

000 INTELLIASK: LEARNING TO ASK CRITICAL QUESTIONS WITH HUMAN-ALIGNED REWARDS

001
002
003
004
005 **Anonymous authors**

006 Paper under double-blind review

007 008 009 ABSTRACT

010
011 Peer review relies on substantive, evidence-based questions, but existing LLM-
012 based approaches often generate surface-level queries. We find that LLM-
013 generated questions take over 50% of their question tokens from a paper’s first
014 page, while human reviewers draw on the full text. Human questions are also more
015 insightful, showing effort and grounding, whereas LLM questions mostly reflect
016 surface style. To address this, we extract 151k candidate questions from ICLR
017 2024 reviews and filter them through a multi-stage filtering process into Probe-
018 15K, a set of 15.5k high-quality questions. From this, we create ProbeVote-500,
019 where human annotators score questions along effort, evidence, and grounding.
020 Using these labels, we train IntelliReward, a reward model built from a frozen
021 Autoregressive LLM with trainable multi-head transformers over the final 50 to-
022 ken states. This architecture outperforms API-based SFT finetuning (Gemini 2.5
023 Flash, GPT-4.1) as baselines for reward. Applying DAPO with IntelliReward, we
024 train IntelliAsk, a question-generation model aligned with human preferences and
025 substantially stronger than existing fine-tuned review models. Finally, by releasing
026 Probe-15K, ProbeVote-500, and IntelliReward, we provide an automatic evalua-
027 tion benchmark for reviewer questions that measures groundedness, effort, and
028 evidence. We release our implementation and datasets to ensure reproducibility
029 (code available at <https://anonymous.4open.science/r/IntelliA-3E09/>).

030 1 INTRODUCTION

031
032 Asking critical and well-reasoned questions is essential for advancing research, as such questions
033 help clarify ideas, reveal limitations, and inspire new directions. In academic publishing, peer review
034 plays a key role in this process, relying on reviewers to raise questions that improve the quality
035 and impact of scientific work. However, as the number of submissions to major conferences has
036 grown, the quality of reviewer feedback has declined. Many reviewers are overloaded and face tight
037 deadlines, leading some to rely on large language models (LLMs) to draft questions and comments
038 (Liang et al., 2024). While LLMs can produce fluent text, the questions they generate often lack
039 technical depth, proper reasoning, or contextual understanding of the work.

040 **Why existing resources are not enough.** Most of the recent research works propose methods to
041 improve the review generation capabilities of the LLMs. However, there’s no focus on the quality of
042 critic and the questions in the review generated by the models trained using these techniques, hence
043 rendering the review useless. Closer to our setting, Idahl & Ahmadi (2025), fine-tunes LLaMA-
044 8B on 79k reviews, but the generated questions extracted from the peer review just mimic the tone
045 of reviewer style (See Section 4). The generated questions “sound” human, without offering a
046 comprehensible and thoughtful question. Chitale et al. (2025) uses a Graph based approach for
047 generating peer reviews. While the graph structure helps organize paper content, the model still
048 relies on simple supervised fine-tuning and produces questions that lack critical depth, remaining
049 shallow imitations of human phrasing. Moreover, both Idahl & Ahmadi (2025) & Chitale et al.
050 (2025) evaluate their systems primarily with automated review-quality scores from LLM judges,
051 without incorporating human-in-the-loop assessments to measure whether the questions are actually
052 useful to authors. Similarly, Dasigi et al. (2021) uses only titles and abstracts to generate questions,
053 limiting the scope for creating technically detailed peer questions that are meaningful to authors.
Overall, these approaches frame the task too broadly - treating it as generic review or QA generation-
without explicitly modeling what makes reviewer questions effortful, evidence-based, and grounded.

054 **Challenges.** Generating effective review questions is not the same task as producing generic QA
 055 pairs based on the available content. LLM-generated questions often lack a clear understanding of
 056 technical content, resulting in questions that may be verbose and lengthy but unhelpful or already
 057 answered in the paper. Our own experiments highlight this gap: we conducted an experiment where
 058 four expert annotators evaluated the questions generated by 3 strong baseline LLMs. They rated four
 059 variants of questions (3 model-generated and 1 human-written question from Openreview) each from
 060 o3, Gemini 2.5 Pro, Qwen2.5-32B and compared them to real human-authored questions. When
 061 evaluated with our rubric, **humans scored 0.78 points higher on average than the strongest model**
 062 and **1.53 points higher on average than the lowest scoring model** (see Table 3). The results show
 063 that human-written questions were consistently more relevant and useful. They were categorized
 064 to be written with more effort, contained evidence from the paper and weren't just framed using
 065 keywords from the paper, while the converse was true for the questions asked by the LLMs.
 066

067 **Our Work.** In this paper, we address the challenge of generating critical, well-reasoned review
 068 questions. We curate **Probe-15K**, a dataset of 15,500 reviewer questions from ICLR 2024 sub-
 069 missions, filtered for technical depth, clarity, and diversity. We then introduce **ProbeVote-500**, an
 070 expert-annotated set of question–paper pairs scored on Effort, Evidence, and Grounding, and use
 071 it to train **IntelliReward**, a reward model that serves as a scalable evaluation benchmark aligned
 072 with expert judgments. Finally, we show that while supervised fine-tuning (SFT) mostly imitates
 073 reviewer style, reinforcement learning (RL with DAPO) guided by IntelliReward achieves closer
 074 alignment with human-authored questions.
 075

076 Our contributions are as follows:
 077

078 1. **Probe-15K dataset:** A rigorously filtered collection of 15,500 reviewer questions drawn
 079 from 151K OpenReview entries for ICLR 2024, via multi-stage filtering- assessing sub-
 080 stance, effort, and grounding, to arrive at 15K questions.
 081
 082 2. **ProbeVote-500 and IntelliReward:** ProbeVote-500, A dataset of expert-annotated ques-
 083 tion–paper pairs using three criteria - Effort, Evidence, and Grounding. From it, we build
 084 IntelliReward, an auto-eval benchmark that aligns more closely with human judgment and
 085 outperforms API based LLM-as-judge baselines tuned using SFT. To validate our reward
 086 model, we trained a 7B and 32B Model using IntelliReward for quality critical question
 087 generation.
 088

089 2 PROBE-15K

090 2.1 LARGE-SCALE EXTRACTION OF QUESTIONS FROM OPENREVIEW REVIEWS

091 We collected a dataset of reviewer feedback by scraping all publicly available reviews from ICLR
 092 2024 using the OpenReview API. For each paper, we retrieved the corresponding metadata and
 093 downloaded the main PDF (excluding supplementary materials), limiting the maximum length to
 094 nine pages.
 095

096 An Openreview submission includes several structured fields: *Summary*, *Strengths*, *Weaknesses*,
 097 *Questions*, *Limitations*, *Ethical Concerns*, numerical scores for *Soundness* and *Overall Evaluation*,
 098 and the reviewer's *Confidence*. In practice, however, reviewers do not consistently confine their
 099 questions to the *Questions* field. To characterize variability in question placement, we manually
 100 annotated a random sample of 100 reviews, observing that questions frequently appeared outside the
 101 designated *Questions* section, sometimes they are present within the *Weaknesses* or, less frequently,
 102 the *Strengths* (See Fig12 in A.9). In some cases, the *Questions* section points to other sections (e.g.,
 103 “See Weaknesses”), or mixed multiple questions with commentary.
 104

105 To address this variability and extract reviewer questions, we used Gemini 2.0 and prompted it with
 106 the concatenated text of the *Questions*, *Strengths*, and *Weaknesses* sections from each review. The
 107 prompt explicitly instructed the model to copy questions verbatim, preserving their original phrasing
 108 and tone. (see A.13.4 in A for the full prompt). When a reviewer wrote multiple independent queries
 109 in a single sentence, the model split them into separate entries. To verify the accuracy, we manually
 110 inspected 500 extracted questions to ensure that the model consistently retained the original phrasing
 111 and did not hallucinate content.
 112

108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161

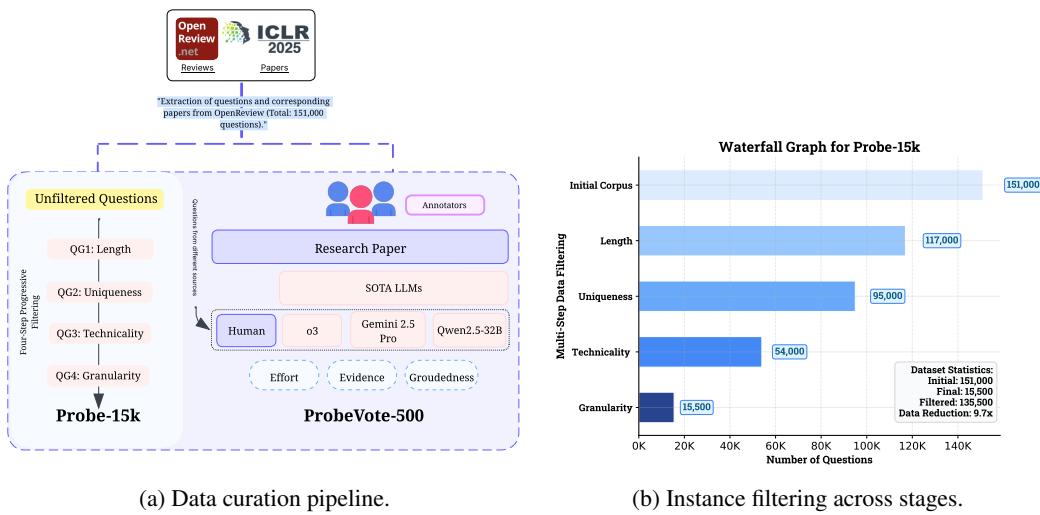


Figure 1: Overview of the dataset construction process: (a) multi-stage data filtering steps for curation, and (b) waterfall diagram illustrating progressive filtering at each stage

2.2 CONSTRUCTING PROBE-15K AND ITERATIVE QUALITY ENHANCEMENT

To construct a dataset suitable for downstream modeling, we applied a series of filtering steps guided by best practices from CVPR reviewer slides (Davis, 2021), NeurIPS (NeurIPS, 2023) and ICLR reviewer guidelines (ICLR, 2025), prior work on LLM feedback for reviews (Thakkar et al., 2025), and our own manual inspection of roughly 2,000 reviews. The initial extraction produced about 151,000 questions. Our goal was not simply to maximize quantity but to ensure that the retained questions were clear, specific, and technically relevant. Each filtering stage systematically removed low-quality or redundant entries. After every stage, we manually checked a random sample of about 1,000 questions to confirm that the filtering criteria were effective and that valid questions were not being discarded.

Length-Based Filtering. We first excluded questions under 100 characters. Manual analysis showed that short questions typically contained superficial comments or clarifications readily apparent in the submission text. This filtering step removed 34,000 entries, resulting in a subset of 117,000 questions. We then proceed to remove semantically similar questions.

Eliminating Semantically Redundant Questions. Numerous questions were semantically identical apart from minor variations in wording. Training on highly redundant content increases the risk of overfitting and limits output diversity. To address this, we applied clustering using Stella with a cluster size of $k=5$. This reduced the dataset to 95,000 questions. After this stage of filtering there were still many questions which were non-technical and not relevant to the content of the paper for which we employ another stage of filtering described further.

Filtering Non-Technical and Irrelevant Content. Manual review identified many questions unrelated to the technical content, including remarks on grammar, formatting, typographic errors, and unprofessional or subjective comments. Prior work (Liang et al., 2024) has shown that reviews containing certain keywords (e.g., “commendable,” “innovative”) are often generated by language models. To mitigate this, we developed a prompt specifying six exclusion criteria, detailed in the Appendix (See A.13.1). Importantly, we provided Gemini 2.0 Flash with both the review text and the corresponding paper as context, ensuring that ungrounded or off-topic questions could be more reliably detected and filtered. This process removed 41,000 questions. Even after this stage, we observed remaining questions that were purely opinion-based or that dismissed techniques without justification, which were addressed in the subsequent filtering stage.

Filtering for Specificity and Actionability. The final stage removed questions that were vague or speculative. We targeted two categories: (i) incomplete, rhetorical, or opinion-based questions without supporting evidence; (ii) unsupported assertions that a technique would fail or had been

162 previously published (See A.13.2 in A). Questions were sequentially evaluated, retaining only those
 163 that satisfied all criteria. This step removed 38,500 questions, resulting in a final corpus of approxi-
 164 mately 15,500 diverse, technically relevant entries.

165 After filtering, the final dataset contained 15.5k questions drawn from 5,841 unique pa-
 166 pers. The train dataset contains 13.2k questions and the test dataset contains 2.3k ques-
 167 tions. To prepare the corresponding paper content for evaluation and training, we applied
 168 olmOCR (allenai/olmOCR-7B-0825-FP8) (Poznanski et al., 2025) to extract structured text
 169 from the first nine pages of each paper.

171 3 BENCHMARKING SOTA REASONING LLMs AGAINST HUMANS

172 LLMs are capable of generating reviews when provided with a complete paper, however, where
 173 they tend to fall short is in asking compelling questions that involve critical thinking about the
 174 content of the paper and as well as the domain knowledge of the paper under consideration. To
 175 study this, we conduct a human annotation study comparing questions extracted from OpenReview
 176 reviews with those generated by several state-of-the-art LLMs.

177 We primarily do this for below two reasons:

- 180 1. To benchmark and quantify the gap between human and LLM-generated questions
- 181 2. To create the preference data required to train a reward model to scale annotation.

184 3.1 PROBEVOTE-500: HUMAN PREFERENCE AND ANNOTATION DATA

185 **Experimental Setup.** ProbeVote-500 consists of 572 annotated question–paper pairs abstracted
 186 from 300 randomly sampled ICLR 2025 submissions on Openreview. For each paper, the full text
 187 was provided as input to the following large language models : Gemini 2.5 Flash (Reasoning model),
 188 o3 (Reasoning model), Qwen2.5-32B , under an identical prompting template (see A.13.3), yielding
 189 one model-generated question per system. In parallel, the corresponding human-authored reviewer
 190 question from Openreview was included as the reference. To eliminate source bias, all questions
 191 were anonymized before annotation. Human evaluators read each paper in full, including text, fig-
 192 ures, and equations, to ensure proper context (See A.10 for the User-Interface used by Annotators).
 193 If a paper was entirely outside an annotator’s domain expertise, it was marked as skipped and re-
 194 assigned. Annotators then scored each anonymized question according to the rubric introduced in
 195 Section 3.2, which evaluates three binary dimensions: Effort, Evidence, and Grounding.

197 3.2 RUBRICS FOR ASSESSING QUESTION QUALITY: EFFORT, EVIDENCE, AND GROUNDING

198 To evaluate question quality, we design a rubric with three binary metrics: Effort, Evidence, and
 199 Grounding. Each metric is scored as 0/1, keeping the evaluation simple and consistent across an-
 200 notators. We chose a binary scheme to reduce ambiguity and to focus on whether a question meets
 201 the essential qualities of being thoughtful and useful for authors. See A.11 for examples of each
 202 category.

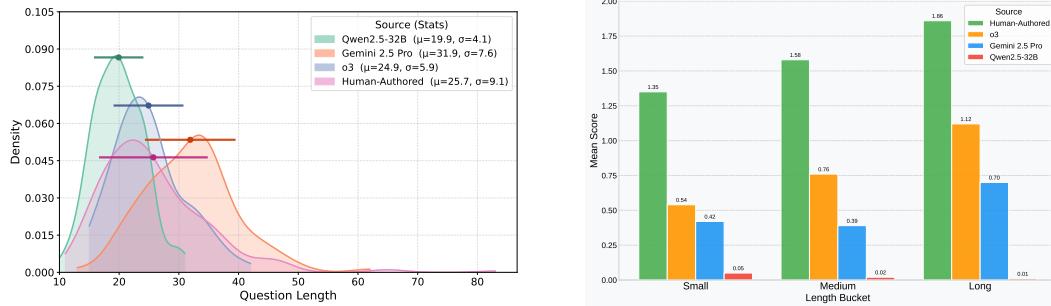
- 204 1. **Effort:** Does the question demand real thought to answer? Low-effort questions can be
 205 answered by directly quoting the paper or restating surface-level details, whereas a high-
 206 effort question requires the reader to synthesize ideas, connect sections, or identify non-
 207 obvious implications beyond what is stated.
- 208 2. **Evidence:** Is the question backed by specific content from the paper? High-evidence ques-
 209 tions point to particular results, assumptions, or arguments in the work and probe them
 210 critically. Low-evidence questions raise points without support, making them speculative
 211 or unhelpful.
- 212 3. **Grounding:** Is the question anchored in the actual content of the paper? Grounded ques-
 213 tions refer to concrete methods, experiments, or claims across sections of the paper. Un-
 214 grounded questions rely only on generic phrasing, keywords, or broad statements that could
 215 apply to almost any paper. For example: What if we increase the depth of the neural net-
 216 work ?

216 3.3 ANALYSIS FROM PROBEVOTE-500
217

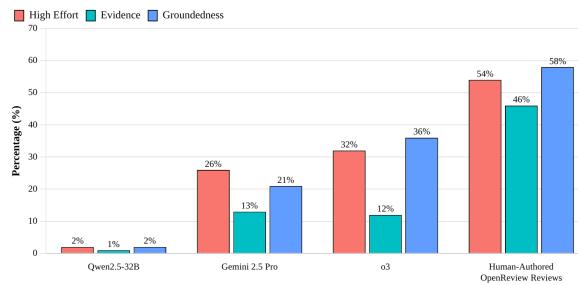
218 **Source Vs Score.** Blind annotation results show that the Qwen2.5-32B model received the lowest
219 scores, while highest quality human-authored questions from Openreview achieved the highest (see
220 Table 3). The mean cumulative score is calculated by taking an average of all the axis of the rubric,
221 with the highest possible score being 3 and lowest 0. This gap becomes even clear when looking at
222 the specific categories scores in Fig 3.

223 **Keyword Coverage.** We measure keyword coverage as the fraction of words in the question that
224 originate from the paper’s first page. This tests whether models rely disproportionately on intro-
225 ductory text when framing questions. A high score indicates surface-level dependence, while lower
226 scores suggest engagement with the full paper. Qwen2.5-32B shows the strongest dependence, with
227 **55%** of question words coming from the first page alone (Table 3). In contrast, Human-authored
228 questions, o3, and Gemini 2.5 Pro achieve relatively low scores, indicating that they draw more
229 evenly from later sections of the paper when constructing questions.

230 **Question Length vs Source.** Figure 2 compares question lengths across sources. Qwen2.5-32B
231 produces the shortest questions, while Gemini 2.5 Pro generates the longest. The average length
232 of o3’s questions is close to that of Human-authored ones, but Humans show the highest variance,
233 reflecting greater diversity and less reliance on fixed phrasing patterns.



246 Figure 2: (Left) Distribution of question lengths across sources. Kernel density estimates are
247 shown for human-authored questions and those generated by Qwen2.5-32B, o3, and Gemini 2.5
248 Pro. (Right) Mean Score Distribution (0-3) vs Length Bucket for different sources



262 Figure 3: The figures show the distribution of votes on Effort, Evidence and Factual metrics for
263 various sources of questions.

266 **Question Length vs Score.** Comparing Human-authored questions with o3 reveals clear gaps in
267 quality. For short questions (< 20 characters), Human-authored ones are more than **2× richer** in
268 quality (effort + evidence + groundedness) than those from o3. The largest gap is in groundedness,
269 where Humans outperform o3 by over **10×**. Effort is also substantially lower for o3, suggesting that
even its concise questions often lack depth and framing.

270 4 WHY SFT FAILS ON THIS TASK
271

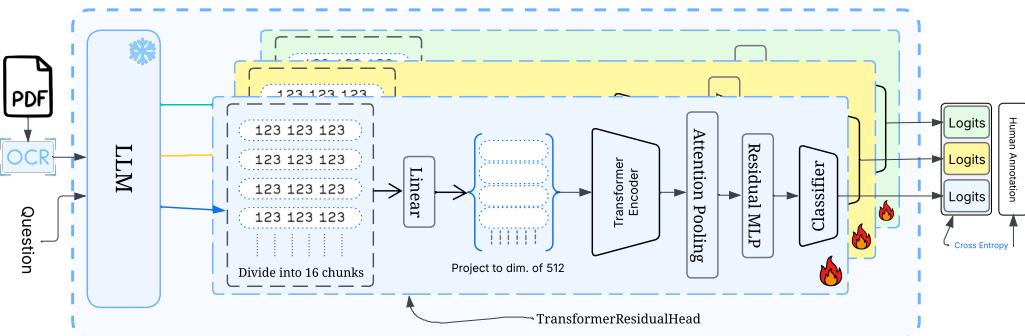
272 We fine-tuned *Qwen/Qwen2.5-7B-Instruct-1M* on Probe-15K using Human-authored questions as
273 the reference for reviewer-style generation. Training ran on four H200 GPUs for 24 hours with an
274 input length of 14K tokens per paper. For evaluation, we held out a test split of 2200 samples from
275 Probe-15K and used the same prompts as in our human annotation study to ensure fairness.

276 The fine-tuned model learned to mimic the phrasing and tone of reviewers but did not improve
277 in producing meaningful questions: depth, reasoning, and grounding remained weak compared
278 to Human-authored questions (see Table 3). We also tested existing SFT-trained reviewer
279 models (OpenReviewer, DeepReviewer, AutoRev) by extracting the *Questions* section of their outputs.
280 Their results were fluent in style but shallow in substance, lacking the critical depth of Human-
281 written questions (See A.1).

282 These findings show that SFT captures style but not reasoning. High-quality reviewer questions re-
283 quire more than surface imitation, motivating our next step: reinforcement learning with IntelliRe-
284 ward, a reward model trained to capture human preferences along Effort, Evidence, and Grounding.
285

286 5 TRAINING INTELLIASK: A SPECIALIZED MODEL FOR ASKING CRITICAL
287 QUESTIONS
288

290 As shown in Section 3 and table 3, SFT does not improve the model’s performance on the critical
291 question generation task. This limitation is consistent with recent findings showing that SFT often
292 memorizes training data and struggles with out-of-distribution scenarios. Because of this tendency,
293 it struggles to adapt to new situations. Reinforcement learning (RL), on the other hand, encourages
294 exploration and learning from feedback, which helps it generalize better and handle tasks that require
295 complex reasoning (Chu et al., 2025).

296 5.1 REWARD MODEL : INTELLIREWARD
297310 Figure 4: Architecture and training of the IntelliReward .
311

312 Evaluating all 15,500 questions with human annotators across three rubrics is costly and risks bias
313 from fatigue. This highlights the need for a reliable automatic evaluation benchmark to support
314 the scaling of our experiments. To reduce reliance on manual effort, we tested leading closed-
315 source LLMs on the reward prediction task. However, they showed weak predictive accuracy (Table
316 1), required large inputs, and incurred high inference costs, making them unsuitable for large-scale
317 benchmarking. To overcome this, we trained IntelliReward on ProbeVote-500 to serve as an efficient
318 and scalable substitute for human judgment. The architecture and training procedure are described
319 in the following subsection.

320 5.2 REWARD MODEL ARCHITECTURE AND TRAINING
321

322 **Reward Model Architecture.** Our reward model handles multiple objectives by pairing a causal
323 LLM with per-objective Transformer heads. We use `gpt-oss-20b` (medium reasoning) as the

324 Table 1: Reward prediction accuracy of different models. We compare off-the-shelf closed-source
 325 LLMs, the same models tuned with SFT using their APIs, an open-source baseline (Qwen2.5-7B-
 326 Instruct-1M), and our trained model IntelliReward on the test split of ProbeVote-500.

Candidate reward model	Checkpoint	Effort (%)	Evidence (%)	Grounded (%)	Mean Accuracy(%)
<i>Closed-source LLMs (off-the-shelf)</i>					
Gemini 2.5 Flash	Original/API	57	25	29	37
GPT-4.1	Original/API	44	22	30	32
GPT-5	Original/API	56	54	49	53
<i>Closed-source LLMs (tuned with SFT via API)</i>					
Gemini 2.5 Flash	SFT/API	61	53	45	53
GPT-4.1	SFT/API	52	25	31	36
<i>Open-source baseline</i>					
Qwen2.5-7B-Instruct-1M	Original	30	26	28	28
<i>Our trained reward model</i>					
IntelliReward (ours)	–	70	76	70	72

343 base. Given an input (e.g., paper OCR, generated question, task prompt), the LLM encodes it into a
 344 fixed representation. We extract the pooled hidden states of the last 50 output tokens and pass it to
 345 our per-objective Transformer head, which empirically improves performance as compared to using
 346 MLP head. (see Table 2). The resulting representation is denoted as
 347

$$r \in \mathbb{R}^H, \quad H = 2880,$$

348 where r is the pooled hidden representation of the LLM outputs and H is its dimensionality.

349 Each evaluation objective $j \in \{1, \dots, k\}$ has an independent head $f_j(\cdot)$ producing logits $\ell_j \in \mathbb{R}^{C_j}$,
 350 where k is the total number of objectives and C_j is the number of classes (or possible labels) for
 351 objective j . Each `TransformerResidualHead` first chunks r into n segments and projects
 352 them to dimension d_{model} , then processes the sequence through L Transformer encoder layers. A
 353 learnable attention query pools the sequence into a vector $z \in \mathbb{R}^{d_{\text{model}}}$, which is refined via a residual
 354 two-layer feedforward network (MLP):
 355

$$z' = \text{LayerNorm}(z + \text{FFN}(z)),$$

356 where $\text{FFN}(\cdot)$ is the feedforward transformation and $\text{LayerNorm}(\cdot)$ denotes layer normalization.
 357 Finally, the refined vector is mapped to logits:

$$\ell_j = W_j z' + b_j,$$

358 where $W_j \in \mathbb{R}^{C_j \times d_{\text{model}}}$ and $b_j \in \mathbb{R}^{C_j}$ are learnable weights and biases for head j .
 359

360 **Training Objective and Inference** During training, the model minimizes the total loss $\mathcal{L} =$
 361 $\sum_{j=1}^k \text{CE}(\ell_j, y_j)$, where CE denotes cross-entropy and y_j is the ground-truth label for objective
 362 j . During inference, each head predicts $\hat{y}_j = \arg \max \ell_j$, and the final score is computed as
 363 $S = \sum_{j=1}^k \hat{y}_j$.
 364

365 **Reward Model Training.** We train IntelliReward using preference annotations from ProbeVote-
 366 500. The frozen LLM provides representations, while only the per-objective heads $f_j(\cdot)$ are updated.
 367 Training follows the cross-entropy loss defined above. We optimize with AdamW (learning rate
 368 2×10^{-5} , batch size 8, weight decay 0.01) for 5 epochs on a single NVIDIA L40S GPU. End-to-end
 369 training completes within 30 minutes. The Per-objective Head is lightweight and only takes total of
 370 300MB of GPU VRAM during inference.
 371

372 5.3 RL USING INTELLIREWARD REWARD MODEL

373 As shown in Section 4, supervised fine-tuning (SFT) performs poorly for review question generation:
 374 the model copies surface style but does not produce questions with real effort, evidence,
 375

378 Table 2: Ablation study comparing **MLP** vs. **TransformerResidualHead** (Ours). We group results
 379 by head architecture to show the impact of using MLP vs TransformerResidualHead, Pool-k = mean
 380 pooling over last k output tokens.

Configuration		Accuracy Metrics (%)			
Base Model	Pooling	Effort	Evidence	Grounding	Mean
Head Architecture: Standard MLP					
Frozen	None	61	64	61	62
Frozen	Pool50	64	67	64	65
Trainable	None	64	65	60	63
Trainable	Pool50	65	69	67	67
Trainable	Pool128	64	68	66	66
Head Architecture: Transformer Residual (Ours)					
Frozen	None	68	68	70	69
Frozen	Pool50	70	76	70	72
Frozen	Pool128	69	77	67	71
Trainable	None	71	69	70	70
Trainable	Pool50	71	78	70	73
Trainable	Pool128	70	78	68	72

402 or grounding. To address this, we use our reward model, **IntelliReward**, to align generation with
 403 human preferences. Fig 5 shows the difference in reward curve for both Qwen2.5-7B-1M and
 404 IntelliAsk.



417 Figure 5: The figures show the difference in reward curves for Qwen2.5-7B (SFT) and IntelliAsk
 418 during training.

421 We train with the DAPO(Yu et al., 2025) for IntelliAsk-7B and using GRPO for IntelliAsk-32B
 422 algorithm: for each paper, the model generates several candidate questions, which are scored by
 423 IntelliReward, and these scores are used as rewards to guide optimization. Training follows the
 424 standard DAPO and GRPO setup (batch sizes, sequence length, gradient clipping, learning rate
 425 schedule; see Appendix A.12). The resulting model, **IntelliAsk-32B**, consistently outperforms SFT-
 426 only baselines (See table 3) by producing questions that are more evidence-based, better grounded,
 427 and require greater effort. IntelliAsk-32B also shows strong gains on external benchmarks.

6 RELATED WORK

431 Recent research has increasingly explored the use of large language models (LLMs) to automate
 432 aspects of peer review. Several works train models on large corpora of reviews, often through super-

432 Table 3: Comparison of reward-model evaluated scores and human-evaluated scores. **Bold** indicates
 433 best performance; underline indicates second best. Component scores (Effort, Evidence, Grounding)
 434 range from 0 to 1. Total score ranges from 0 to 3 (sum of components). For Coverage, **lower**
 435 percentages (↓) indicate better performance.

437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466	437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466	Model / Source	Reasoning	Scores [0–1]			Total [0–3]	Coverage (%) ↓
				Effort	Evidence	Grounding		
Reward-Model Evaluated: Large Models								
gpt-oss-120b	Medium	0.08	0.15	0.12	0.35	22.99		
gpt-4.1	No	0.07	0.12	0.12	0.31	31.73		
gpt-5	Default	0.09	0.20	0.16	0.45	18.63		
o3	Medium	0.28	0.14	0.30	0.72	16.81		
claude-3.7-sonnet	No	0.09	<u>0.18</u>	0.15	0.42	45.14		
claude-3.7-sonnet	Default	0.08	0.16	0.13	0.37	47.13		
gemini-2.5-flash	No	0.08	0.15	0.15	0.38	39.06		
gemini-2.5-pro	Default	0.22	0.11	0.18	0.51	25.75		
llama-4-maverick	No	0.09	0.17	0.15	0.41	48.48		
grok-4	No	0.07	0.14	0.12	0.33	35.47		
deepseek-chat-v3.1	Default	0.11	0.20	0.17	0.48	36.83		
Reward-Model Evaluated: Small Open-Source Models ($\leq 32B$)								
gpt-oss-20b	Medium	0.06	0.11	0.10	0.27	24.81		
Qwen2.5-7B SFT	No	0.00	0.01	0.02	0.03	42.11		
OpenReviewer	No	0.00	0.00	0.10	0.10	51.14		
DeepReviewer	No	0.00	0.00	0.10	0.10	48.14		
IntelliAsk-32B (Ours)	Default	<u>0.23</u>	0.12	<u>0.20</u>	<u>0.68</u>	21.37		
IntelliAsk-7B (Ours)	No	0.03	0.07	0.07	0.17	27.44		
Human-Evaluated Scores								
Human reviewer questions	–	0.54	0.46	0.57	1.57	28.21		
o3	Medium	0.32	0.12	0.36	0.80	16.81		
Gemini 2.5 Pro	Default	0.26	0.13	0.21	0.60	25.75		
IntelliAsk-32B (Ours)	Default	0.27	0.13	0.26	0.66	21.37		
Qwen2.5-32B	No	0.02	0.01	0.02	0.05	54.96		

467
 468
 469 vised fine-tuning (SFT). For instance, Idahl & Ahmadi (2025) introduce *OpenReviewer*, fine-tuning
 470 LLaMA-8B on 79K reviews to produce fluent and structured assessments, while Zhu et al. (2025)
 471 develop *DeepReview*, a multi-stage pipeline that integrates retrieval and self-reflection, supported by
 472 the curated DeepReview-13K dataset. Similarly, Tan et al. (2025) propose *ReviewMT*, a dataset of
 473 110K review comments enabling multi-turn, role-based review dialogue. While these systems im-
 474 prove stylistic fluency and tone, they primarily focus on generating full reviews rather than isolating
 475 and producing the probing questions or issue-driven feedback that most benefits authors.

476 Other approaches explore multi-agent frameworks. D’Arcy et al. (2024) propose *MARG*, which dis-
 477 tributes paper sections across specialized agents (e.g., clarity, experiments, impact) that collaborate
 478 to generate comprehensive feedback, mitigating context-length limitations and improving coverage.
 479 Similarly, Chamoun et al. (2024) introduce *SWIFT²T*, which decomposes review generation into plan-
 480 ner, investigator, reviewer, and controller modules to provide focused, actionable comments. These
 481 approaches enhance specificity and helpfulness relative to earlier baselines that mostly generate
 482 general feedback or superficial style corrections.

483 Several datasets and evaluation frameworks also relate closely. Baumgärtner et al. (2025), Sundar
 484 et al. (2024), and Singh et al. (2024) harvest reviewer questions and author responses—facilitating
 485 tasks such as answer generation or content retrieval rather than explicit question generation itself.
 On the evaluation side, recent work such as GEM PiCO (Ning et al., 2025), and ReviewCritique

486
 487 Table 4: This table compares the performance of IntelliAsk-32B and Qwen3-32B across various
 488 external benchmarks to show generalization of IntelliAsk across different domains. The detailed
 489 categorical results for WritingBench can be found in A.7 and additional results for the Trait Bench-
 490 mark are included in A.8.

Benchmark	Scores		Primary Skill Tested	Metric
	IntelliAsk-32B	Qwen3-32B		
Reasoning & Comprehension				
Eluther/DROP	95.1	93.3	Discrete reading comprehension, numerical reasoning	F1/Acc
MuSR	68.3	64.7	Multistep soft reasoning (e.g., mysteries, allocation)	Accuracy
BoolQ	90.0	90.0	Contextual reading comprehension, Yes/No QA	Accuracy
GPQA-Diamond	69.1	68.4	Graduate-level expert scientific reasoning	Accuracy
Writing & Generation				
WritingBench	8.31	8.07	Core writing domains (creative, persuasive, technical)	0–10
Arena Hard	94.1	93.8	Alignment with human preference.	0-100
Domain Generalization				
Conf. Mix '25*	0.65	0.07	Generalization to unseen conference domains	Score (0–3)

500
 501 *We sampled 100 random papers from a pool of ICML 2025, NeurIPS 2025, and CVPR 2025 from different
 502 tracks. The purpose is to test generalization to other conferences and domains using the IntelliReward Score
 503 (0–3).

504
 505
 506 (Du et al., 2024) analyze the quality of reviews via off-the-shelf LLM judges or annotated corpora,
 507 focusing on fluency, coverage, consistency, and groundedness. Almost all of these works rely on
 508 SFT or prompting, and none explicitly train a model purely for reviewer-style question generation
 509 using human-labeled question data.

510
 511
 512
 513 Despite this progress, existing research overwhelmingly treats peer review as a problem of generating
 514 full reviews or answering reviewer questions. Very little attention has been given to *question*
 515 *generation itself*—the actionable and constructive element of peer feedback. Moreover, the dominant
 516 reliance on SFT or LLM-as-judge evaluations leaves a gap in aligning generation with the
 517 qualities that authors value most: effortful engagement, grounded critique, and context-aware probing.
 518 Our work directly addresses this gap by introducing a human-annotated dataset of reviewer-
 519 style questions, and by training with supervised fine-tuning to generate them, thereby offering a
 520 new benchmark and model geared specifically toward generating probing, useful questions in peer
 521 review.

522 7 CONCLUSION

523
 524 We introduce Probe-15K, a large-scale dataset of high-quality reviewer questions; ProbeVote-500,
 525 a human-annotated evaluation set for effort, evidence, and grounding; and IntelliReward, a reward-
 526 model-based benchmark that more faithfully captures expert preferences than existing LLM-judge
 527 baselines. Building on these resources, we train IntelliAsk, a model that generates reviewer-style
 528 questions which are deeper and more useful than those produced by SFT-based models. Together,
 529 these contributions provide both a scalable benchmark and a specialized model for advancing the
 530 quality of automated peer review. Looking ahead, we plan to scale IntelliAsk to larger foundation
 531 models, extend our framework across conferences and domains, and incorporate multimodal inputs
 532 such as figures and tables, advancing toward richer and more comprehensive critiques of scientific
 533 work.

540 USE OF LARGE LANGUAGE MODELS
541542 We used LLMs to check for grammar, and rewrite some sentence for better readability. We did not
543 use LLMs for generating technical content or any idea.
544545 ETHICAL CONSIDERATION AND DATA LICENSING
546547 The Probe-15K dataset was created from reviewer comments on ICLR papers that are publicly avail-
548 able on OpenReview.net. We restricted the collection to text that is already accessible to the public
549 and removed any metadata that could identify reviewers. As OpenReview content is distributed un-
550 der the Creative Commons Attribution 4.0 International (CC BY 4.0) license, our use and release of
551 these comments complies with the original license terms.
552553 The ProbeVote-500 dataset was constructed through additional human annotation on top of the col-
554 lected review comments. These annotations are original contributions by our paper and are released
555 under the same CC BY 4.0 license.556 We do not claim copyright over the original review texts or paper excerpts used in our datasets.
557558 REPRODUCIBILITY STATEMENT
559560 We have taken steps to make our results reproducible. The datasets (Probe-15K and ProbeVote-500),
561 the reward model (IntelliReward), and the question generation model (IntelliAsk) will be released
562 publicly with clear documentation. The preprocessing steps and filtering procedures are described in
563 Section 2 and the Appendix A.13.1. The architectures, training procedures, and hyperparameters for
564 IntelliReward and IntelliAsk are described in Section 3 and Appendix A. We used standard training
565 frameworks and provide implementation details sufficient for replication.
566567 We release the code, datasets, and trained model checkpoints under the CC BY 4.0 license to support
568 further research and ensure that our results can be independently verified.
569570 REFERENCES
571572 Tim Baumgärtner, Ted Briscoe, and Iryna Gurevych. PeerQA: A scientific question answering
573 dataset from peer reviews. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings*
574 *of the 2025 Conference of the Nations of the Americas Chapter of the Association for Compu-*
575 *tational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 508–544,
576 Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-
577 89176-189-6. doi: 10.18653/v1/2025.naacl-long.22. URL <https://aclanthology.org/2025.naacl-long.22/>.
578579 Eric Chamoun, Michael Schlichtkrull, and Andreas Vlachos. Automated focused feedback gener-
580 ation for scientific writing assistance. In Andre Ku, Lun-Wei ands Martins and Vivek Srikumar
581 (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9742–9763,
582 Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/
583 v1/2024.findings-acl.580. URL <https://aclanthology.org/2024.findings-acl.580/>.
584585 Maitreya Prafulla Chitale, Ketaki Mangesh Shetye, Harshit Gupta, Manav Chaudhary, and Vasudeva
586 Varma. Autorev: Automatic peer review system for academic research papers. *arXiv preprint*
587 *arXiv*: 2505.14376, 2025.588 Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V
589 Le, Sergey Levine, and Yi Ma. SFT memorizes, RL generalizes: A comparative study of founda-
590 tion model post-training. In *Forty-second International Conference on Machine Learning*, 2025.
591 URL <https://openreview.net/forum?id=dYur3yabMj>.
592593 Mike D’Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. Marg: Multi-agent review generation
594 for scientific papers, 2024. URL <https://arxiv.org/abs/2401.04259>.

594 Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset
 595 of information-seeking questions and answers anchored in research papers. *arXiv preprint arXiv:*
 596 *2105.03011*, 2021.

597

598 D. Davis. Cvpr 2021 training materials: Reference slides, 2021. URL <http://luthuli.cs.uiuc.edu/~daf/CVPR21TrainingMaterials/RefSlides.pdf>.

600 Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng
 601 Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, Haoran Ranran Zhang, Vipul
 602 Gupta, Yinghui Li, Tao Li, Fei Wang, Qin Liu, Tianlin Liu, Pengzhi Gao, Congying Xia, Chen
 603 Xing, Cheng Jiayang, Zhaowei Wang, Ying Su, Raj Sanjay Shah, Ruohao Guo, Jing Gu, Hao-
 604 ran Li, Kangda Wei, Zihao Wang, Lu Cheng, Surangika Ranathunga, Meng Fang, Jie Fu, Fei
 605 Liu, Ruihong Huang, Eduardo Blanco, Yixin Cao, Rui Zhang, Philip S. Yu, and Wenpeng
 606 Yin. LLMs assist NLP researchers: Critique paper (meta-)reviewing. In Yaser Al-Onaizan,
 607 Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical
 608 Methods in Natural Language Processing*, pp. 5081–5099, Miami, Florida, USA, November
 609 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.292. URL
 610 <https://aclanthology.org/2024.emnlp-main.292/>.

611 ICLR. Leveraging llm feedback to enhance review quality, Apr 2025. URL <https://blog.iclr.cc/2025/04/15/leveraging-llm-feedback-to-enhance-review-quality/>.

612

613

614 Maximilian Idahl and Zahra Ahmadi. OpenReviewer: A specialized large language model for
 615 generating critical scientific paper reviews. In Nouha Dziri, Sean (Xiang) Ren, and Shizhe
 616 Diao (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of
 617 the Association for Computational Linguistics: Human Language Technologies (System Demos-
 618 trations)*, pp. 550–562, Albuquerque, New Mexico, April 2025. Association for Compu-
 619 tational Linguistics. ISBN 979-8-89176-191-9. doi: 10.18653/v1/2025.naacl-demo.44. URL
 620 <https://aclanthology.org/2025.naacl-demo.44/>.

621

622 Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao
 623 Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel A. McFarland, and James Y. Zou. Monitoring
 624 ai-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews.
 625 In *ICML*, 2024. URL <https://openreview.net/forum?id=bX3J7ho18S>.

626

627 Reichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher
 628 Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou,
 629 Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. Webgpt:
 630 Browser-assisted question-answering with human feedback, 2022. URL <https://arxiv.org/abs/2112.09332>.

631

632 NeurIPS. Reviewer guidelines, 2023. URL <https://neurips.cc/Conferences/2023/ReviewerGuidelines>.

633

634 Kun-Peng Ning, Shuo Yang, Yuyang Liu, Jia-Yu Yao, Zhenhui Liu, Yonghong Tian, Yibing Song,
 635 and Li Yuan. PiCO: Peer review in LLMs based on consistency optimization. In *The Thirteenth
 636 International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=sfQ6XpApfS>.

637

638 Jake Poznanski, Jon Borchardt, Jason Dunkelberger, Regan Huff, Daniel Lin, Aman Rangapur,
 639 Christopher Wilhelm, Kyle Lo, and Luca Soldaini. olmOCR: Unlocking Trillions of Tokens
 640 in PDFs with Vision Language Models, 2025. URL <https://arxiv.org/abs/2502.18443>.

641

642 Shruti Singh, Nandan Sarkar, and Arman Cohan. SciDQA: A deep reading comprehension dataset
 643 over scientific papers. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Pro-
 644 ceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp.
 645 20908–20923, Miami, Florida, USA, November 2024. Association for Computational Linguis-
 646 tics. doi: 10.18653/v1/2024.emnlp-main.1163. URL <https://aclanthology.org/2024.emnlp-main.1163/>.

648 Anirudh Sundar, Jin Xu, William Gay, Christopher Gordon Richardson, and Larry Heck. cPAPERS:
 649 A dataset of situated and multimodal interactive conversations in scientific papers. In *The Thirty-*
 650 *eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*,
 651 2024. URL <https://openreview.net/forum?id=DfhcOelEnP>.
 652

653 Cheng Tan, Dongxin Lyu, Siyuan Li, Zhangyang Gao, Jingxuan Wei, Siqi Ma, Zicheng Liu, and
 654 Stan Z. Li. Peer review as a multi-turn and long-context dialogue with role-based interactions:
 655 Benchmarking large language models, 2025. URL <https://openreview.net/forum?id=uV3Gdoq2ez>.
 656

657 Nitya Thakkar, Mert Yuksekgonul, Jake Silberg, Animesh Garg, Nanyun Peng, Fei Sha, Rose Yu,
 658 Carl Vondrick, and James Zou. Can llm feedback enhance review quality? a randomized study of
 659 20k reviews at iclr 2025. *arXiv preprint arXiv: 2504.09737*, 2025.
 660

661 Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai,
 662 Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guang-
 663 ming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu,
 664 Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao
 665 Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingx-
 666 uan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL
 667 <https://arxiv.org/abs/2503.14476>.
 668

669 Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. DeepReview: Improving LLM-based
 670 paper review with human-like deep thinking process. In Wanxiang Che, Joyce Nabende, Eka-
 671 terina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting*
 672 *of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29330–29355,
 673 Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-
 674 251-0. doi: 10.18653/v1/2025.acl-long.1420. URL <https://aclanthology.org/2025.acl-long.1420/>.
 675

676

A APPENDIX

A.1 EXAMPLES QUESTIONS GENERATED FROM OPENREVIEWER, DEEPREVIEWER

Paper 1:

Human-Authored: Concerning the computation requirement, one gradient descent step seems to become at least slightly more computational expensive as compared to DQN. In the paper it is mentioned at the very end of the discussion section that "with K=5 (it) only requires 1 to 2 times more time to run". Can this quantification be more accurate or why would it "sometimes" same compute time and "sometimes" double the compute time?

IntelliAsk-32B: The analysis in Section 5 uses Theorem 5.1 to bound the gap between the learned and optimal value functions, showing that iDQN's error depends on the sum of approximation errors over K Bellman iterations. However, the bound also contains problem-dependent constants C_{K,γ,R_γ} and $F(r; K, \rho, \gamma)$, whose scaling with K is not explicitly characterized. How do these terms grow as K increases, and what does this imply for the trade-off between reduced approximation error and potential amplification of concentrability coefficients or other problem-dependent factors? In particular, does increasing K asymptotically improve the bound, or does it introduce a non-trivial optimization trade-off when choosing K in practice?

OpenReviewer: Why did you not compare to other DQN variants?

DeepReview: Could the authors provide a clear explanation of the proposed algorithm, and how it compares to existing algorithms?

Paper 2:

Human-Authored: Could you provide more theoretical justification for the choice of an asymmetric loss function over other well-established loss functions like Huber loss or quantile loss? How does it specifically benefit the forecasting model in capturing the cost of errors?

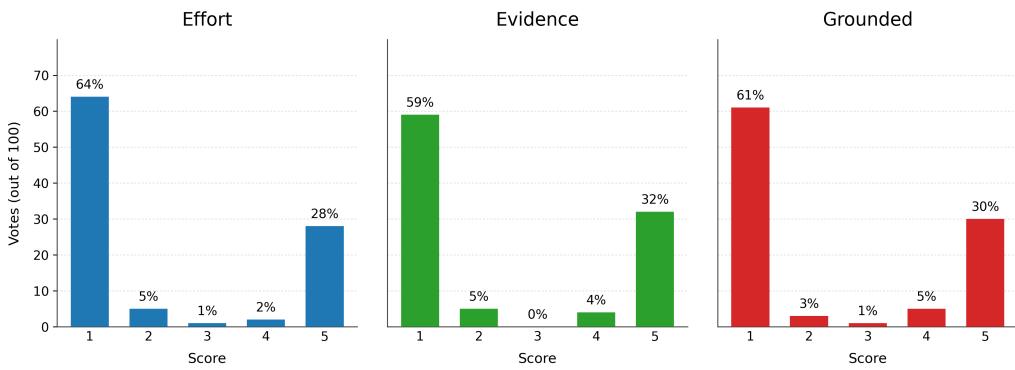
702 **OpenReviewer:** The authors should identify a research gap in electrical load forecasting and
 703 attempt to fill the gap with a novel method. The authors should clearly state what is novel about
 704 their package.

705 **DeepReviewer:** What is the definition of "domain-specific feature engineering"? How does it
 706 differ from general feature engineering techniques?

708 *The weights for AutoRev aren't open-sourced so we referred to the questions presented in the paper
 709 for evaluating the quality of questions.

710 A.2 OBSERVATIONS FROM LIKERT SCORING IN RUBRIC

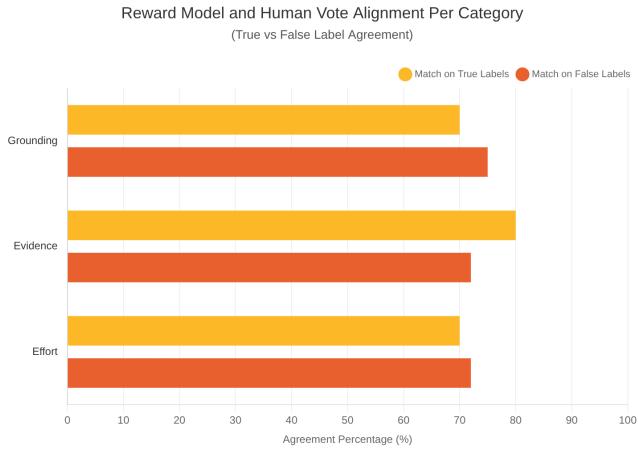
712 We explored Likert scoring initially. In our pilot phase, annotators used a 1-5 scale for Effort, Evidence,
 713 and Grounding. However, once we did 25% of annotation, we observed a strong bimodal
 714 pattern: more than 85% of ratings clustered at the extremes (1 or 5), with very sparse use of inter-
 715 mediate values which can be seen in Figure 6



729 Figure 6: The graphs shows distribution of votes on different categories during pilot annotation.
 730 This clearly shows the votes clustered at the extremes (1 or 5).

732 A.3 REWARD MODEL AND HUMAN VOTE ALIGNMENT

734 Figure 7 illustrates the agreement between our reward model and human annotators. We evaluated
 735 this alignment across three key categories: Grounding, Evidence, and Effort. The results show that
 736 the model consistently matches human judgment, achieving over 70% agreement on both 'True' and
 737 'False' labels for every category.



754 Figure 7: Comparison of reward model agreement with human annotations across three evaluation
 755 dimensions on Positive and Negative labels.

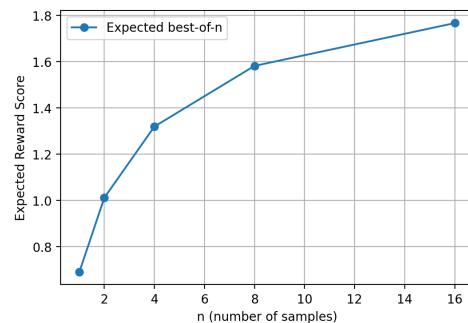
756 A.4 REJECTION SAMPLING
757

758 Referring to the setup used in (Nakano et al., 2022), we performed rejection sampling by generating
759 16 completions for each of 300 prompts from the ProbeVote-500 test set. We set the temperature
760 to **0.9** and computed best-of- n for $n \in \{1, 2, 4, 6, 8, 16\}$. Completions were generated using GPT-5
761 and Gemini-2.5-Pro.

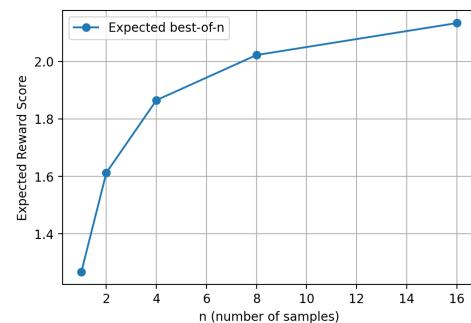
762 The annotators then manually inspected these samples to verify whether the reward scores matched
763 the actual quality of the generated questions. Below, we include few examples from this analysis in
764 Table 5. We have summarized the best-of- n results for both the models in Table 6 along with the
765 expected reward curve for Gemini 2.5 Pro and GPT-5 in Fig 8.

766 Table 5: Comparison of generated questions. The gray headers indicate the specific model being
767 evaluated.

770 Question	771	772 Score
GPT-5		
773 In Algorithm 1, Eq. (2) appears to subtract identical terms at x_{t-1} ; was the intended SPIDER-style 774 recursion $u_t^s = u_{t-1}^s + (1/ A) \sum_{j \in A} [\nabla f_{sj}(x_t; \xi_{sj}) - \nabla f_{sj}(x_{t-1}; \xi_{sj})]$, and if so, can you show 775 why this estimator yields an unbiased λ_t -weighted common descent direction?	3.0	
776 Why is permutation invariance inappropriate for Event Cloud processing, and how do PEPNet's tai- 777 lored hierarchical structure with temporal attention aggregation achieve state-of-the-art relocalization 778 accuracy?	0.0	
Gemini 2.5 Pro		
781 How does the paper's decomposition of the Bayes-Adaptive MDP's Q-value into an 'Incremental 782 Value of Information' and a 'Value of Opportunity' explain why different classes of reward shaping 783 functions are effective?	2.0	
784 How does the proposed framework enhance the robustness of reinforcement learning agents against 785 adversarial state perturbation-inference techniques tailored for different types of environments?	0.0	



(a) Gemini 2.5 Pro



(b) GPT-5

800 Figure 8: Expected Reward Score using Best-of- n scaling
801
802803 A.5 QUESTION PREFERENCE: INTELLIASK-32B VS GPT-4.1, GEMINI-2.5 FLASH,
804 QWEN3-32B
805

806 To assess model quality, we conducted pairwise human preference evaluations. We compared
807 IntelliAsk-32B against three strong baselines: Gemini 2.5-Flash, GPT-4.1, and Qwen3-32B. Across
808 all comparisons, IntelliAsk-32B achieved higher preference rates, winning between 81% and 96%
809 of evaluated pairs (See Fig 9. These results highlight a substantial advantage in human-aligned
behavior relative to other models.

Table 6: Best-of- n Performance Scaling: Gemini 2.5 Pro vs. GPT-5

<i>n</i>	Gemini 2.5 Pro		GPT-5	
	Mean Reward Score	Gain vs $n=1$	Mean Reward Score	Gain vs $n=1$
1	0.6896	–	1.2667	–
2	1.0114	+0.3218	1.6125	+0.3458
4	1.3192	+0.6296	1.8649	+0.5982
8	1.5816	+0.8920	2.0222	+0.7555
16	1.7667	+1.0771	2.1333	+0.8667

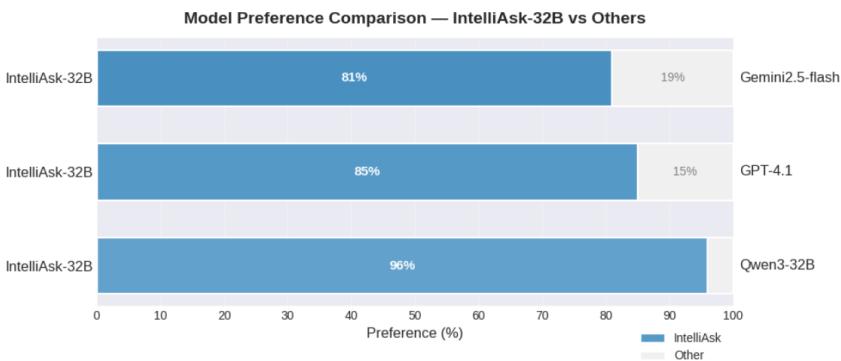


Figure 9: Human evaluators consistently favor IntelliAsk-32B over other leading models. Across comparisons with Gemini-2.5-Flash, GPT-4.1, and Qwen3-32B, IntelliAsk-32B receives 81–96% of total preferences.

A.6 INTER-ANNOTATOR AGREEMENT ON PROBEVOTE-500

We also report the Inter-annotator agreement during the annotation phase and found that, across the three final attributes: Effort, Evidence, and Grounding, the annotators achieved stable and consistent reliability. Figure 10 reports the Cohen’s kappa scores for each question source, demonstrating consistent agreement levels among human annotators.

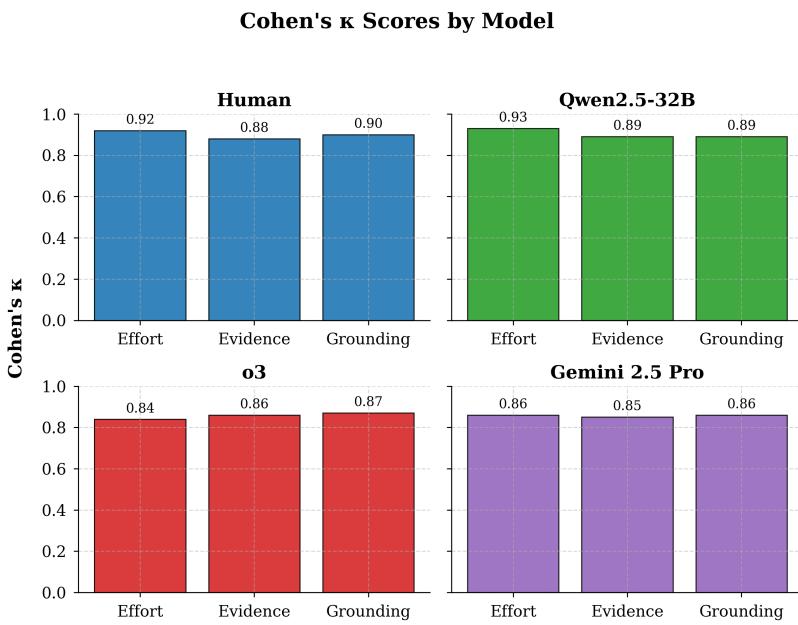


Figure 10: Cohen's κ agreement scores across three evaluation categories (Effort, Evidence, Grounding)

A.7 SCORE DISTRIBUTION IN WRITINGBENCH

Table 7 provides a detailed breakdown of the score distribution of IntelliAsk-32B and Qwen3-32B on WritingBench. The results indicate that **IntelliAsk-32B** demonstrates dominating performance, surpassing Qwen3-32B in the vast majority of evaluated domains and document categories.

Table 7: Detailed Performance Comparison by Domain (Higher Score Bolded) out of 10

Category	Score: IntelliAsk-32B	Score: Qwen3-32B
Academic & Engineering	8.325	8.093
Finance & Business	8.216	8.039
Politics & Law	8.292	8.015
Literature & Arts	8.405	8.155
Education	8.268	8.089
Advertising & Marketing	8.371	8.176
Abstract	8.000	7.947
Introduction	8.000	7.845
Contributions	8.667	8.338
Limitations	8.360	8.169
Conclusion	8.600	8.257
Literature Review	8.300	8.305
Experiments	8.533	8.110
Defense Presentation	7.933	7.751
Defense Script	7.960	7.739
Technical Documentation	8.450	8.305
Research Proposal	8.333	7.817
Internship Report	8.800	8.599
Engineering Report	8.700	8.403
Patent	8.300	8.305

Continued on next page...

918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
Table 7 – continued from previous page

Metric	Score: IntelliAsk-32B	Score: Qwen3-32B
Contract	8.164	7.941
Test Report	8.350	8.012
User Research	7.933	7.719
Meeting Minutes	8.400	8.305
Briefing	8.367	8.045
Financial Reports	7.969	7.787
Tender Document	8.178	7.991
Bid Proposal	8.257	7.761
Requirements Specification	8.450	8.354
Product Proposal	8.314	8.179
Investment Analysis	8.175	8.208
Risk Management	8.167	8.176
Market Analysis	7.960	8.110
Human Resource Management	8.400	8.240
Market Research	8.400	8.305
Recruitment	8.300	8.208
Pitch Deck	8.433	8.176
Event Planning	8.320	8.130
Business Correspondence	8.000	7.622
Party Membership Application	9.000	8.745
Mean	8.306	8.070

A.8 TRAIT BENCHMARK

949
950
951
952
953
954
The scores for the TRAIT benchmarks are reported in Table 8. For Neuroticism and Dark Triad
955 traits, lower scores are generally considered "better" or safer for AI alignment.

956
Table 8: Detailed comparison of personality trait scores between IntelliAsk-32B and Qwen3-32B
957 on the Trait Benchmark. The traits are categorized into the Big Five and the Dark Triad.

Trait	IntelliAsk-32B	Qwen3-32B
Big Five Traits		
Openness	0.679	0.611
Conscientiousness	0.714	0.754
Extraversion	0.364	0.485
Agreeableness	0.667	0.781
Neuroticism	0.160	0.209
Dark Triad Traits		
Machiavellianism	0.115	0.258
Narcissism	0.105	0.115
Psychopathy	0.000	0.016

A.9 IDENTIFYING QUESTIONS WITHIN REVIEWS

Summary: This paper proposes a multi-modal LLM, called any-to-any MM-LLM, to extend the multi-modality of LLM to a state where there is no limitation on the input and output modality combinations. To achieve this goal, the authors (1) propose a lightweight alignment learning technique to achieve an effective semantic alignment across different modalities with limited trainable parameters and (2) annotate a modality-switching instruction tuning dataset. The displayed results and visualizations suggest the promising performance of the tuned any-to-any MM-LLM.

Soundness: 3 good

Presentation: 4 excellent

Completeness: 3 good

Strengths:

- Extending the multi-modal LLMs free of limitation on the input/output modalities is an important research question that can facilitate a wider range of applications.
- The introduced dataset, if made publicly available, would be a good contribution to the community.
- Various evaluation benchmarks are used to benchmark the proposed model with existing solutions.
- The writing is clean and easy to follow

Weaknesses:

- The proposed alignment learning technique is a bit naive and does not consider much about the challenge introduced by the any-to-any modality, such as how to balance the performance across different modalities
- Although introducing contents from different modalities during tuning is considered to improve the overall performance of the model, in the experiment section, it seems introducing these additional modalities actually leads to worse performance on benchmark datasets. Does this indicate the alignment technique is not effective enough as expected?

Questions:

Will the pretrained model and dataset be released to the public?

Flag For Ethics Review: Yes, Discrimination / bias / fairness concerns, Yes, Privacy, security and safety, Yes, Responsible research practice (e.g., human subjects, data release)

Details Of Ethics Concerns:

The proposed dataset's content may need a deeper look from experts to check its content. And the content generated by the model may need further checking to make sure there are no harmful contents generated.

Rating: 6: marginally above the acceptance threshold

Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

Code Of Conduct: Yes

Add: [Public Comment](#)

Figure 11: Variability in the occurrence of questions in a review

A.10 UI OF HUMAN ANNOTATION TOOL

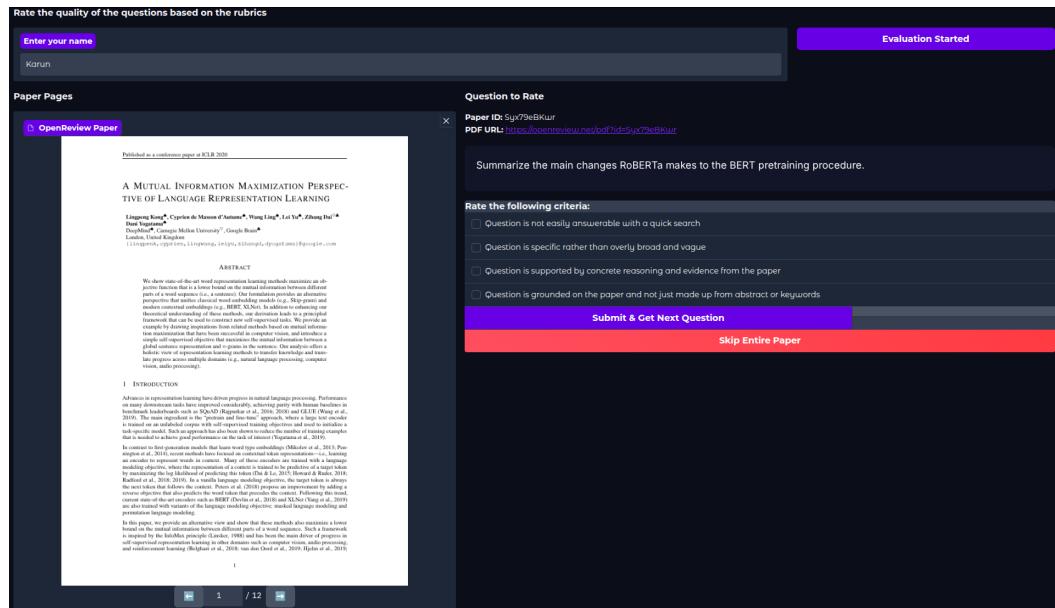


Figure 12: UI of Human Annotation tool (Figure shows dummy data.)

A.11 EXAMPLES OF EFFORTFUL, SUBSTANTIVE, AND EVIDENCE-BASED QUESTIONS

1026

1027

1028

1029

1030

1031

1032

1033

1034

1035

1036

1037

1038

1039

1040

1041

1042

1043

1044

1045

1046

1047

1048

1049

1050

1051

1052

1053

1054

1055

1056

1057

1058

1059

1060

1061

1062

1063

Table 9: Analysis of Peer Review Questions by Quality Dimension

	High	Low
Effort	Why is the training time of NoLA with shared random basis similar to that of LoRA when the training time of NoLA with a unique random basis is higher? Aren't the number of coefficients being trained the same in both cases? This is a high-effort question because it requires reasoning about subtle implementation differences between NoLA and LoRA variants. Answering it in-depth involves connecting training dynamics and model design choices, which goes beyond what is explicitly stated in the paper.	How does the proposed Δ -SGD method adapt to the heterogeneity in local data across different clients and datasets compared to other optimization methods as shown in the experimental results? This is a low-effort question because the abstract and results already explain how Δ -SGD adapts to heterogeneous client data. The answer can be found by directly quoting or restating surface-level details, without requiring deeper reasoning or synthesis.
Evidence	'This way, we transform the optimization-based estimation into a feed-forward prediction, thus bypassing the time-consuming gradient computation and avoiding sub-optimality via large-scale training on a wide spectrum of distributions.' — For MINE, we do need to update NNs' parameters. But InfoNet also needs gradient ascent. How to understand 'bypassing the time-consuming gradient computation'? This is a high-evidence question because it cites a specific claim from the paper and directly challenges a possible inconsistency. The reasoning is grounded in the authors' own statement, making the critique precise and well-supported.	What specific improvements or changes in the recommendation system's architecture or methodology did the authors implement to achieve improved performance compared to traditional item and user embedding-based recommendation systems? This is a low-evidence question because it asks broadly about improvements without pointing to any specific claim, experiment, or section of the paper. It raises a generic point without evidence-based grounding.

Continued on next page

1064

1065

1066

1067

1068

1069

1070

1071

1072

1073

1074

1075

1076

1077

1078

1079

1080
 1081
 1082
 1083
 1084
 1085
 1086
 1087
 1088
 1089
 1090
 1091
 1092
 1093
 1094
 1095
 1096
 1097
 1098
 1099
 1100

Table 9 – continued from previous page

	High	Low
Grounding	In section 4.2 you mentioned that you used LORA to inject low-rank matrices into attention weights Q, K and V only and freeze all other weights inside the Transformer, given that there are other large MLPs inside it, what is the rationale of only applying LoRA to Q, K and V? This is a high-grounding question because it explicitly refers to a section of the paper and to concrete implementation choices (applying LoRA only to Q, K, and V). It probes a methodological decision directly anchored in the text.	How does the proposed Deep Reinforcement Learning (DRL) framework in this paper address the trade-off between minimizing taxi delays and ensuring sufficient runway throughput in mixed-mode runway operations, and how does this compare to existing methods like DRL in Ali et al. (2022)? This is a low-grounding question because the comparison to prior work is posed in generic terms and does not engage with specific details of the method described. The reference to Ali et al. is already mentioned in the paper, so the question does not add a deeper, paper-specific probe.

1101

A.12 HYPERPARAMETERS FOR TRAINING INTELLIAK

1102
 1103

Listing 1: Training parameters for IntelliAsk

```

1104 # =====
1105 # Experiment metadata
1106 # =====
1107 project_name="DAPo-QReward"
1108 exp_name="DAPo-Qwen2.5-7B"
1109 MODEL_PATH=/path/to/hf_cache/Qwen/Qwen2.5-7B-Instruct-1M
1110 CKPTS_DIR=/path/to/ckpts/${project_name}/${exp_name}
1111
1112 # =====
1113 # Core training hyperparameters
1114 # =====
1115 adv_estimator="grpo"
1116 clip_ratio_low=0.20
1117 clip_ratio_high=0.28
1118
1119 max_prompt_length=14000
1120 max_response_length=$((1024 * 20))
1121
1122 enable_overlong_buffer=True
1123 overlong_buffer_len=$((max_prompt_length + 1024))
1124 overlong_penalty_factor=1.0
1125
1126 loss_agg_mode="token-mean"
1127
1128 enable_filter_groups=True
1129 filter_groups_metric="acc"
1130
1131 max_num_gen_batches=2
1132 train_prompt_bsz=64
1133 gen_prompt_bsz=$((train_prompt_bsz * 3))
1134 n_resp_per_prompt=8
1135 train_prompt_mini_bsz=2
1136
1137 temperature=1.0

```

```

1134 top_p=1.0
1135 top_k=-1
1136 val_top_p=0.7
1137
1138 sp_size=${NGPUS}
1139 use_dynamic_bsz=True

1140 actor_ppo_max_token_len=$((max_prompt_length + max_response_length))
1141 infer_ppo_max_token_len=$((max_prompt_length + max_response_length))

1142 offload=False
gen_tp=${NGPUS}

1143
1144 # =====
# Optimizer & actor configs
# =====
1145 actor_rollout_ref.actor.optim.lr=1e-6
1146 actor_rollout_ref.actor.optim.lr_warmup_steps=10
1147 actor_rollout_ref.actor.optim.weight_decay=0.1
1148 actor_rollout_ref.actor.ppo_mini_batch_size=${train_prompt_mini_bsz}
1149 actor_rollout_ref.actor.entropy_coeff=0.0
1150 actor_rollout_ref.actor.grad_clip=1.0
1151 actor_rollout_ref.actor.loss_agg_mode=${loss_agg_mode}

1152
1153
1154
1155 A.13 SYSTEM PROMPT
1156
1157 A.13.1 QUALITY GATE 3
1158
1159 You are an expert evaluator assessing Questions asked by the reviewers at
1160 top conferences from the CVPR, NeurIPS, ICML, ICLR, EMNLP, after
1161 reading a scientific paper for their suitability in a specialized
1162 dataset aimed at training Large Language Models for advanced
1163 reasoning.
1164
1165 **Goal:** Filter the provided Question to determine if it is a Valid
1166 Question. The question will be a Valid Question if it passes through
1167 all the rules, without getting rejected, resulting in keep = true.
1168
1169 **Input Format:** You will receive a JSON object representing a single
1170 question with fields like 'review_id', 'question'..
1171
1172 **Output Format:** Respond with a JSON object containing two fields:
1173 1. 'keep': A boolean value ('true' or 'false').
1174 2. 'reason': A concise string explaining your decision based on the
1175 specific criteria and rule number(s) below. (e.g., "REJECT: Rule 2-
1176 Question states to correct the caption.", "KEEP: A Valid Question
1177 passed through all the rules.").
1178
1179 **Core Task:** Evaluate the question based primarily* the rules
1180 mentioned below to check their validity and importance in a dataset
1181 used to train a Large Language Model:
1182
1183 **Filtering Criteria & Rules (Apply strictly in this order):**
1184 **Rule 1:** REJECT the questions asking for changes/additions/formatting
1185 that require substantial effort
1186 **Rule 2:** REJECT the questions asking for Edits, Summaries, correcting
1187 typos
1188 Examples of Questions to REJECT under this rule:
1189 Question: In Table 2, it probably needs to be noticed that for COCO
1190 instance segmentation, Mask R-CNN is used
1191 Question: Correct the typo made on page 4, line 3 and add a caption for
1192 figure 3.
1193 **Rule 3:** REJECT the questions if it asks to refer to other sections
1194 like 'See weakness section for questions.

```

```

1188 **Rule 4**: REJECT the questions if it contains unprofessional or
1189 inappropriate remarks in the review and giving personal opinions on
1190 the paper quality
1191 Examples of Questions to REJECT under this rule:
1192 Question: I spend several hours and still can not get an intuitive
1193 understanding about why such a claim hold. For instance, why A and B
1194 are 'irrelevant' according to footnote 6?
1195 Question: The current contribution feels like just \"another score
1196 function\" with no guarantees of identifiability.
1197 Question: Theoretical analysis in main paper seems under developed and
1198 not sure how its useful."
1199
1200
1201 **Rule 5**: REJECT the question if keywords such as review process,
1202 conflict of interest, anonymity, rebuttal, etc.. appear.
1203
1204
1205 **Rule 6**: REJECT the Question if it contains words like commendable and
1206 innovatively since these reviews are most likely generated by LLMs
1207
1208 **Decision Logic Summary:**
1209 * A question MUST pass ALL applicable rules (1 -6) to be kept ('keep:
1210   true').
1211 * Failure at any rule stage leads to rejection ('keep: false').
1212
1213
1214 A.13.2 QUALITY GATE 4

```

```

1215 You are an expert evaluator assessing Questions asked by the reviewers at
1216 top conferences from the CVPR, NeurIPS, ICML, ICLR, EMNLP, after
1217 reading a scientific paper for their suitability in a specialized
1218 dataset aimed at training Large Language Models for advanced
1219 reasoning.
1220
1221 **Goal:** Filter the provided Question to determine if it is a Valid
1222 Question. The question will be a Vaild Question if it passes through
1223 all the rules, without getting rejected, resulting in keep = true.
1224
1225 **Input Format:** You will receive a JSON object representing a single
1226 question with fields like 'review_id', 'question'..
1227
1228 **Output Format:** Respond with a JSON object containing two fields:
1229 1. 'keep': A boolean value ('true' or 'false').
1230 2. 'reason': A concise string explaining your decision based on the
1231   specific criteria and rule number(s) below. (e.g., "REJECT: Rule 2-
1232   Question states to correct the caption.", "KEEP: A Valid Question
1233   passed through all the rules.").
1234
1235 **Core Task:** Evaluate the question based *primarily* the rules
1236 mentioned below to check their validity and importance in a dataset
1237 used to train a Large Language Model:
1238
1239 **Filtering Criteria & Rules (Apply strictly in this order):**
1240 **Group A: Low Specificity / Generic Content**
1241 **Rule 1: REJECT vague or low-specificity questions**
1242 Questions that consist of broad or unclear comments without actionable
1243   suggestions (e.g., Can you elaborate on the methodology?) should be
1244   REJECTED.
1245
1246 **Rule 2: REJECT generic questions about limitations or future work**

```

1242 REJECT questions that ask casually about limitations or future
 1243 directions without referencing a specific issue, weakness, or
 1244 observation in the paper.
 1245 REJECT questions that:
 1246 Casually ask about limitations or future directions without pointing to a
 1247 specific issue, weakness, or observation in the paper.
 1248 Use broad or vague phrasing like "Can you discuss the limitations...", "
 1249 How could future work address this...", or "What are the next steps?"
 1250 without context or justification.

1251 Examples of Questions to REJECT under this rule:
 1252 Question: Can you discuss the limitations of your benchmarking tool, and
 1253 how future research could address these limitations to further
 1254 advance the field of PINNs
 1255 Only keep such questions if they are tied to concrete findings, results,
 1256 or gaps explicitly discussed in the paper.

1257 ****Rule 3: REJECT superficial or generic feedback****
 1258 REJECT out comments that offer only brief praise or criticism without
 1259 actionable insight. Reviewers sometimes provide only a few lines of
 1260 text with little actionable criticism, or simply assign a score
 1261 without justification. This is irrelevant and low quality
 1262 Examples of Questions to REJECT under this rule:
 1263 : Great work! with no follow-up question.

1264 : Writing too bad or not state of the art or too niche etc.. without
 1265 justification.

1266

1267

1268

1269 ****Group B: Incomplete, Speculative, or Opinion-Based Content****
 1270 ****Rule 4: REJECT incomplete or context-less questions****
 1271 REJECT questions that are missing sufficient context or phrasing to be
 1272 actionable and do not make sense.

1273

1274 Example: Not really large-scale.

1275

1276

1277 Example: Ablation studies are missing.
 1278 Question: Besides, 'IGB' is not really ***large-scale*** while some datasets
 1279 like 'ogbn-products' and 'ogbn-papers100M' have millions or hundred
 1280 millions of nodes.

1281

1282 ****Rule 5: Exclude speculative or rhetorical questions****
 1283 REJECT vague or rhetorical speculation without a clear, answerable
 1284 prompt.

1285

1286 Example: I assume they come from different sources...

1287

1288

1289 Example: Would this method fail if we used another model?
 1290 Question: I assume they come from different sources and thus require
 1291 different techniques and efforts to get rid of (if possible

1292

1293 ****Rule 6: Remove personal opinion or preference-based comments****
 1294 REJECT questions/comments that express a personal view without backing
 1295 or relevance.

```

1296
1297 Example: ...which is not that necessary, in my opinion.
1298
1299
1300 **Rule 7: REJECT questions asking for unreported or hypothetical
1301 experiments**
1302 REJECT questions that request speculative experiments beyond the papers
1303 scope, such as trying different models, datasets, or parameters.
1304 Specifically REJECT questions that request unreported experiments or
1305 conjectures beyond the scope of the paper (e.g., "Could this work
1306 better with another model?", "What happens if we try Z instead?").
1307 Examples of Questions to REJECT under this rule:
1308 Question: Compared to Hits@10, Hits@1 could be more critical in the real-
1309 world applications, especially for tail nodes with very few neighbors
1310 . I wonder if the authors can also provide the Hits@1 performance.
1311 Question: Would the method fail if using a non-contrastive pre-trained
1312 model?
1313 The paper mainly focuses on 4-bit and 5-bit quantization, leaving
1314 questions about the performance and relevance of other bit
1315 quantizations
1316
1317 **Rule 8: Exclude questions framed as unsupported suggestions**
1318 REJECT questions like Did you consider X? if they are isolated and not
1319 grounded in the papers content, especially if surrounded by
1320 uninformative praise or vague critique.
1321
1322
1323 Make sure to be strict so that no poor quality question passes through.
1324
1325 **Decision Logic Summary:**
1326 * A question MUST pass ALL applicable rules (1 -6) to be kept ('keep:
1327 true').
1328 * Failure at any rule stage leads to rejection ('keep: false').
1329
1330
1331 A.13.3 QUESTION GENERATION
1332 The prompt shown below was used uniformly across all models for question generation
1333
1334 {"role": "system", "content": "You are expert at asking unique questions
1335 based on the OCR text of a research paper. So given the text,
1336 generate one high quality question now."},
1337 {"role": "user", "content": f"Here's the text of the complete research
1338 paper and now generate a question based on it. \n{ocr_output}"}
1339
1340
1341 A.13.4 EXTRACTION OF QUESTIONS
1342
1343 """You are a highly experienced professor from Stanford University with
1344 extensive experience in reviewing and publishing research papers. You
1345 will be provided with a peer review containing a heading called
1346 Questions and another section called Mixed Content. The Questions
1347 section contains multiple questions without any indication/ separator
1348 for a new question and the Mixed Content has a mix of questions that
1349 might not have a ? to indicate a question. It can simply be a
suggestion, an edit, a clarification required from the author etc.

```

1350 Task: Your Primary task is to Extract Questions first from the Questions
 1351 section and then from the Mixed Content section. Perform verbatim
 1352 extraction. I.e. Word-for-Word
 1353 By Questions I mean all the questions explicitly or implicitly asked
 1354 that the author needs to answer the reviewer based on the review text
 1355
 1356
 1357 1) Extract all the questions from the Questions section in a way all the
 1358 sentences are retained. Do not miss any sentence or words from the
 1359 original content in the section and output multiple Questions you
 1360 have found, you need to break the Questions properly. If someone
 1361 concatenates the multiple questions you have formed, they must get
 1362 the Questions section as it is.
 1363 2) While breaking the questions from the Question section, you might
 1364 encounter nested questions. If both the parts are related keep them
 1365 as a single question but if one part is an independent question, make
 1366 them as separate questions.
 1367 3) Extract all the questions that are present in the Mixed Content
 1368 section. The questions might not be direct, it might include the
 1369 reviewer telling what made him arrive at this question and then pose
 1370 the question. It can also be some clarification he/she needs from a
 1371 content in the paper. So include the complete context and dont simply
 1372 output just the question.
 1373 4) In some cases, the Questions section will direct you to refer the
 1374 Mixed Content section by asking you to refer the weakness. That
 1375 simply is your hint to find questions in the Mixed Content section.
 1376 5) The "Mixed Content" section might have general observations or
 1377 weaknesses of the paper, so only pick up questions, reviewer's
 1378 suggestion for edits, reviewer seeking clarification BUT dont include
 1379 general observations. This is the rule for "Mixed Content" section.
 1380
 1381 Note: The Questions section will always have question present in it until
 1382 unless it is blank or only asking you to refer to the weakness. The
 1383 Mixed Content section might or might not have questions in it, so
 1384 check very carefully. Learn from the zero-shot example below.
 1385 Note 2: Important: When the questions that you form from "Questions"
 1386 section are concatenated, it should form the original and complete
 1387 content of the "Questions" section. This rule of concatenation is
 1388 important and ONLY for "Questions" Section ONLY.
 1389
 1390 Remember: Your task is just extraction of Questions and Not Rephrasing.
 1391
 1392
 1393
 1394
 1395
 1396
 1397
 1398
 1399 Example 1:
 1400
 1401 Input:
 1402
 1403 Paper Id: : Asdho34

```

1404 Review Id: ioedh45
1405 Questions :
1406 I have questions about the learning process of the 11 conv layer in
1407 equation (5). How is it exactly trained? And is it sensitive to the
1408 training sample size?\n
1409 - Will instance normalization also work in text-to-image tasks? It will
1410 be interesting to see if it could generate higher fidelity images
1411 with semantic meaning more aligned with the provided text prompts
1412
1413 Mixed Content :
1414 The proposed method is a systematic approach for image translation tasks
1415 incorporating different components. A potential drawback is its
1416 inference speed. It would be beneficial if the authors could compare
1417 inference speed with other image translation tasks.\n
1418 - The comparison with methods like SDEdit, Prompt2Prompt, and
1419 InstructPix2Pix is somehow unfair since they do not require an
1420 additional segmentation network.\n
1421 - The quantitative evaluation is only the proposed dataset, which
1422 contains fine-grained edit instructions. The effectiveness of DVP
1423 could be further proved by evaluating simple or even ambiguous
1424 instructions
1425
1426 Overall, the paper is well-organized and easy to follow. The figures and
1427 tables are informative.\n
1428
1429 - The performance of the proposed method is promising. Figures 4, 6
1430 clearly demonstrate the superiority of DVP.\n
1431
1432
1433
1434
1435 Output:
1436 {
1437 Questions: [
1438 {
1439 \\"Paper_id\\": Asdho34,
1440 \\"review_id\\": ioedh45,
1441 \\"Q_Number\\": 1,
1442 \\"Question\\": I have questions about the learning process of the 11 conv
1443 layer in equation (5). How is it exactly trained? And is it
1444 sensitive to the training sample size?
1445 },
1446 {
1447 \\"Paper_id\\": Asdho34,
1448 \\"review_id\\": ioedh45,
1449 \\"Q_Number\\": 2,
1450 \\"Question\\": Will instance normalization also work in text-to-image
1451 tasks? It will be interesting to see if it could generate higher
1452 fidelity images with semantic meaning more aligned with the provided
1453 text prompts
1454 },
1455 {
1456 \\"Paper_id\\": Asdho34,
1457 \\"review_id\\": ioedh45,
1458 \\"Q_Number\\": 3,
1459

```

```

1458  \\"Question\\": A potential drawback is its inference speed. It would be
1459  beneficial if the authors could compare inference speed with other
1460  image translation tasks
1461  },
1462
1463  {
1464  \\"Paper_id\\": Asdho34,
1465  \\"review_id\\": ioedh45,
1466  \\"Q_Number\\": 4,
1467  \\"Question\\": The quantitative evaluation is only the proposed dataset,
1468  which contains fine-grained edit instructions. The effectiveness of
1469  DVP could be further proved by evaluating simple or even ambiguous
1470  instructions
1471  }
1472  ]
1473
1474  }
1475
1476
1477
1478 Example 2
1479
1480
1481 Input:
1482
1483 Paper Id: : Asdho34
1484 Review Id: ioedh45
1485 Questions :
1486 Please comment on the weaknesses outlined above.\-
1487 - Figures 10 and 11, right: Why is adaptation slower for OC-GFN than GFN
1488 in the first few thousand iterations? This is surprising since one
1489 would hope pretraining helps bootstrap downstream performance as in
1490 vision / language / RL. If its an exploration phase, did you validate
1491 it and is there a way to side-step it?
1492
1493 Mixed Content :
1494 There should be a discussions of assumptions behind the OC-GFNs
1495 pretraining. Namely, that transfer is only possible when the reward
1496 function changes but not if the action-space or the state-space
1497 change. Moreover, the goal-conditioning requires a well specified set
1498 of outcomes  $Y$  presumably not all states  $s$  are terminal states
1499 which makes the proposed method not truly unsupervised. These
1500 limitations (together with the applicability mentioned at the end of
1501 A.2) could be stated explicitly in the main text, and left to future
1502 work.\-
1503 - While there are enough benchmarks, I believe none include continuous
1504 action/state spaces. Moreover, the experiments only one GFN variant
1505 the detailed-balance one, which is also used for OC-GFN. It would
1506 help validate the generality of OC if we had experiments showing it
1507 worked on these different settings. Moreover, Id be curious to know
1508 how other pretrained amortized sampling baselines (eg, VAEs,
1509 normalizing flows) fare against OC-GFN \xa0and what about
1510 pretraining a GFN on task A (without OC) and fine-tuning it on task B
1511 ?\-
1512 - (minor) The second and fourth paragraphs of Section 4.2 mention the
1513 reasoning potential of GFNs, and that intractable marginalization
1514 leads to slow thinking. Are these anthropomorphisms really needed for
1515 this paper?\-

```

1512 - (minor) I wished the preliminaries (Section 2) included a training
 1513 objective like Eq. 5 & 9, and that these more clearly specified which
 1514 are the optimization variables.\
 1515 - Some typos, there maybe more:\
 1516 - p. 3: multi-objective what?\
 1517 - p. 4: given a reward R a posterior as a function\
 1518 - p. 4: autotelicly autotelically?\
 1519 - p. 5: in log-scale obtained from Eq. (5) should be Eq. 4?'
 1520
 1521 The exposition is generally clear, and I enjoyed reading the paper. The
 1522 authors first present the goal-conditioning idea and how it applies
 1523 to GFNs, then walk the reader through their derivation and
 1524 assumptions for amortized adaptation. I especially appreciated
 1525 Section 2 which gave a clear and concise background.\
 1526 - The paper tackles an impactful problem for GFNs. While the pretraining
 1527 solution is not particularly novel, its a neat application of goal-
 1528 condition RL to an amortized sampling problem. The authors also
 1529 figured out how to make it work on a wide range of problems, and
 1530 provide several ablations in the main text and the appendix.\
 1531 - The insight that a new sampling policy can be readily obtained from an
 1532 outcome-conditioned flow is neat and, as far as I can tell, novel.
 1533 This could spawn interest in outcome-conditioned flows and different
 1534 ways to amortize Eq. 6.
 1535
 1536 Output:
 1537 {
 1538 Questions: [
 1539 {
 1540 **\"Paper_id\"**: Asdho34,
 1541 **\"review_id\"**: ioedh45,
 1542 **\"Q_Number\"**: 1,
 1543 **\"Question\"**: Please comment on the weaknesses outlined above.\
 1544 - Figures 10 and 11, right: Why is adaptation slower for OC-GFN than GFN
 1545 in the first few thousand iterations? This is surprising since one
 1546 would hope pretraining helps bootstrap downstream performance as in
 1547 vision / language / RL. If its an exploration phase, did you validate
 1548 it and is there a way to side-step it?
 1549 },
 1550 {
 1551 **\"Paper_id\"**: Asdho34,
 1552 **\"review_id\"**: ioedh45,
 1553 **\"Q_Number\"**: 2,
 1554 **\"Question\"**: There should be a discussions of assumptions behind the OC
 1555 -GFNs pretraining. Namely, that transfer is only possible when the
 1556 reward function changes but not if the action-space or the state-
 1557 space change
 1558 },
 1559 {
 1560 **\"Paper_id\"**: Asdho34,
 1561 **\"review_id\"**: ioedh45,
 1562 **\"Q_Number\"**: 3,
 1563 **\"Question\"**: These limitations (together with the applicability
 1564 mentioned at the end of A.2) could be stated explicitly in the main
 1565 text, and left to future work.
 1566 },
 1567 {

```

1566  \\"Paper_id\\": Asdho34,
1567  \\"review_id\\": ioedh45,
1568  \\"Q_Number\\": 4,
1569  \\"Question\\": While there are enough benchmarks, I believe none include
1570  continuous action/state spaces. Moreover, the experiments only one
1571  GFN variant the detailed-balance one, which is also used for OC-GFN.
1572  It would help validate the generality of OC if we had experiments
1573  showing it worked on these different settings
1574  },
1575
1576
1577
1578
1579  {
1580  \\"Paper_id\\": Asdho34,
1581  \\"review_id\\": ioedh45,
1582  \\"Q_Number\\": 5,
1583  \\"Question\\": Id be curious to know how other pretrained amortized
1584  sampling baselines (eg, VAEs, normalizing flows) fare against OC-GFN
1585  \\"\\xa0and what about pretraining a GFN on task A (without OC) and fine
1586  -tuning it on task B?
1587  },
1588
1589  {
1590  \\"Paper_id\\": Asdho34,
1591  \\"review_id\\": ioedh45,
1592  \\"Q_Number\\": 6,
1593  \\"Question\\": The second and fourth paragraphs of Section 4.2 mention
1594  the reasoning potential of GFNs, and that intractable marginalization
1595  leads to slow thinking. Are these anthropomorphisms really needed
1596  for this paper?
1597  },
1598
1599  {
1600  \\"Paper_id\\": Asdho34,
1601  \\"review_id\\": ioedh45,
1602  \\"Q_Number\\": 7,
1603  \\"Question\\": I wished the preliminaries (Section 2) included a training
1604  objective like Eq. 5 & 9, and that these more clearly specified
1605  which are the optimization variables},
1606
1607
1608  {
1609  \\"Paper_id\\": Asdho34,
1610  \\"review_id\\": ioedh45,
1611  \\"Q_Number\\": 8,
1612  \\"Question\\": Some typos, there maybe more:
1613  - p. 3: multi-objective what?
1614  - p. 4: given a reward R a posterior as a function
1615  - p. 4: autotelicly autotelically?
1616  - p. 5: in log-scale obtained from Eq. (5) should be Eq. 4?
1617  },
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2260
2261
2262
2263
2264
2265
2266
2267
2268
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2289
2290
2291
2292
2293
2294
2295
2296
2297
2298
2299
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2310
2311
2312
2313
2314
2315
2316
2317
2318
2319
2320
2321
2322
2323
2324
2325
2326
2327
2328
2329
2330
2331
2332
2333
2334
2335
2336
2337
2338
2339
2340
2341
2342
2343
2344
2345
2346
2347
2348
2349
2350
2351
2352
2353
2354
2355
2356
2357
2358
2359
2360
2361
2362
2363
2364
2365
2366
2367
2368
2369
2370
2371
2372
2373
2374
2375
2376
2377
2378
2379
2380
2381
2382
2383
2384
2385
2386
2387
2388
2389
2390
2391
2392
2393
2394
2395
2396
2397
2398
2399
2400
2401
2402
2403
2404
2405
2406
2407
2408
2409
2410
2411
2412
2413
2414
2415
2416
2417
2418
2419
2420
2421
2422
2423
2424
2425
2426
2427
2428
2429
2430
2431
2432
2433
2434
2435
2436
2437
2438
2439
2440
2441
2442
2443
2444
2445
2446
2447
2448
2449
2450
2451
2452
2453
2454
2455
2456
2457
2458
2459
2460
2461
2462
2463
2464
2465
2466
2467
2468
2469
2470
2471
2472
2473
2474
2475
2476
2477
2478
2479
2480
2481
2482
2483
2484
2485
2486
2487
2488
2489
2490
2491
2492
2493
2494
2495
2496
2497
2498
2499
2500
2501
2502
2503
2504
2505
2506
2507
2508
2509
2510
2511
2512
2513
2514
2515
2516
2517
2518
2519
2520
2521
2522
2523
2524
2525
2526
2527
2528
2529
2530
2531
2532
2533
2534
2535
2536
2537
2538
2539
2540
2541
2542
2543
2544
2545
2546
2547
2548
2549
2550
2551
2552
2553
2554
2555
2556
2557
2558
2559
2560
2561
2562
2563
2564
2565
2566
2567
2568
2569
2570
2571
2572
2573
2574
2575
2576
2577
2578
2579
2580
2581
2582
2583
2584
2585
2586
2587
2588
2589
2590
2591
2592
2593
2594
2595
2596
2597
2598
2599
2600
2601
2602
2603
2604
2605
2606
2607
2608
2609
2610
2611
2612
2613
2614
2615
2616
2617
2618
2619
2620
2621
2622
2623
2624
2625
2626
2627
2628
2629
2630
2631
2632
2633
2634
2635
2636
2637
2638
2639
2640
2641
2642
2643
2644
2645
2646
2647
2648
2649
2650
2651
2652
2653
2654
2655
2656
2657
2658
2659
2660
2661
2662
2663
2664
2665
2666
2667
2668
2669
2670
2671
2672
2673
2674
2675
2676
2677
2678
2679
2680
2681
2682
2683
2684
2685
2686
2687
2688
2689
2690
2691
2692
2693
2694
2695
2696
2697
2698
2699
2700
2701
2702
2703
2704
2705
2706
2707
2708
2709
2710
2711
2712
2713
2714
2715
2716
2717
2718
2719
2720
2721
2722
2723
2724
2725
2726
2727
2728
2729
2730
2731
2732
2733
2734
2735
2736
2737
2738
2739
2740
2741
2742
2743
2744
2745
2746
2747
2748
2749
2750
2751
2752
2753
2754
2755
2756
2757
2758
2759
2760
2761
2762
2763
2764
2765
2766
2767
2768
2769
2770
2771
2772
2773
2774
2775
2776
2777
2778
2779
2780
2781
2782
2783
2784
2785
2786
2787
2788
2789
2790
2791
2792
2793
2794
2795
2796
2797
2798
2799
2800
2801
2802
2803
2804
2805
2806
2807
2808
2809
2810
2811
2812
2813
2814
2815
2816
2817
2818
2819
2820
2821
2822
2823
2824
2825
2826
2827
2828
2829
2830
2831
2832
2833
2834
2835
2836
2837
2838
2839
2840
2841
2842
2843
2844
2845
2846
2847
2848
2849
2850
2851
2852
2853
2854
2855
2856
2857
2858
2859
2860
2861
2862
2863
2864
2865
2866
2867
2868
2869
2870
2871
2872
2873
2874
2875
2876
2877
2878
2879
2880
2881
2882
2883
2884
2885
2886
2887
2888
2889
2890
2891
2892
2893
2894
2895
2896
2897
2898
2899
2900
2901
2902
2903
2904
2905
2906
2907
2908
2909
2910
2911
2912
2913
2914
2915
2916
2917
2918
2919
2920
2921
2922
2923
2924
2925
2926
2927
2928
2929
2930
2931
2932
2933
2934
2935
2936
2937
2938
2939
2940
2941
2942
2943
2944
2945
2946
2947
2948
2949
2950
2951
2952
2953
2954
2955
2956
2957
2958
2959
2960
2961
2962
2963
2964
2965
2966
2967
2968
2969
2970
2971
2972
2973
2974
2975
2976
2977
2978
2979
2980
2981
2982
2983
2984
2985
2986
2987
2988
2989
2990
2991
2992
2993
2994
2995
2996
2997
2998
2999
3000
3001
3002
3003
3004
3005
3006
3007
3008
3009
3010
3011
3012
3013
3014
3015
3016
3017
3018
3019
3020
3021
3022
3023
3024
3025
3026
3027
3028
3029
3030
3031
3032
3033
3034
3035
3036
3037
3038
3039
3040
3041
3042
3043
3044
3045
3046
3047
3048
3049
3050
3051
3052
3053
3054
3055
3056
3057
3058
3059
3060
3061
3062
3063
3064
3065
3066
3067
3068
3069
3070
3071
3072
3073
3074
3075
3076
3077
3078
3079
3080
3081
3082
3083
3084
3085
3086
3087
3088
3089
3090
3091
3092
3093
3094
3095
3096
3097
3098
3099
3100
3101
3102
3103
3104
3105
3106
3107
3108
3109
3110
3111
3112
3113
3114
3115
3116
3117
3118
3119
3120
3121
3122
3123
3124
3125
3126
3127
3128
3129
3130
3131
3132
3133
3134
3135
3136
3137
3138
3139
3140
3141
3142
3143
3144
3145
3146
3147
3148
3149
3150
3151
3152
3153
3154
3155
3156
3157
3158
3159
3160
3161
3162
3163
3164
3165
3166
3167
3168
3169
3170
3171
3172
3173
3174
3175
3176
3177
3178
3179
3180
3181
3182
3183
3184
3185
3186
3187
3188
3189
3190
3191
3192
3193
3194
3195
3196
3197
3198
3199
3200
3201
3202
3203
3204
3205
3206
3207
3208
3209
3210
3211
3212
3213
3214
3215
3216
3217
3218
3219
3220
3221
3222
3223
3224
3225
3226
3227
3228
3229
3230
3231
3232
3233
3234
3235
3236
3237
3238
3239
3240
3241
3242
3243
3244
3245
3246
3247
3248
3249
3250
3251
3252
3253
3254
3255
3256
3257
3258
3259
3260
3261
3262
3263
3264
3265
3266
3267
3268
3269
3270
3271
3272
3273
3274
3275
3276
3277
3278
3279
3280
3281
3282
3283
3284
3285
3286
3287
3288
3289
3290
3291
3292
3293
3294
3295
3296
3297
3298
3299
3300
3301
3302
3303
3304
3305
3306
3307
3308
3309
3310
3311
3312
3313
3314
3315
3316
3317
3318
3319
3320
3321
3322
3323
3324
3325
3326
3327
3328
3329
3330
3331
3332
3333
3334
3335
3336
3337
3338
3339
3340
3341
3342
3343
3344
3345
3346
3347
3348
3349
3350
3351
3352
3353
3354
3355
3356
3357
3358
3359
3360
3361
3362
3363
3364
3365
3366
3367
3368
3369
3370
3371
3372
3373
3374
3375
3376
3377
3378
3379
3380
3381
3382
3383
3384
3385
3386
3387
3388
3389
3390
3391
3392
3393
3394
3395
3396
3397
3398
3399
3400
3401
3402
3403
3404
3405
3406
3407
3408
3409
3410
3411
3412
3413
3414
3415
3416
3417
3418
3419
3420
3421
3422
3423
3424
3425
3426
3427
3428
3429
3430
3431
3432
3433
3434
3435
3436
3437
3438
3439
3440
3441
3442
3443
3444
3445
3446
3447
3448
3449
3450
3451
3452
3453
3454
3455
3456
3457
3458
3459
3460
3461
3462
3463
3464
3465
3466
3467
3468
3469
3470
3471
3472
3473
3474
3475
3476
3477
3478
3479
3480
3481
3482
3483
3484
3485
3486
3487
3488
3489
3490
3491
3492
3493
3494
3495
3496
3497
3498
3499
3500
3501
3502
3503
3504
3505
3506
3507
3508
3509
3510
3511
3512
3513
3514
3515
3516
3517
3518
3519
3520
3521
3522
3523
3524
3525
3526
3527
3528
3529
3530
3531
3532
3533
3534
3535
3536
3537
3538
3539
3540
3541
3542
3543
3544
3545
3546
3547
3548
3549
3550
3551
3552
3553
3554
3555
3556
3557
3558
3559
3560
3561
3562
3563
3564
3565
3566
3567
3568
3569
3570
3571
3572
3573
3574
3575
3576
3577
3578
3579
3580
3581
3582
3583
3584
3585
3586
3587
3588
3589
3590
3591
3592
3593
3594
3595
3596
3597
3598
3599
3600
3601
3602
3603
3604
3605
3606
3607
3608
3609
3610
3611
3612
3613
3614
3615
3616
3617
3618
3619
3620
3621
3622
3623
3624
3625
3626
3627
3628
3629
3630
3631
3632
3633
3634
3635
3636
3637
3638
3639
3640
3641
3642
3643
3644
3645
3646
3647
3648
3649
3650
3651
3652
3653
3654
3655
3656
3657
3658
3659
3660
3661
3662
3663
3664
3665
3666
3667
3668
3669
3670
3671
3672
3673
3674
3675
3676
3677
3678
3679
3680
3681
3682
3683
3684
3685
3686
3687
3688
3689
3690
3691
3692
3693
3694
3695
3696
3697
3698
3699
3700
3701
3702
3703
3704
3705
3706
3707
3708
3709
3710
3711
3712
3713
3714
3715
3716
3717
3718
3719
3720
3721
3722
3723
3724
3725
3726
3727
3728
3729
3730
3731
3732
3733

```