539**	**.516**	.397	**.667**	.345
Open LLMs														
Llama 2	-.316	-.304	-.125	-.345	.165	.123	.028	.200	.289	-.111	.229	-.006	.211	.273
	-.304	-.125	-.345	.089	.225	.221	.165	-.111	.097	-.039	.023	.153	.222	.063
Llama 3	-.222	.375	.219	-.105	.176	.079	.023	.216	.105	-.111	.147	.007	.229	.223
	-.316	-.105	.376	.285	.347	**.459**	**.879**	.342	.397	.150	**.516**	.238	.345	.504
Llama 3.1	-.023	.783	.027	.108	.329	.147	-.022	.368	.309	.111	.358	.063	.252	.386
	-.571	**.836**	-.189	-.105	.359	.255	.245	.327	.771	.111	.441	.291	.268	.595
DeepSeek-R1	.375	-.625	-.179	**.830**	.068	.040	-.044	.036	-.029	.011	.059	-.380	.204	.181
	.096	.074	-.522	.812	-.002	.116	-.098	-.069	-.764	.111	.178	.197	.542	.078
Qwen 2.5	.111	.091	.500	-.909	-.047	.028	-.027	.005	-.297	-.111	.021	.254	-.169	.007
	.182	.329	**.636**	-.687	-.030	.016	-.064	-.009	-.294	-.333	.133	.016	.070	-.034
Llama 3-8B	-.387	.000	-.020	.045	-.029	-.034	-.278	-.111	.095	.023	-.062	.054	.003	.021
	-.035	-.727	-.015	.035	-.074	.021	-.413	-.111	.129	.002	-.208	.119	.014	.023
Open-Source LLMs														
Prometheus-13B	.021	-.223	-.25	.182	-.015	.006	.007	.010	.021	-.111	-.012	-.018	-.058	-.154
	-.334	-.514	.269	.267	.027	-.076	.029	-.116	-.307	.111	.112	-.093	.093	.029

Table 6: Prompt-level QWK scores on FCE under zero-shot (shaded) and few-shot (unshaded) settings.