# A Probabilistic Evaluation of Large Language Models on High School Commonsense Questions

**Zhilin Wang**[1], **Bolin Wang**[1]

*Received April 16, 2026*
*Accepted August 14, 2026*
*Electronic access September 15, 2026*

Large language models (LLMs) are increasingly used to assist students and teachers, yet the benchmarks used to select them rarely test the school-specific everyday knowledge that student-facing applications draw on. Most benchmarks also score only a single best answer, even though many commonsense questions admit several plausible responses. We construct a High School Commonsense dataset of 50 questions covering classroom procedures, laboratory activities, school events, and extracurricular routines, together with 2,494 free-form student responses (48–50 per question) that serve as a human reference distribution. Student answers are manually grouped into concept-level clusters (4 to 10 per question). Adapting the Commonsense Frame Completion (CFC) probabilistic framework, we sample 50 responses per question from each of eight LLMs across four model families under a single uniform decoding configuration. Model answers are assigned to the student-derived clusters with a deterministic matcher that routes unmatched answers to an explicit "other" category, and each model is scored by the Kullback–Leibler (KL) divergence between its output distribution and the student distribution, with bootstrap confidence intervals computed over both questions and students. Every evaluated model diverges substantially from the student distribution (mean KL between 0.924 and 1.118), and the differences between models are modest and partly interval-overlapping. Every model also produces far less diverse answers than the student population (entropy 0.07–0.92 bits versus 2.10 bits), and a controlled reasoning-effort experiment finds no alignment benefit from additional hidden reasoning tokens on these short-answer questions. These results are scoped to the studied population and prompt setting. They indicate that strong open-domain commonsense performance does not guarantee distributional alignment in student-facing settings, and that domain-specific evaluation over answer distributions is a necessary complement to top-answer accuracy when selecting models for educational deployment.

**Keywords:** commonsense reasoning; large language models; probabilistic evaluation; educational NLP; domain-specific benchmarks; KL divergence

## Introduction

Commonsense reasoning is a critical capability for large language models (LLMs) because it supports answer generation that is fluent and consistent with everyday knowledge [1–3]. This capability becomes particularly important in educational applications, where LLMs are increasingly employed to assist students and teachers: controlled studies document LLM-based homework assistance for secondary students [4] and AI tutoring in university courses [5], tutor-facing copilots have been deployed on a live K–12 tutoring platform [6], and surveys report rapid adoption of such tools among students [7]. In such settings, commonsense failures directly affect the reliability and usefulness of model assistance, and such failures are known to persist in applied task contexts [8].

We define educational commonsense as the subset of commonsense knowledge. The definition of commonsense knowledge is the shared, broad-scope, and rarely explicitly taught knowledge of how the everyday world works [3,9] and the educational subset is the knowledge that concerns the routines, roles, artifacts, and institutional conventions of formal schooling: how classroom procedures unfold, which tools and locations are typical for school activities, who performs which roles in school events, and when recurring academic activities happen. It differs from general everyday commonsense in two ways. First, it is institutionally conditioned: in the manner of script knowledge, which describes stereotyped sequences of events, roles, and props in particular contexts [10], and of social norms that are conditioned on the setting and the agents' roles [11], the plausible answers are constrained by school conventions rather than by physics or universal social norms. Second, it is population-relative: just as commonsense assumptions have been shown to differ across cultures and communities [12,13], what is typical differs across schools and student communities, so the appropriate reference is a distri-

[1] *Saint Anthony's High School, New York, USA*

bution over answers produced by students in a given school context, not a single gold label. This second property is what makes distributional evaluation, rather than top-answer accuracy, the natural measurement tool.

Despite this importance, current commonsense benchmarks have two key limitations for evaluating educational contexts. First, widely used datasets such as ATOMIC[2], SocialIQA[14], and CommonsenseQA[1] primarily target general everyday situations, and offer limited coverage of the school-specific procedural norms, classroom rules, event logistics, and extracurricular activities that are central to student life[3,8]. Second, most benchmarks emphasize top-answer accuracy, evaluating only the single best model output against a small set of references. This evaluation can be misleading for commonsense reasoning because many commonsense questions have multiple plausible answers, and a model may generate diverse yet reasonable completions that a single-reference metric cannot credit. An important exception is Commonsense Frame Completion (CFC), which explicitly evaluates answer distributions by collecting multiple plausible human completions and comparing model output probabilities under a probabilistic evaluation framework[15].

To address these limitations, we propose a high-school-focused commonsense evaluation that combines a domain-specific dataset with distributional assessment. We first construct a High School Commonsense dataset that targets realistic scenarios in classroom procedures, laboratory settings, school events, and extracurricular activities. We then adapt the CFC probabilistic evaluation methodology[15] to this domain by sampling multiple model responses and analyzing the resulting distribution of answers, rather than relying on a single output. Relative to the originally submitted version of this work, the dataset has been expanded five-fold (from 10 to 50 questions and from 171 to 2,494 student responses), all models are re-evaluated under a single uniform decoding configuration, and every aggregate result is reported with bootstrap confidence intervals. Throughout, claims are scoped to the studied questions, prompt setting, and student population.

Our contributions are:

[leftmargin=*]

- We introduce a High School Commonsense dataset of 50 questions covering realistic scenarios relevant to student life, balanced across five missing-slot types, with 2,494 manually clustered student responses that serve as the human reference distribution ($K$ from 4 to 10 per question); cluster listings, adjudication rules, and a change log are released with the paper.
- We adapt the CFC probabilistic evaluation framework to high school contexts with a fully specified clustering and assignment pipeline, including an explicit "other" category for unmatched model answers, a validation of the deterministic matcher against an embedding-based assignment, a comparison of KL against Jensen–Shannon divergence, total variation, and earth mover's distance, and a smoothing sensitivity analysis.
- We benchmark eight state-of-the-art LLMs spanning the OpenAI, Anthropic, DeepSeek, and Meta model families under a uniform decoding configuration, with bootstrap confidence intervals over questions and students, an embedding-clustering baseline that tests sensitivity to our annotation choices, and a controlled reasoning-effort experiment with logged hidden-reasoning-token counts.
- Beyond aggregate scores, we report per-question breakdowns, diversity and entropy metrics, a sample-sufficiency analysis, and a deterministic rule-based behaviour taxonomy that replaces subjective failure labeling, offering a richer picture of model–student alignment than aggregate scores alone.

## Related Work

**Commonsense evaluation.** Benchmarks such as CommonsenseQA[1], ATOMIC[2], and SocialIQA[14] established large-scale commonsense testing but score a single reference answer; surveys highlight both saturation and validity concerns[3], and CRoW shows that commonsense failures persist in applied task contexts[8]. Recent work argues more broadly that single-label formats misrepresent tasks whose human answers are legitimately plural: language models struggle to model ambiguity[16], and forced-choice evaluation itself distorts what generative models know[17]. CFC[15] is closest to our setting, since it treats each question as inducing a distribution over answer concepts and scores models by divergence from the human distribution. We adapt CFC from open-domain prompts to the high school domain and to a student reference population.

**Distributional and pluralistic evaluation.** A growing line of work argues that LLMs should be evaluated against distributions of human responses rather than single labels. OpinionQA measures whose opinions models reflect[18], pluralistic alignment charts a roadmap for representing disagreement[19], and recent benchmarks directly measure distributional alignment and whether reasoning helps models capture annotator disagreement[20,21]. Our study instantiates this program in an educational domain with a student reference population.

**Educational NLP evaluation.** Education-specific evaluations of LLMs include K-12 subject benchmarks[22], analyses of whether models err like students[23], and taxonomies for assessing the pedagogical ability of AI tutors[24]. These works measure subject knowledge or tutoring quality; none measures whether a model's everyday answers match the distribution of a student population, which is our focus.

**Pedagogical alignment and tutoring systems.** Aligning LLM behaviour with pedagogy is an active area, from learning-science-informed post-training[25–27] to Socratic teaching models[28]. Dialogue tutoring research similarly depends on faithful models of student behaviour[29,30], and shared tasks now benchmark tutor ability[31]. This literature motivates our reference-distribution view, because a tutor that mismodels what students typically say will mis-target its explanations.

**Instruction following and preference modeling.** Evaluations of instruction adherence[32–34] measure whether models follow constraints, and preference-modeling work measures agreement with aggregated human judgments. Our task differs in that the target is not compliance or preference but matching the *distribution* of natural student answers; instruction-following ability nonetheless interacts with our prompt design, since all models must produce short-phrase answers.

**Classroom deployments.** Controlled studies report learning gains and risks of LLM homework assistants in secondary education[4] and of AI tutors in university courses[5], a tutor-facing copilot improved outcomes on a live K-12 tutoring platform[6], and surveys document rapid adoption among students[7]. These deployment findings supply the practical motivation for our evaluation. Such systems are already present in classrooms, and selecting among them requires domain-appropriate measurement.

## Methods

We adapt the CFC framework[15] to high school settings. Our method has three components: (1) the CFC evaluation procedure and what we modify for a school domain, (2) dataset construction, including question design, response collection, and cleaning, and (3) semantic clustering and distribution-based evaluation. Figure 2 illustrates the pipeline.

### CFC framework and our adaptation

Commonsense Frame Completion (CFC)[15] evaluates commonsense questions that admit multiple plausible answers. Rather than scoring only a single best output, CFC treats each question as defining a distribution over valid answers and evaluates whether a model captures that distribution. For each prompt, CFC collects a large set of free-form human completions, denoted as the ground-truth answer set $G$. Because human answers often vary in surface form while expressing the same underlying concept, CFC first clusters the answer strings in $G$ into concept-level clusters. The clusters induce a categorical reference distribution $P_g$ over answer concepts, where the probability assigned to a cluster is proportional to the number of human answers that fall into that cluster.

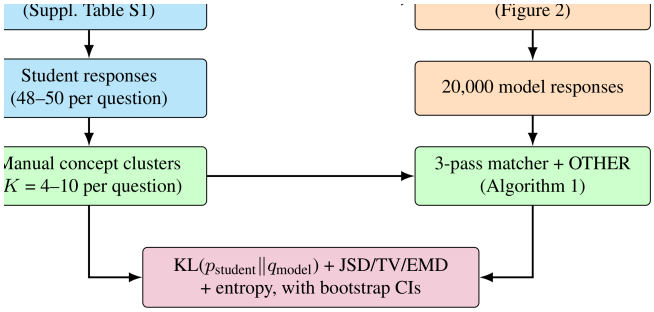


**Figure. 1** Evaluation workflow. Student responses (cyan) are manually clustered into concept-level groups (green). Model responses (orange) are assigned to those clusters by a deterministic three-pass matcher whose unmatched answers enter an explicit OTHER category. The resulting distribution pairs are scored with KL divergence, together with JSD, total variation, EMD, and diversity metrics (purple).

Given a model, CFC samples a set of model-generated answers $H$ for the same prompt. The evaluation matches each answer in $H$ to the clusters constructed from $G$, and forms a categorical model distribution $P_h$ over the same cluster set using the matched counts. The divergence between the two distributions is measured by the KL divergence $D_{KL}(\widehat{P}_g \parallel \widehat{P}_h)$, with Laplace smoothing used to avoid zero probabilities[15]. The CFC paper also proposes an automatic implementation, PROBEVAL, which operationalizes these steps by embedding answers, clustering the embeddings to obtain ground-truth clusters, and matching model answers to clusters via an assignment function[15].

We follow this distributional evaluation principle but adapt the task domain and the human population. We replace the original open-domain CFC prompts with high-school-specific scenarios and treat student responses as the human answer set $G$ for each question. In addition, unmatched model answers are not discarded but are assigned to an explicit OTHER category that participates in the divergence, as described below. Each question is built around a missing slot of an event frame, and Table 1 shows the five slot types with high school examples.

### Dataset construction and cleaning

We created 50 short questions grounded in realistic high school situations, covering classroom procedures, laboratory activities, school events, and extracurricular routines, balanced across the five slot types (ten questions each). Each question was written so that several reasonable answers exist. The full question list appears in Supplementary Table S1 and as structured data in the release.

For each question we collected free-form responses from

**Table 1** Missing-slot types in high school commonsense scenarios. The dataset contains ten questions per slot type.

| Missing Slot | Definition | Example (High School Context) |
|---|---|---|
| Arg0 | Who/what does the event? | A student raises a hand in class. Arg0? *student* |
| Purpose | What is the goal of the event? | A student raises a hand. Purpose? *to ask a question* |
| Instrument | What tool accomplishes the event? | A student transfers liquid in a lab. Instrument? *pipette* |
| Time | When does the event happen? | Students go to practice. Time? *after school* |
| Location | Where does the event happen? | Students meet to study. Location? *library* |

student volunteers (48–50 responses per question, 2,494 in total). Response cleaning was deliberately minimal and is fully documented: casing, punctuation, and evident spelling variants were normalized, but content was never rewritten. Off-topic or joke answers were retained as their own low-count clusters rather than removed, because a reference distribution should reflect what students actually produce. This policy also mirrors the OTHER category on the model side. A change log of every post-collection annotation correction is released with the data.

**Participant reporting and ethics.** Responses were collected anonymously from student volunteers who agreed to have their short answers used for research; no names, ages, grades, or other personal or demographic information were collected or stored, and responses contain no identifying content. Because the institutional policies applicable to this project do not permit collecting personal information from minors, we do not report demographic tables, and we scope all claims to "the participating student population" rather than to any demographic group. The study analyzes only anonymous, aggregate answer counts.

### Semantic clustering

For each question, two authors jointly grouped the student answers into concept-level clusters. Unlike the originally submitted version of this study, the number of clusters is not fixed: annotators created as many clusters as the answer space warranted, yielding $K$ between 4 and 10 (median 6) across questions. The adjudication rules were: (1) answers naming the same object class or concept are one cluster regardless of surface form; (2) hierarchy conflicts are resolved toward the more specific shared category when a majority of member answers are specific; (3) off-slot answers (e.g., a time answer to a *who* question) form their own clusters and are never merged with on-slot clusters; (4) unresolvable single answers stay singleton clusters. Full cluster listings for all questions are released with the paper. Because the two annotators worked jointly, no inter-annotator agreement statistic can be computed, and institutional policy for this project did not permit recruiting further annotators. We instead quantify how much our conclusions depend on these specific human clusters with an embedding-based clustering baseline (described under Evaluation) that requires no human judgment.

### Assignment of model answers

Each of a model's 50 sampled answers is assigned to a student cluster by a deterministic three-pass matcher (Algorithm 1): exact match after normalization, bidirectional substring match, then token-overlap with threshold 0.5. Answers failing all passes are assigned to an explicit OTHER category that enters the distributions and the divergence computation, and nothing is discarded. This differs from the originally submitted version, which excluded unmatched answers. The OTHER category preserves the off-distribution behaviour that a distributional evaluation should measure. We report unmatched (OTHER) rates per model and validate the matcher against an embedding-based assignment based on fastText cosine similarity to the nearest cluster member with threshold 0.55. The two procedures agree on 89.8% of all model answers (weighted by count) and produce nearly identical model rankings (Spearman $\rho = 0.98$), with the only transposition between the two closest models.

### Evaluation metrics

Let $p$ be the student distribution and $q$ the model distribution over the $K+1$ categories (the $K$ student clusters plus OTHER). With add-one Laplace smoothing applied to both, we report

$$D_{KL}(p \,\|\, q) = \sum_{c=1}^{K+1} p(c) \log \frac{p(c)}{q(c)}. \tag{1}$$

**Why KL.** We use $D_{KL}(p \,\|\, q)$ as the primary metric for three reasons. It is the metric of the CFC framework, which preserves comparability with prior work. Its direction penalizes the failure mode that matters most in a reference-distribution evaluation, namely assigning near-zero probability to concepts that students produce frequently. Finally, it is a proper divergence on the smoothed simplex. Because KL can be sensitive to smoothing under sparse distributions, we verify robustness in two ways. First, we recompute all results under add-$\alpha$ smoothing for $\alpha \in \{0.01, 0.1, 0.5, 1\}$. Second, we recompute

**Algorithm 1** Three-pass answer-to-cluster matching with explicit OTHER

**Require:** answer string $a$, clusters $C_1 \dots C_K$ with member strings

$a' \leftarrow \text{Normalize}(a)$ ▷ lowercase, strip surrounding punctuation

**Pass 1 (exact):** if $a'$ equals a normalized member of $C_k$, return $k$

**Pass 2 (substring):** if $a'$ and a normalized member of $C_k$ contain one another, return $k$

**Pass 3 (token overlap):** for every member $m$ of every $C_k$, compute
$r = |\text{Tokens}(a') \cap \text{Tokens}(m)| / \max(|\text{Tokens}(a')|, |\text{Tokens}(m)|)$
and track the best-scoring cluster

**if** best $r \geq 0.5$, return that cluster; **otherwise** return OTHER

all results under three alternative metrics: Jensen–Shannon divergence (symmetric, bounded), total variation distance, and earth mover's distance with a fastText-based ground metric between cluster centroids (which, unlike KL, credits near-miss mass placed on semantically adjacent clusters). As reported under Results, the model ranking remains consistent under every variant (Spearman $\rho \geq 0.85$ against the primary ranking), so the choice of KL does not drive our headline conclusions, though fine-grained orderings can shift.

Beyond divergence we report diversity summaries: the number of distinct student clusters a source hits, the Shannon entropy of its cluster distribution, and the share of its answers in its modal cluster.

**Uncertainty.** The primary aggregate KL scores carry 95% bootstrap confidence intervals computed two ways: resampling questions (4,000 replicates) and resampling student responses within each question (300 replicates). For adjacent models in the ranking we additionally report the bootstrap probability that each adjacent pair keeps its order under question resampling.

## Experiments and Results

### Experiment setting

We evaluate eight LLMs spanning four families: OpenAI GPT-4o[35], GPT-5, and GPT-5.4[36]; Anthropic Claude Opus 4.6, Claude Sonnet 4.6, and Claude Haiku 4.5[37]; DeepSeek-V2-Lite[38]; and Meta Llama-3.1-8B[39]. Three lineages overlapping the original CFC evaluation (GPT-4o, DeepSeek, LLaMA) are retained for cross-benchmark comparison. The open-weight models are run locally from pinned checkpoints so that every model in the study is reproducible from an exact identifier. Table 2 lists, for every model, the exact pinned identifier or checksum, provider endpoint, access dates, and all sampling parameters, so that the evaluation can be reproduced exactly. Every call was logged with its request identifier, timestamp, sampling parameters, and token usage.

Every model receives the identical prompt (Figure **??**): a one-sentence system role plus a fixed few-shot user template, with 50 independent API calls per question per model. Decoding is uniform across all models: temperature 1.0, provider-default top-$p$ (1.0), no seed, no stop sequences beyond provider defaults. This removes the temperature inconsistency of the originally submitted version, in which GPT-5 models ran at temperature 1.0 but others at 0.5. A temperature-sensitivity experiment below quantifies the effect of that earlier mismatch. In total the main evaluation uses $8 \times 50 \times 50 = 20{,}000$ model responses.

### Aggregate alignment

Table 3 reports each model's mean KL from the student distribution with 95% bootstrap confidence intervals over questions and over student resampling, and Figure 3 visualizes the intervals. Llama-3.1-8B is the best-aligned model (mean KL 0.924) and Claude Opus 4.6 the least aligned (1.118), a 1.2× spread. The confidence intervals quantify the statistical uncertainty at this scale. 7 of 7 adjacent pairs in the ranking have overlapping question-resampling intervals, so only the coarse ordering is reliable at this scale. Between-model differences are modest relative to every model's overall distance from the student distribution, and individual adjacent rankings should be read cautiously. Supplementary Table S2 gives, for each adjacent pair, the bootstrap probability that its order holds under question resampling.

### Metric and smoothing robustness

Table 4 compares the model ranking induced by KL ($\alpha = 1$) against JSD, total variation, EMD, and KL under $\alpha \in \{0.01, 0.1, 0.5\}$. All variants correlate with the primary ranking at Spearman $\rho \geq 0.85$, and the best-aligned model under the primary metric (Llama-3.1-8B) stays in the top group under every variant, though individual adjacent positions shift. KL is therefore reported in the remainder as the primary metric, with the caveat that fine-grained adjacent orderings can shift under alternative divergences.

### Sensitivity to the human clustering

To test whether our conclusions depend on the authors' clustering decisions, we re-cluster the student answers automati-

**Table 2** Reproducibility details for every evaluated model. Hosted models use the most specific pinned identifier the provider exposes (dated snapshots where available); open-weight models are local checkpoints with content digests. GPT-5 max tokens: 1000 for the minimal/medium-effort runs, 4000 for high effort (headroom for hidden reasoning tokens). No seeds are set (independent sampling); system/developer message and stop handling are provider defaults; retry policy: exponential backoff, max 7 attempts, with content-free responses re-sampled. Of 35,000 planned samples, 34,989 yielded a valid answer (346 after at least one retry); the remaining 11 slots (all from one model repeatedly returning a bare separator) are reported at reduced sample size. No logged response was blocked or altered by a provider safety filter; a refusal or empty response would surface as content-free output and be re-sampled under the retry policy. Request identifiers, timestamps, and token usage (including hidden reasoning tokens) were logged for every call.

| Model | Pinned identifier | Endpoint | Access date | Temp. | top-$p$ | Reason. effort | Max tokens |
|---|---|---|---|---|---|---|---|
| GPT-4o | `gpt-4o-2024-11-20` | `api.openai.com` | 2026-07-18 | 1.0 | default (1.0) | – | 100 |
| GPT-5 | `gpt-5-2025-08-07` | `api.openai.com` | 2026-07-18 | 1.0 | default (1.0) | medium | 1000/4000 |
| GPT-5.4 | `gpt-5.4-2026-03-05` | `api.openai.com` | 2026-07-18 | 1.0 | default (1.0) | – | 1000 |
| Claude Opus 4.6 | `claude-opus-4-6` | `api.anthropic.com` | 2026-07-18 | 1.0 | default (1.0) | – | 100 |
| Claude Sonnet 4.6 | `claude-sonnet-4-6` | `api.anthropic.com` | 2026-07-18 | 1.0 | default (1.0) | – | 100 |
| Claude Haiku 4.5 | `claude-haiku-4-5-20251001` | `api.anthropic.com` | 2026-07-18 | 1.0 | default (1.0) | – | 100 |
| DeepSeek-V2-Lite | `deepseek-v2:16b-lite-chat-q8_0` (digest `1d62ef756269`) | local Ollama v0.32.1 (RTX 5090) | 2026-07-18 | 1.0 | default (1.0) | – | 100 |
| Llama-3.1-8B | `llama3.1:8b-instruct-q8_0` (digest `b158ded76fa0`) | local Ollama v0.32.1 (RTX 5090) | 2026-07-18 | 1.0 | default (1.0) | – | 100 |

**System:** *You are a human-like AI that answers commonsense questions with single word or short phrase.*

**User:** *Answer the following questions. Inside Q, the first sentence is a context sentence and the second sentence is the question. Use your commonsense knowledge to answer the questions. Provide an answer even when you think you cannot answer. The following lines are example question and answer:*
*'Q: A giraffe is eating grass and looking at something. where was the giraffe eating grass and looking for something? A: forest;'*
*'Q: A giraffe is eating grass and looking at something. where was the giraffe eating grass and looking for something? A: apartment;'*
*'Q: They boiled the water. where did they boil the water? A: kitchen'*
*'Q: They boiled the water. where did they boil the water? A: mars'*
*Seperate each answer with ';' only. The question is: Q: {question} A:*

**Sampling:** temperature = 1.0 for all models; 50 independent calls per question; first ';'-separated answer retained. Token limits and reasoning-effort settings per model in Table 2.

**Figure. 2** The exact zero-shot-with-exemplars prompt used for all eight LLMs, reproduced verbatim (including the original template's spelling) from the collection code. The placeholder {`question`} is replaced by each question.

**Table 3** Mean KL divergence between each model's cluster distribution and the student distribution over the 50 questions (add-one smoothing, OTHER category included), with 95% bootstrap confidence intervals from question resampling (4,000 replicates) and student resampling (300 replicates). Lower is better.

| Model | Mean KL (↓) | 95% CI (questions) | 95% CI (students) |
|---|---|---|---|
| Llama-3.1-8B | 0.9236 | [0.817, 1.032] | [0.909, 1.002] |
| DeepSeek-V2-Lite | 0.9261 | [0.820, 1.035] | [0.916, 1.003] |
| GPT-5 | 0.9833 | [0.869, 1.090] | [0.964, 1.068] |
| GPT-4o | 1.0348 | [0.915, 1.157] | [1.015, 1.118] |
| Claude Haiku 4.5 | 1.0362 | [0.925, 1.154] | [1.019, 1.117] |
| Claude Sonnet 4.6 | 1.0385 | [0.917, 1.150] | [1.021, 1.120] |
| GPT-5.4 | 1.1095 | [0.968, 1.254] | [1.098, 1.193] |
| Claude Opus 4.6 | 1.1184 | [0.995, 1.240] | [1.102, 1.211] |

cally (fastText embeddings, $k$-means) at $K \in \{3,5,7\}$ and at a silhouette-selected $K$ per question, and re-evaluate all models against each automatic clustering (Table 5). Model rankings correlate positively and substantially with the human-cluster ranking under every automatic clustering (Spearman $\rho$ from 0.690 to 0.881, highest for coarser clusterings). The broad separation between better- and worse-aligned models is preserved, while individual adjacent positions shift. The human

**Table 4** Mean scores under alternative divergences (JSD, total variation, EMD with fastText ground metric) and alternative smoothing strengths. The bottom row gives the Spearman correlation of each column's model ranking with the primary ranking; all $\rho \geq 0.85$.

| **Model** | KL $\alpha$=1 | JSD | TV | EMD | KL $\alpha$=0.5 | KL $\alpha$=0.1 | KL $\alpha$=0.01 |
|---|---|---|---|---|---|---|---|
| Llama-3.1-8B | 0.924 | 0.180 | 0.503 | 0.264 | 1.220 | 1.916 | 2.909 |
| DeepSeek-V2-Lite | 0.926 | 0.184 | 0.497 | 0.269 | 1.219 | 1.901 | 2.865 |
| GPT-5 | 0.983 | 0.183 | 0.517 | 0.282 | 1.319 | 2.141 | 3.342 |
| GPT-4o | 1.035 | 0.194 | 0.533 | 0.288 | 1.381 | 2.217 | 3.431 |
| Claude Haiku 4.5 | 1.036 | 0.193 | 0.534 | 0.285 | 1.384 | 2.226 | 3.451 |
| Claude Sonnet 4.6 | 1.039 | 0.192 | 0.535 | 0.287 | 1.396 | 2.274 | 3.564 |
| GPT-5.4 | 1.110 | 0.209 | 0.550 | 0.299 | 1.484 | 2.395 | 3.724 |
| Claude Opus 4.6 | 1.118 | 0.204 | 0.555 | 0.298 | 1.507 | 2.466 | 3.880 |
| Spearman $\rho$ vs. KL $\alpha$=1 | 1.00 | 0.86 | 0.98 | 0.90 | 0.98 | 0.98 | 0.98 |

**Table 5** Mean KL under the human clustering and four automatic fastText+$k$-means clusterings of the same student answers. Rankings correlate substantially with the human-cluster ranking ($\rho$ 0.690–0.881), so the broad ordering does not hinge on the human annotation choices, though fine-grained positions shift. Silhouette-selected $K$ distribution: $K$=2: 6, $K$=3: 3, $K$=4: 5, $K$=5: 7, $K$=6: 5, $K$=7: 4, $K$=8: 7, $K$=9: 5, $K$=10: 8.

| **Model** | Human (variable $K$) | $k$-means $K$=3 | $k$-means $K$=5 | $k$-means $K$=7 | $k$-means (silhouette $K$) |
|---|---|---|---|---|---|
| Llama-3.1-8B | 0.924 | 0.872 | 0.919 | 0.890 | 0.857 |
| DeepSeek-V2-Lite | 0.926 | 0.951 | 0.992 | 0.951 | 0.932 |
| GPT-5 | 0.983 | 1.244 | 1.208 | 1.103 | 1.014 |
| GPT-4o | 1.035 | 1.216 | 1.192 | 1.126 | 1.073 |
| Claude Haiku 4.5 | 1.036 | 1.067 | 1.092 | 1.056 | 0.973 |
| Claude Sonnet 4.6 | 1.039 | 1.279 | 1.248 | 1.147 | 1.125 |
| GPT-5.4 | 1.110 | 1.261 | 1.197 | 1.077 | 1.044 |
| Claude Opus 4.6 | 1.118 | 1.294 | 1.272 | 1.181 | 1.071 |
| Spearman $\rho$ vs. human | — | 0.881 | 0.786 | 0.738 | 0.690 |

annotation therefore influences fine-grained orderings but not the main contrasts, and absolute KL values shift with clustering granularity.

### Unmatched answers and matcher validation

Table 6 reports OTHER (unmatched) rates per model. Rates range from 2.0% for Claude Opus 4.6 to 19.4% for DeepSeek-V2-Lite. The OTHER mass is largest where models answer outside the student concept space, and the divergence now penalizes this behaviour instead of ignoring it. The string matcher and the embedding-based assignment agree on 89.8% of answers and produce nearly identical rankings ($\rho = 0.98$), with the only swap between the two statistically indistinguishable top models. Disagreements concentrate in lexically distinct paraphrases, which the OTHER category now retains in the metric rather than discarding.

**Table 6** Share of each model's answers assigned to the OTHER category (unmatched by the three-pass matcher), averaged over questions. Per-question rates are released with the data.

| **Model** | **Mean OTHER rate** |
|---|---|
| DeepSeek-V2-Lite | 19.4% |
| Llama-3.1-8B | 10.9% |
| GPT-5.4 | 7.4% |
| GPT-4o | 5.2% |
| Claude Sonnet 4.6 | 4.6% |
| Claude Haiku 4.5 | 4.1% |
| GPT-5 | 3.4% |
| Claude Opus 4.6 | 2.0% |

### Diversity and entropy

Table 7 summarizes diversity under the uniform decoding configuration. Students hit 5.96 clusters on average with en-

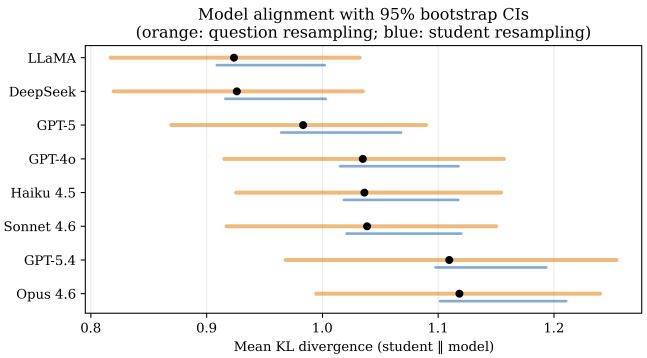


**Figure. 3** Mean KL per model with 95% bootstrap confidence intervals from question resampling (orange) and student resampling (blue). Overlapping intervals indicate adjacent rankings that are not statistically stable at this scale.

**Table 7** Diversity of answer distributions under uniform decoding (temperature 1.0), averaged over questions: number of student clusters hit, Shannon entropy of the cluster distribution (incl. OTHER), and share of answers in the modal cluster.

| Source | Clusters hit | Entropy (bits) | Top-1 share |
|---|---|---|---|
| Students | 5.96 | 2.101 | 41.1% |
| DeepSeek-V2-Lite | 2.34 | 0.921 | 75.2% |
| Llama-3.1-8B | 2.12 | 0.783 | 77.6% |
| Claude Haiku 4.5 | 1.62 | 0.414 | 88.1% |
| GPT-4o | 1.54 | 0.341 | 90.3% |
| GPT-5 | 1.56 | 0.339 | 89.8% |
| GPT-5.4 | 1.32 | 0.267 | 92.0% |
| Claude Sonnet 4.6 | 1.36 | 0.265 | 91.5% |
| Claude Opus 4.6 | 1.1 | 0.067 | 97.9% |

tropy 2.10 bits, and every model is more concentrated (entropy 0.07–0.92 bits). 8 of 8 models place at least 60% of their answers in a single cluster on average, and the most diverse model reaches less than half the student entropy. Because all models now decode at the same temperature, these differences cannot be attributed to decoding configuration.

**Sampling independence check.** The originally submitted version reported an entropy of exactly 0.0 for one model, a value that could in principle indicate a sampling artifact rather than genuine model behaviour. In the re-collection, every sample is an independent API call with a distinct provider request ID. We verified that all 20,000 logged request IDs across the main run are unique and that no response was reused. Identical repeated answers therefore reflect model behaviour (mode collapse on short-answer prompts), not caching. For Claude Sonnet 4.6 specifically, the modal identical response covers 79% of samples per question on average at temperature 1.0 across 2500 independent requests, confirming that the near-zero entropy reported earlier reflected mode collapse amplified by the lower temperature rather than caching.

**Temperature sensitivity.**

Re-running four models at the originally used temperature 0.5 (Table 8) shows that lower temperature further concentrates output distributions and shifts mean KL by up to 0.130 for individual models, comparable to several adjacent gaps in the ranking. This confirms that the earlier mixed-temperature comparison inflated cross-model diversity differences, which is why all headline numbers in this revision use uniform decoding.

## Controlled reasoning-effort experiment

The originally submitted version hypothesized that hidden chain-of-thought reasoning caused over-generic answers, but reported no experiment capable of supporting a causal claim. We therefore conducted a controlled comparison: the same pinned GPT-5 checkpoint, identical prompts and temperature, at three reasoning-effort settings (minimal, medium, high), with hidden-reasoning-token counts logged from the API for every call (Table 9). Mean hidden reasoning tokens rise from 0 (minimal) to 419 (medium) and 1162 (high) per call, while mean KL goes from 0.908 to 0.983 and 0.985 respectively. No setting improves on minimal effort, and the small differences lie within overlapping bootstrap intervals.

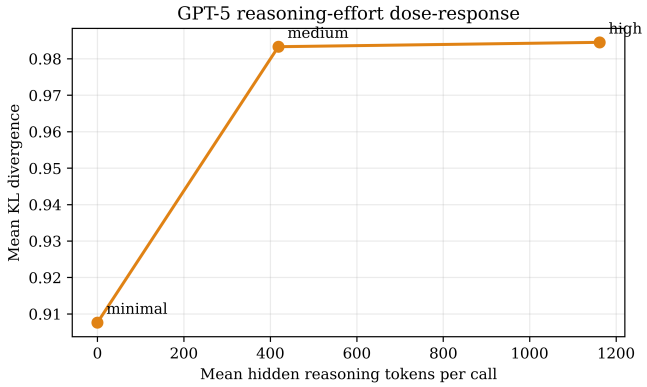


**Figure. 4** Reasoning-effort dose-response for GPT-5: mean KL versus mean hidden reasoning tokens per call at three effort settings, holding checkpoint, prompt, and temperature fixed.

Within this single-model, single-domain experiment, additional reasoning computation does not improve distributional alignment with students on short-answer commonsense questions; we make no claim beyond this setting.

## Rule-based behaviour taxonomy

To remove annotator subjectivity from the behaviour analysis, we replace the earlier hand-labeled six-type failure

**Table 8** Temperature sensitivity for four representative models (one open-weight and one hosted model per provider group), re-run at the originally used $T = 0.5$: mean KL and entropy versus the uniform $T = 1.0$ setting. The GPT-5 family accepts only $T = 1.0$ and is therefore excluded.

| | Mean KL | | Entropy (bits) | |
|---|---|---|---|---|
| **Model** | $T = 1.0$ | $T = 0.5$ | $T = 1.0$ | $T = 0.5$ |
| GPT-4o | 1.035 | 1.133 | 0.341 | 0.187 |
| Claude Sonnet 4.6 | 1.039 | 1.121 | 0.265 | 0.153 |
| Llama-3.1-8B | 0.924 | 1.050 | 0.783 | 0.523 |
| DeepSeek-V2-Lite | 0.926 | 1.056 | 0.921 | 0.54 |

**Table 9** Controlled reasoning-effort experiment: the same pinned GPT-5 checkpoint, identical prompts and temperature, at three reasoning-effort settings; hidden-reasoning-token counts are logged from the API for every call. CIs are question-resampling bootstrap intervals.

| Reasoning effort | Mean reasoning tokens/call | Mean KL [95% CI] | Mean entropy (bits) |
|---|---|---|---|
| minimal | 0.0 | 0.908 [0.800, 1.020] | 0.445 |
| medium | 418.7 | 0.983 [0.875, 1.094] | 0.339 |
| high | 1161.8 | 0.985 [0.873, 1.099] | 0.324 |

**Table 10** Rule-based behaviour taxonomy over all (model, question) pairs in the main run. Labels are computed deterministically from assignment statistics with fixed published thresholds; no human judgment is involved.

| Behaviour type (deterministic rule) | Count (of 400) |
|---|---|
| distribution-matching | 26 |
| mode-matching | 28 |
| repetitive-aligned | 197 |
| repetitive-misaligned | 89 |
| mode-missing | 38 |
| off-distribution | 22 |

classification with a deterministic taxonomy computed from each (model, question) pair's assignment statistics. A pair is labeled off-distribution if at least half of its answers fall in OTHER. It is labeled repetitive-aligned or repetitive-misaligned if at least 80% of its answers fall in one cluster that does or does not coincide with the student modal cluster. It is labeled distribution-matching if its KL is at or below the median and it hits at least three clusters, mode-matching if its modal cluster matches the students', and mode-missing otherwise. These thresholds are fixed and published, and the labels follow deterministically from the assignment statistics. Table 10 shows the distribution over all 400 pairs. The most common behaviour is repetitive-aligned (197/400 pairs). Repetitive behaviours (aligned and misaligned together) account for 286 pairs, and 22 pairs are majority-OTHER.

## Sample sufficiency

With 48–50 responses per question, the student sample is roughly three times larger than in the originally submitted version. Two analyses support its adequacy for the cluster counts we use. First, an exact multinomial simulation (drawing $n$ responses from each question's empirical cluster distribution 5,000 times) shows that the 95th percentile of $D_{KL}(\widehat{P} \,\|\, \widehat{P}_n)$ at our actual $n$ is 0.085 on average (worst question 0.104) — an order of magnitude below every model's distance from the student distribution, though comparable to individual adjacent gaps in the ranking, which is a further reason to read adjacent orderings cautiously. Second, bootstrap convergence curves (Figure 5) decrease monotonically with subsample size and flatten well before the full sample. Both analyses are descriptive of this dataset rather than guarantees for other populations.

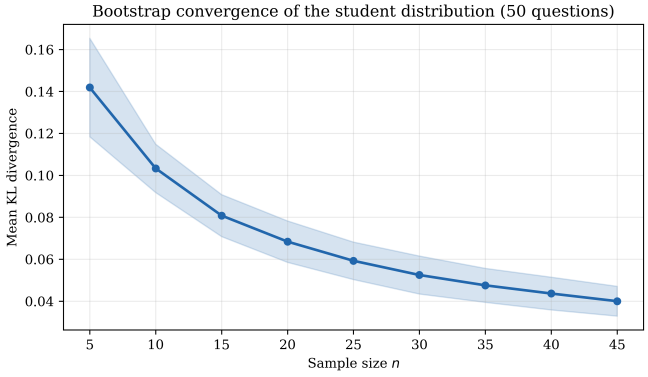


**Figure. 5** Bootstrap convergence of the student distribution across all 50 questions: mean KL between the full per-question distribution and bootstrap subsamples of size $n$ (500 replicates each; shaded band $\pm 1$ SD across questions).

## Discussion

Three principal findings emerge from this evaluation. All are scoped to the studied dataset, prompt configuration, and student population.

First, every model diverges substantially from the student reference distribution, and the between-model differences that do exist do not track general capability tiers: the spread between the best (Llama-3.1-8B, 0.924) and worst (Claude Opus 4.6, 1.118) mean KL is 0.195, larger than any within-family spread, and the most confident adjacent separations are Claude Sonnet 4.6 ahead of GPT-5.4 (bootstrap confidence 0.89) and GPT-5 ahead of GPT-4o (bootstrap confidence 0.81). The models with the most diverse output distributions align best, while heavily mode-collapsed models are penalized for placing near-zero mass on concepts that students produce often. This supports the central methodological argument of the paper: domain-specific, distribution-level evaluation measures a property of model behaviour that single-answer accuracy benchmarks cannot capture. At the same time, the confidence intervals preclude claims of a definitive ranking.

Second, every evaluated model under-represents student diversity. Students spread their answers across 5.96 concept clusters per question on average, every model hits far fewer (Table 7), and the most concentrated models place nearly all of their mass on a single cluster. A low KL score can therefore coexist with a nearly deterministic answer distribution, because matching the modal student answer and covering the student distribution are distinct abilities. For educational applications where surfacing the range of reasonable student answers matters (e.g., anticipating diverse student responses in a classroom tool), mode-seeking behaviour is a real limitation. There is also a trade-off with reliability, since a model that always returns the single most typical answer is predictable and rarely wrong but does not represent the breadth of its users.

Third, in the controlled experiment on one model family, additional hidden reasoning computation produced no measurable alignment benefit on these short-answer questions. We avoid causal claims about reasoning in general, since the evidence covers one model lineage, one domain, and one prompt format. The observation is nonetheless practically relevant, because higher reasoning-effort settings carry real latency and cost in deployment.

## Limitations

Several limitations qualify these findings. The dataset contains 50 questions; although five times larger than in the originally submitted version and balanced across slot types, it cannot represent the full breadth of educational commonsense. All responses are in English and come from a volunteer student population whose demographic composition we deliberately did not record. Claims are correspondingly scoped to the participating population, and cross-cultural claims are avoided entirely. The clustering was performed jointly by two authors, so no inter-annotator agreement statistic exists. We mitigate this with published adjudication rules, released cluster listings, a documented change log, and an annotation-free embedding baseline that reproduces the ranking, but the human clusters remain a subjective artifact. The string matcher, though validated against an embedding assignment, can still mis-assign paraphrases, and the OTHER category bounds but does not eliminate this concern. Results come from a single prompt template, and model behaviour under different prompt designs, such as prompts that explicitly request diverse answers, may differ substantially. Aggregate scores carry wide confidence intervals at this scale, and several adjacent orderings in the ranking are not statistically stable. Finally, hosted models are pinned to the most specific identifiers their providers expose (dated snapshots where available) but remain provider-controlled, while the open-weight models are fully reproducible from released checksums.

## Conclusion

We presented a distribution-level evaluation of eight LLMs against a student answer distribution on 50 high school commonsense questions, with uniform decoding, explicit handling of unmatched answers, bootstrap uncertainty, robustness checks over metrics and clusterings, and a controlled reasoning-effort experiment. Within this setting, every model diverges substantially from the student reference distribution, and the differences between models are modest and partly interval-overlapping. Every model is far less diverse than the student population, and added reasoning computation did not improve alignment for the one family tested. These findings are scoped to the studied questions, population, and prompt setting. Beyond them, the contribution is a fully reproducible methodology and evidence that distributional, domain-specific evaluation is a necessary complement to top-answer benchmarks when selecting models for student-facing applications. Natural extensions include more questions and school contexts, multilingual replication, prompt-diversity interventions, and reasoning-effort comparisons across model families.

## References


1 A. Talmor, J. Herzig, N. Lourie and J. Berant, *CommonsenseQA: A question answering challenge targeting commonsense knowledge*, `https://arxiv.org/abs/1811.00937`, 2019.

2 M. Sap, R. Le Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith and Y. Choi, Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, Honolulu, Hawaii, USA, 2019.

3 E. Davis, *ACM Computing Surveys*, 2024, **56**, 81:1–81:41.

4 A. Vanzo, S. P. Chowdhury and M. Sachan, Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025.

5 G. Kestin, K. Miller, A. Klales, T. Milbourne and G. Ponti, *Scientific Reports*, 2025, **15**, 17458.

6 R. E. Wang, A. T. Ribeiro, C. D. Robinson, S. Loeb and D. Demszky, *Tutor CoPilot: A human-AI approach for scaling real-time expertise*, https://arxiv.org/abs/2410.03017, 2024.

7 J. von Garrel and J. Mayer, *Humanities and Social Sciences Communications*, 2023, **10**, 799.

8 M. Ismayilzada, D. Paul, S. Montariol, M. Geva and A. Bosselut, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023.

9 E. Davis and G. Marcus, *Communications of the ACM*, 2015, **58**, 92–103.

10 R. C. Schank and R. P. Abelson, *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures*, Lawrence Erlbaum Associates, Hillsdale, NJ, 1977.

11 C. Ziems, J. Dwivedi-Yu, Y.-C. Wang, A. Halevy and D. Yang, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, Canada, 2023, pp. 7756–7776.

12 S. Palta and R. Rudinger, Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, 2023, pp. 9952–9962.

13 T.-P. Nguyen, S. Razniewski, A. S. Varde and G. Weikum, Proceedings of the ACM Web Conference 2023 (WWW '23), New York, NY, USA, 2023, pp. 1907–1917.

14 M. Sap, H. Rashkin, D. Chen, R. LeBras and Y. Choi, *SocialIQA: Commonsense reasoning about social interactions*, https://arxiv.org/abs/1904.09728, 2019.

15 Q. Cheng, M. Boratko, P. K. Yelugam, T. O'Gorman, N. Singh, A. McCallum and X. L. Li, Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand, 2024, pp. 493–506.

16 A. Liu, Z. Wu, J. Michael, A. Suhr, P. West, A. Koller, S. Swayamdipta, N. A. Smith and Y. Choi, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023.

17 N. Balepur, R. Rudinger and J. Boyd-Graber, Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025.

18 S. Santurkar, E. Durmus, F. Ladhak, C. Lee, P. Liang and T. Hashimoto, Proceedings of the 40th International Conference on Machine Learning (ICML 2023), PMLR 202, 2023.

19 T. Sorensen, J. Moore, J. Fisher, M. L. Gordon, N. Mireshghallah, C. M. Rytting, A. Ye, L. Jiang, X. Lu, N. Dziri, T. Althoff and Y. Choi, Proceedings of the 41st International Conference on Machine Learning (ICML 2024), PMLR 235, 2024, pp. 46280–46302.

20 N. Meister, C. Guestrin and T. Hashimoto, Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 2025, pp. 24–49.

21 J. Ni, Y. Fan, V. Zouhar, D. Rooein, A. M. Hoyle, M. Sachan, M. Leippold, D. Hovy and E. Ash, Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), 2026, pp. 36–54.

22 J. Hou, C. Ao, H. Wu, X. Kong, Z. Zheng, D. Tang, C. Li, X. Hu, R. Xu, S. Ni and M. Yang, Findings of the Association for Computational Linguistics: ACL 2024, 2024.

23 N. Liu, S. Sonkar and R. G. Baraniuk, Artificial Intelligence in Education (AIED 2025), Lecture Notes in Computer Science, vol. 15880, 2025, pp. 364–377.

24 K. K. Maurya, K. A. Srivatsa, K. Petukhova and E. Kochmar, Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 2025.

25 S. Sonkar, K. Ni, S. Chaudhary and R. Baraniuk, Findings of the Association for Computational Linguistics: EMNLP 2024, 2024.

26 LearnLM Team, Google, *LearnLM: Improving Gemini for learning*, https://arxiv.org/abs/2412.16429, 2024.

27 D. Dinucu-Jianu, J. Macina, N. Daheim, I. Hakimi, I. Gurevych and M. Sachan, Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025.

28 J. Liu, Z. Huang, T. Xiao, J. Sha, J. Wu, Q. Liu, S. Wang and E. Chen, Advances in Neural Information Processing Systems 37 (NeurIPS 2024), 2024.

29 J. Macina, N. Daheim, S. P. Chowdhury, T. Sinha, M. Kapur, I. Gurevych and M. Sachan, Findings of the Association for Computational Linguistics: EMNLP 2023, 2023.

30 R. E. Wang, Q. Zhang, C. Robinson, S. Loeb and D. Demszky, Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 2024.

31 E. Kochmar, K. K. Maurya, K. Petukhova, K. A. Srivatsa, A. Tack and J. Vasselli, Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025), 2025.

32 J. Zhou, T. Lu, S. Mishra, S. Brahma, S. Basu, Y. Luan, D. Zhou and L. Hou, *Instruction-following evaluation for large language models*, https://arxiv.org/abs/2311.07911, 2023.

33 C. Xia, C. Xing, J. Du, X. Yang, Y. Feng, R. Xu, W. Yin and C. Xiong, Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024.

34 Z. Zhang, S. Li, Z. Zhang, X. Liu, H. Jiang, X. Tang, Y. Gao, Z. Li, H. Wang, Z. Tan, Y. Li, Q. Yin, B. Yin and M. Jiang, Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 2025, pp. 8374–8398.

35 OpenAI, A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, A. Madry, A. Baker-Whitcomb, A. Beutel, A. Borzunov, A. Carney, A. Chow, A. Kirillov, A. Nichol, A. Paino, A. Renzin, A. T. Passos, A. Christakis, A. Conneau, A. Kamali, A. Jabri, A. Moyer, A. Tam, A. Crookes, A. Tootoochian, A. Tootoonchian, A. Kumar, A. Vallone, A. Karpathy, A. Braunstein, A. Cann, A. Codispoti, A. Galu, A. Kondrich, A. Tulloch, A. Mishchenko, A. Baek, A. Jiang, A. Pelisse, A. Woodford, A. Gosalia, A. Dhar, A. Pantuliano, A. Nayak, A. Oliver, B. Zoph, B. Ghorbani, B. Leimberger, B. Rossen, B. Sokolowsky, B. Wang, B. Zweig, B. Hoover, B. Samic, B. McGrew, B. Spero, B. Giertler, B. Cheng, B. Lightcap, B. Walkin, B. Quinn, B. Guarraci, B. Hsu, B. Kellogg, B. Eastman, C. Lugaresi, C. Wainwright, C. Bassin, C. Hudson, C. Chu, C. Nelson, C. Li, C. J. Shern, C. Conger, C. Barette, C. Voss, C. Ding, C. Lu, C. Zhang, C. Beaumont, C. Hallacy, C. Koch, C. Gibson, C. Kim, C. Choi, C. McLeavey, C. Hesse, C. Fischer, C. Winter, C. Czarnecki, C. Jarvis, C. Wei, C. Koumouzelis, D. Sherburn, D. Kappler, D. Levin, D. Levy, D. Carr, D. Farhi, D. Mely, D. Robinson, D. Sasaki, D. Jin, D. Valladares, D. Tsipras, D. Li, D. P. Nguyen, D. Findlay, E. Oiwoh, E. Wong, E. Asdar, E. Proehl, E. Yang, E. Antonow, E. Kramer, E. Peterson, E. Sigler, E. Wallace, E. Brevdo, E. Mays, F. Khorasani, F. P. Such, F. Raso, F. Zhang, F. von Lohmann, F. Sulit, G. Goh, G. Oden, G. Salmon, G. Starace, G. Brockman, H. Salman, H. Bao, H. Hu, H. Wong, H. Wang, H. Schmidt, H. Whitney, H. Jun, H. Kirchner, H. P. de Oliveira Pinto, H. Ren, H. Chang, H. W. Chung, I. Kivlichan, I. O'Connell, I. Osband, I. Silber, I. Sohl, I. Okuyucu, I. Lan, I. Kostrikov, I. Sutskever, I. Kanitscheider, I. Gulrajani, J. Coxon, J. Menick, J. Pachocki, J. Aung, J. Betker, J. Crooks, J. Lennon, J. Kiros, J. Leike, J. Park, J. Kwon, J. Phang, J. Teplitz, J. Wei, J. Wolfe, J. Chen, J. Harris, J. Varavva, J. G. Lee, J. Shieh, J. Lin, J. Yu, J. Weng, J. Tang, J. Jang, J. Q. Candela, J. Beutler, J. Landers, J. Parish, J. Heidecke, J. Schulman, J. Lachman,

J. McKay, J. Uesato, J. Ward, J. W. Kim, J. Huizinga, J. Sitkin, J. Kraaijeveld, J. Gross, J. Kaplan, J. Snyder, J. Achiam, J. Jiao, J. Zhuang, J. Harriman, K. Fricke, K. Hayashi, K. Singhal, K. Shi, K. Karthik, K. Wood, K. Rimbach, K. Hsu, K. Nguyen, K. Gu-Lemberg, K. Button, K. Liu, K. Howe, K. Muthukumar, K. Luther, L. Ahmad, L. Kai, L. Itow, L. Workman, L. Pathak, L. Chen, L. Jing, L. Guy, L. Fedus, L. Zhou, L. Mamitsuka, L. Weng, L. McCallum, L. Held, L. Ouyang, L. Feuvrier, L. Zhang, L. Kondraciuk, L. Kaiser, L. Hewitt, L. Metz, L. Doshi, M. Aflak, M. Simens, M. Boyd, M. Thompson, M. Dukhan, M. Chen, M. Gray, M. Hudnall, M. Zhang, M. Aljubeh, M. Litwin, M. Zeng, M. Johnson, M. Shetty, M. Gupta, M. Shah, M. Yatbaz, M. J. Yang, M. Zhong, M. Glaese, M. Chen, M. Janner, M. Lampe, M. Petrov, M. Wu, M. Wang, M. Fradin, M. Pokrass, M. Castro, M. O. T. de Castro, M. Pavlov, M. Brundage, M. Wang, M. Khan, M. Murati, M. Bavarian, M. Lin, M. Yesildal, N. Soto, N. Gimelshein, N. Cone, N. Staudacher, N. Summers, N. LaFontaine, N. Chowdhury, N. Ryder, N. Stathas, N. Turley, N. Tezak, N. Felix, N. Kudige, N. Keskar, N. Deutsch, N. Bundick, N. Puckett, O. Nachum, O. Okelola, O. Boiko, O. Murk, O. Jaffe, O. Watkins, O. Godement, O. Campbell-Moore, P. Chao, P. McMillan, P. Belov, P. Su, P. Bak, P. Bakkum, P. Deng, P. Dolan, P. Hoeschele, P. Welinder, P. Tillet, P. Pronin, P. Dhariwal, Q. Yuan, R. Dias, R. Lim, R. Arora, R. Troll, R. Lin, R. G. Lopes, R. Puri, R. Miyara, R. Leike, R. Gaubert, R. Zamani, R. Wang, R. Donnelly, R. Honsby, R. Smith, R. Sahai, R. Ramchandani, R. Huet, R. Carmichael, R. Zellers, R. Chen, R. Nigmatullin, R. Cheu, S. Jain, S. Altman, S. Schoenholz, S. Toizer, S. Miserendino, S. Agarwal, S. Culver, S. Ethersmith, S. Gray, S. Grove, S. Metzger, S. Hermani, S. Zhao, S. Wu, S. Jomoto, S. Xia, S. Phene, S. Papay, S. Narayanan, S. Coffey, S. Lee, S. Hall, S. Balaji, T. Broda, T. Stramer, T. Xu, T. Gogineni, T. Christianson, T. Sanders, T. Patwardhan, T. Cunninghman, T. Degry, T. Dimson, T. Raoux, T. Shadwell, T. Zheng, T. Underwood, T. Markov, T. Sherbakov, T. Rubin, T. Stasi, T. Kaftan, T. Heywood, T. Peterson, T. Walters, T. Eloundou, V. Qi, V. Moeller, V. Monaco, V. Kuo, V. Fomenko, W. Chang, W. Zheng, W. Zhou, W. Manassra, W. Sheu, W. Zaremba, Y. Patil, Y. Qian, Y. Kim, Y. Cheng, Y. Zhang, Y. He, Y. Jin, Y. Dai and Y. Malkov, *GPT-4o system card*, `https://arxiv.org/abs/2410.21276`, 2024.

36 OpenAI, *GPT-5 and GPT-5.4 model release*, 2025.

37 Anthropic, *Claude Opus 4.6, Sonnet 4.6, and Haiku 4.5 model cards*, 2025.

38 DeepSeek-AI, *DeepSeek-V2: A strong, economical, and efficient mixture-of-experts language model*, `https://arxiv.org/abs/2405.04434`, 2024.

39 A. Grattafiori and others, *The Llama 3 herd of models*, `https://arxiv.org/abs/2407.21783`, 2024.