OPTIMSYN: INFLUENCE-GUIDED RUBRICS OPTI-MIZATION FOR SYNTHETIC DATA GENERATION

Anonymous authors

000

001

002003004

010 011

012

013

014

015

016

017

018

019

021

024

025

026

027

028

029

031

034

037 038

039

040

041

042

043

044

046 047

048

051

052

Paper under double-blind review

ABSTRACT

Large language models (LLMs) achieve strong downstream performance largely due to abundant supervised fine-tuning (SFT) data that imparts problem-solving capabilities. However, as applications expand, high-quality SFT data in knowledge-intensive domains (e.g., humanities and social sciences, medicine, law, finance) is exceedingly scarce: expert curation is costly, privacy constraints are strict, and label consistency is hard to guarantee. Recent work turns to synthetic data, typically prompting a generator over domain documents and filtering with handcrafted rubrics. Yet, rubric design is expert-dependent and rarely transfers across domains; moreover, prevalent heuristic optimization follows a brittle loop (write rubric \rightarrow synthesize \rightarrow train \rightarrow inspect \rightarrow guess tweaks) that lacks reliable, quantitative feedback about a rubric's true contribution to downstream performance. We argue for assessing synthetic data quality through its training utility on the target model, using this feedback to guide data generation. Inspired by classic influence estimations, we repurpose an optimizer-aware estimator that uses gradient information to quantify each synthetic sample's contribution to the objective of a given target model on specific tasks. Our analysis reveals a gap: although synthetic and real samples may be close in embedding space, their *influence* on learning can differ substantially. Building on this insight, we propose an optimizationbased synthetic data framework that adapts rubrics with target-model feedback. Instead of manually engineering domain rubrics, we supply lightweight guiding text and delegate rubric generation to a rubric-specialized model conditioned on the task; crucially, we employ influence score as reward and optimize the rubric generator with reinforcement learning. Empirically, the framework yields consistent gains across domains (HSS and health), target models (e.g., Qwen and Llama families), and data generators, demonstrating broad generalization and engineering portability without task-specific tuning.

1 Introduction

Large language models (LLMs) now excel across a wide range of downstream tasks (Achiam et al., 2023; Comanici et al., 2025; Guo et al., 2025), due in no small part to training on vast and diverse corpora. In particular, supervised fine-tuning (SFT) has endowed LLMs with strong instruction-following and problem-solving capabilities (Peng et al., 2023; Maeng et al., 2017). However, as LLMs are deployed in increasingly specialized scenarios, *high-quality*, *real* SFT data has become acutely scarce—especially in knowledge-intensive verticals such as the humanities and social sciences, medicine, law, and finance. The barriers are both practical and structural: domain expertise is costly and limited, privacy constraints are stringent, and large-scale manual annotation is difficult to standardize, expensive to procure, and hard to keep consistent.

Synthetic data has recently emerged as a promising way to alleviate this bottleneck (Wang et al., 2022; Maeng et al., 2017). Most current pipelines bootstrap from raw materials (e.g., documents), query a teacher model to produce question–answer pairs, and then filter (or steer) generations with pre-defined *rubrics* (rules or prompts) (Penedo et al., 2024; Jiang et al., 2025; Yue et al., 2024). Yet this paradigm faces two fundamental limitations. (i) Limited transferability. Rubric design is largely expert-driven and domain-specific; rubrics that work in one field often fail to generalize to another, undermining the universality of synthesis strategies. (ii) Heuristic optimization. The common loop—"handcraft rubrics \rightarrow synthesize data \rightarrow train model \rightarrow inspect outcomes \rightarrow propose

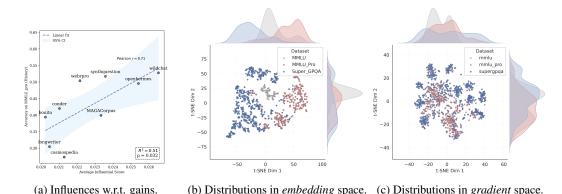


Figure 1: Visualization of influence score as an effective feature for assessing sample utility. (a) Relationship between validation influence scores and training performance. (b) Distribution of three validation sets in the embedding space of Qwen3-8B-Base. (c) Distribution of three validation sets

revisions"—relies on tacit experience and provides little quantitative feedback. Human observers cannot reliably attribute downstream performance changes to specific rubric choices, making the process slow, brittle, and uncertain.

in the loss-gradient embedding space, showing closer alignment and shared optimization directions.

To address these limitations, we propose to directly quantify the *training utility* of each synthetic example for a given target model and task using a gradient-based influence estimator (Xia et al., 2024). Concretely, we approximate the contribution of a candidate QA pair to the validation objective under the same update rule used during fine-tuning (Adam), yielding an influence signal that is aligned with the actual optimization dynamics. Preliminary analyses (Figure 1) reveal a marked gap between representation proximity and training impact: synthetic and real samples can be close in embedding space yet diverge substantially in their estimated influence, which helps explain why seemingly on-distribution samples may still underperform during SFT. Moreover, dataset-level influence aggregates exhibit a strong positive correlation with held-out accuracy, validating influence as a reliable proxy for synthetic data quality. These observations motivate replacing heuristic rubric engineering with model-impact supervision grounded in gradients.

Building on this insight, we introduce a prompter (the policy model) optimization framework that treats rubric construction as a learnable component driven by target-model feedback. Starting from minimal guiding text, a rubric generator proposes seed document-specific rubrics; a generator model, conditioned on the seed document and rubric, synthesizes a QA pair; and the target model supplies a verifiable reward that combines lightweight validity checks with the optimizer-aware influence score. We then update the policy using GRPO algorithm (Guo et al., 2025), thereby closing the synthesis-training loop and explicitly maximizing expected downstream improvement. Empirically, this framework yields consistent gains across knowledge-intensive domains (the humanities and social sciences, and medicine and health), transfers across target model families and scales, and is robust to the choice of data generator, turning rubric design from brittle, expert-crafted heuristics into a portable, model-aligned optimization problem. The contributions of this paper are threefold:

- 1. **Optimizer-aware influence estimation.** We adapt classical influence-estimation ideas to modern synthetic-data pipelines and derive a practical, optimizer-aligned estimator of each synthetic sample's training utility.
- 2. **From heuristics to supervision.** We replace heuristic rubric engineering with a model-impact objective by training a *rubric generator* to maximize estimated downstream benefit (compatible with RL or gradient-based updates), yielding more effective synthetic data.
- 3. **Transferable and efficient across domains.** Empirically, the learned rubrics are *model-and task-conditioned*, reducing reliance on domain expertise and enabling deployment in privacy-constrained, knowledge-intensive fields where manual curation is infeasible.

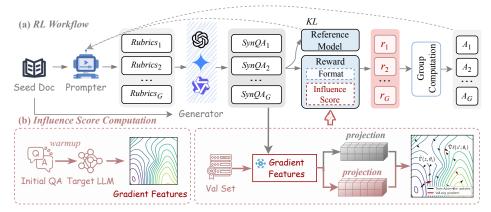


Figure 2: The proposed RL framework for synthetic data generation with optimizer-aware rewards. (a) RL workflow. Starting from a seed document, a *prompter* instantiates G seed-specific rubrics, and a *generator* produces synthetic QA pair. (b) Reward computation. We employ the influence score for each synthetic QA pair as reward, followed by policy update.

2 METHOD

2.1 OVERVIEW

Given a set of seed data $\mathcal{S}=\{S_i\}_{i=1}^N$, the goal of data synthesis is to construct a dataset containing synthetic question—answer pairs $\{(Q_i,A_i)\}_{i=1}^N$. Existing methods typically instruct a teacher model—often a strong external model (e.g., an API)—with carefully designed rubrics to generate such data. However, rubric-based synthesis is constrained by human intuition, highly domain-specific, and thus poorly transferable across tasks and models. To overcome these limitations, our framework employs a dedicated rubric generator that produces a targeted rubric B_i for each seed datum S_i and a specified target model. We then condition the teacher model on (B_i,S_i) to generate a synthetic pair (Q_i,A_i) . In the following, we detail how we quantify the training utility of a synthetic pair (Q_i,A_i) using optimizer-aware influence estimates and how this feedback drives the learning of the rubric generator.

2.2 ESTIMATING THE INFLUENCE OF SYNTHETIC DATA

While human-crafted rubrics appear reasonable, there is no guarantee that they actually lead to effective dataset generation. A promising yet underexplored alternative is to adopt a more principled generation process that does not rely solely on human intuition. Influence functions (Pruthi et al., 2020), which approximate training dynamics via first-order analysis, have been widely used to estimate the impact of individual training examples on held-out performance. Prior work (Xia et al., 2024) has shown their effectiveness in selecting data that benefits downstream tasks. This naturally motivates our use of influence functions as a principled guide for dataset generation.

Preliminaries. The classical influence function (Koh & Liang, 2017) quantifies how individual training points affect a model's parameters and, consequently, its predictions. TracIn (Pruthi et al., 2020) pursues the same goal with a scalable, first-order, trajectory-based estimator: it accumulates gradient inner products of training checkpoints to approximate influence. Because it requires only per-example gradients, learning rates, and saved checkpoints, TracIn is practical at LLM scale. Since contemporary LLMs are typically optimized with Adam (Adam et al., 2014), we adopt an Adam-compatible variant (Xia et al., 2024). Given a training sample z, its influence on an evaluation sample z^{\prime} can be represented as:

$$\operatorname{Inf}_{\operatorname{Adam}}(z, z') = \sum_{i=1}^{T} \bar{\eta}_{i} \cos(\nabla_{\theta} \ell(z'; \theta_{i}), \Gamma(z, \theta_{i})), \tag{1}$$

where $\bar{\eta}_i$ is the average learning rate in epoch i (out of T total epochs) and θ_i is the checkpoint after epoch i. Γ requires the moment statistics $(\boldsymbol{m}, \boldsymbol{v})$, which depend on past gradients, with details provided in Appendix B.

181 182

183

184

185

186

187

188

189

190 191

192

193

194

195

196

197

198 199

200201

202

203

204

205

206

207

208

209210

211

212

213214

215

Algorithm 1 Influence-Guided Data Synthesis Framework

```
163
              Require: Seed data set S, Generator model T, Target model M, Policy model \pi_{\theta}, Validation set \mathcal{D}_{\text{val}}, Number
164
                    of updates K, Rollout group size G, Reference model \pi_{ref} \leftarrow \pi_{\theta};
                1: for k = 1 to K do
                           Sample a mini-batch of seed data \{S_i\}_{i=1}^B \sim \mathcal{X}.
166
               2:
                           for each seed data S_i \in \mathcal{S} do
                3:
167
                                Generate G rubrics B_{i,j} \sim \pi_{\theta}(\cdot|S_i) for j = 1..G
                4:
168
                5:
                                for each rubric B_{i,j} do
169
                                      Obtain (Q_{i,j}, A_{i,j}) \leftarrow T(S_i, B_{i,j})
                6:
170
                7:
                                      Estimate influence score \operatorname{IF}_{Adam}(Q,A) according to Eq. 1
                                      Compute trajectory reward R(\tau) = \operatorname{Valid}(Q, A) \cdot \operatorname{IF}(Q, A) - \lambda (1 - \operatorname{Valid}(Q, A))
171
                8:
               9:
172
              10:
                                Aggregate rewards and compute group-normalized advantages:
173
                                \hat{A}_{i,t} = \left(R(\tau_i) - \frac{1}{G}\sum_{j=1}^G R(\tau_j)\right)/\sqrt{\frac{1}{G}\sum_{j=1}^G \left(R(\tau_j) - \frac{1}{G}\sum_{k=1}^G R(\tau_k)\right)^2 + \delta} Aggregate loss over tokens:
174
              11:
              12:
175
                                Aggregate toss over toens. L(\theta) = -\frac{1}{G} \sum_{i=1}^{G} \frac{1}{|\tau_i|} \sum_{t=1}^{|\tau_i|} \min \left[ r_{i,t}(\theta) \, \hat{A}_{i,t}, \, \operatorname{clip}(r_{i,t}(\theta)) \, \hat{A}_{i,t} \right] - \beta \operatorname{KL}(\pi_\theta \parallel \pi_{\text{ref}}) \right]
176
              13:
177
              14:
                                Update policy \pi_{\theta}
                           end for
              15:
178
              16: end for
179
              17: return Optimized prompter \pi_{\theta}^*
```

Why gradients, not embeddings? Empirically (Fig. 1b and 1c), we observe two phenomena: (i) synthetic sets whose gradient distributions are closer to the validation distribution yield better downstream performance; (ii) the same trend does *not* hold in embedding space, where proximity often fails to predict gains. This explains why aesthetically "high-quality" samples may not train well: they align semantically but steer optimization in suboptimal directions. Let IF(Q, A) denote the Adam-compatible influence score (Eq. 1) of the synthetic pair (Q, A) with respect to the validation set. We further quantify this by correlating IF averages with held-out accuracy (Fig. 1a): IF exhibits a strong positive correlation, validating it as a reliable proxy for synthetic data quality.

Takeaway. Influence functions provide a principled, model-aware signal for judging the *training utility* of synthetic examples, turning data synthesis from heuristic rubric engineering into *model-centric* optimization. Building on this, we close the loop with an RL framework in which a *rubric generator* acts as the agent (replacing ad-hoc human design), explores the rubric space, and receives influence-based rewards that estimate each synthesized sample's contribution to downstream improvement. This alignment of generation with measured model impact yields more useful data with less domain-specific handcrafting, better sample efficiency, and a clear path to automatic refinement across tasks and domains.

2.3 Influence-Guided Reinforcement Learning

We train the *rubric generator* as a policy that explores the rubric space and receives *influence-based* feedback from the target model. Formally, given a seed datum S, the policy LLM π_{θ} produces a rubric $B \sim \pi_{\theta}(\cdot \mid S)$; a teacher model then conditions on (S, B) to synthesize a pair (Q, A); the target model supplies a scalar reward that scores the *training utility* of (Q, A). A trajectory is $\tau = \{S, B, (Q, A)\}$, and the objective is to maximize $\mathbb{E}_{\tau \sim \pi_{\theta}}[R(\tau)]$.

Verifiable, influence-based rewards. We combine automatic validity checks with an optimizer-aware influence estimate. Let $\mathrm{Valid}(Q,A) \in \{0,1\}$ be a conjunction of lightweight verifiers (formatting, non-triviality, safety). The reward is:

$$R(\tau) = \operatorname{Valid}(Q, A) \cdot \operatorname{IF}(Q, A) - \lambda (1 - \operatorname{Valid}(Q, A)), \tag{2}$$

where $\lambda > 0$ penalizes invalid generations. This yields a *verifiable* signal that is aligned with downstream improvement while discouraging degenerate outputs.

Policy optimization. We use a clipped policy-gradient objective (GRPO/PPO-style) with group-relative baselines for variance reduction. For each seed S we sample G rubrics, producing trajecto-

ries $\{\tau_i\}_{i=1}^G$. Denote by $\tau_{i,(t)}$ the t-th token and by $\tau_{i,< t}$ its prefix. The objective is

$$J(\theta) = \frac{1}{G} \sum_{i=1}^{G} \frac{1}{|\tau_{i}|} \sum_{t=1}^{|\tau_{i}|} \min \left[r_{i,t}(\theta) \, \hat{A}_{i,t}, \, \operatorname{clip}(r_{i,t}(\theta), \, 1 - \epsilon, \, 1 + \epsilon) \, \hat{A}_{i,t} \right] - \beta \, \operatorname{KL}(\pi_{\theta} \parallel \pi_{\operatorname{ref}}), (3)$$

with importance ratio $r_{i,t}(\theta) = \pi_{\theta}(\tau_{i,(t)} \mid \tau_{i,< t}) / \pi_{\text{old}}(\tau_{i,(t)} \mid \tau_{i,< t})$, and a group-normalized advantage $\hat{A}_{i,t} = \left(R(\tau_i) - \frac{1}{G}\sum_{j=1}^G R(\tau_j)\right) / \sqrt{\frac{1}{G}\sum_{j=1}^G \left(R(\tau_j) - \frac{1}{G}\sum_{k=1}^G R(\tau_k)\right)^2 + \delta}$, where ϵ is the clipping parameter, β controls KL trust region against a reference policy π_{ref} (e.g., the SFT model), and $\delta > 0$ stabilizes normalization. We add entropy regularization to encourage exploration.

This RL loop closes the synthesis-training feedback cycle: the policy learns rubrics that systematically *maximize measured training impact*, yielding synthetic data that is not only plausible by heuristic criteria but empirically *helpful for learning*.

3 EXPERIMENT

3.1 EXPERIMENTAL SETUP

Models The *teacher LLM* (also denoted as generator), unless otherwise specified, is *Qwen3-235B-Instruct* (Yang et al., 2025), which generates synthetic supervised fine-tuning (SFT) data; in ablation studies, we substitute it with *GPT-4.1* (Achiam et al., 2023) and *Gemini-2.5-Pro* (Comanici et al., 2025) to evaluate the sensitivity of our pipeline to the teacher model choice. The *target LLM* is by default *Qwen3-8B-Base*, trained on the synthesized data; we further examine in two settings: (i) replacing the target with *Llama3-8B-Base* (Dubey et al., 2024) and (ii) Qwen variants—*Qwen3-4B-Base*, *Qwen3-8B-Base*, and *Qwen3-14B-Base*—using the *same* learned prompter. Finally, the *prompter base model* is initialized from *Qwen3-8B-Instruct*, which serves as the base model for rubric optimization throughout training.

Training Dataset We apply our SFT data–synthesis framework to two *Humanities and Social Sciences* and *Medical and Health*. For HSS, we curate seed documents from books: specifically, we select 223 humanities and social sciences books with high public ratings on *Goodreads*, with the domain distribution shown in Fig. 3 (left). Texts are segmented by table-of-contents chapters, and we remove boilerplate, front/back matter, and other non-textual or noisy content. For the medical domain, we draw seed documents from the pretraining corpora released with the Meditron large language model (Chen et al., 2023) and from a subset of open-access PubMed abstracts; the distribution is shown in Fig. 3 (right). Because rubric generation requires feeding each seed document to the prompter



Figure 3: Domain distribution of seed documents used to synthesize questionanswering pairs: *Humanities and Social Sciences* and *Medical and Health*.

model and is constrained by the model's input window, we retain only documents shorter than 8k tokens. After filtering, the HSS and medical corpora contain 18,665 and 26,435 seed documents, respectively. Further details of the raw sources are provided in Appendix D.

Evaluation Benchmarks We evaluate models on 12 benchmarks in two domains. Because the humanities and social sciences (HSS) lack dedicated evaluation suites, we extract HSS-related subsets from three comprehensive datasets: MMLU-pro (*History*) (Wang et al., 2024), SuperGPQA (*History*) (Du et al., 2025), and Humanity's Last Exam (Phan et al., 2025) (*Humanities / Social Science*). We include Big Bench Hard (Kazemi et al., 2025), HellaSwag (Zellers et al., 2019), and DROP to gauge performance on complex comprehension and commonsense reasoning. For the medical domain, we assess capabilities using MMLU-pro (*Health*), SuperGPQA (*Health*), Humanity's Last Exam (*Biology / Medicine*), HealthBench (*Consensus*) (Arora et al., 2025), PubMedQA (Jin et al., 2019), and MedQA (Koh & Liang, 2017). Our evaluation adopts a 5-shot for MMLU-pro and a one-shot for PubMedQA and MedQA, with answer Accuracy as the primary metric. For benchmarks HLE and HealthBench that require an LLM-as-Judge, we use GPT-4.1 as the judge.

Table 1: Results across two domains. All metrics are accuracy (higher is better). Across both HSS and Medical domains, our method consistently upgrades Qwen3-8B-Base beyond widely used open SFT corpora, often rivaling or surpassing Qwen3-8B-Instruct.

Humanities and Social Sciences (HSS)						
Model	MMLU pro	Super GPQA	HLE	ввн	HellaSwag	DROP
Qwen3-8B-Base	22.83	20.77	5.70	35.62	35.71	45.73
Qwen3-8B-Instruct	49.87	23.44	4.66	85.21	<u>78.71</u>	80.73
MAGACorpus	39.90	21.18	8.29	58.21	72.88	62.53
Bonito	39.37	21.36	5.70	45.78	83.00	50.95
Conder	41.99	23.29	6.22	74.21	59.32	41.52
Cosmospedia	27.30	20.03	5.18	55.63	79.33	49.70
Longwriter	30.45	12.02	6.22	34.89	65.09	23.62
Openhermes	49.56	24.60	3.63	73.64	73.41	61.62
Synthquestion	51.71	23.99	5.18	74.24	80.20	61.88
WebrPro	50.39	22.85	7.77	68.32	77.60	58.77
Wildchat	<u>52.76</u>	22.11	6.74	<u>76.32</u>	73.41	54.31
Ours	56.96	26.07	7.85	75.65	78.08	65.42
Medical and Health						
Model	MMLU pro	Super GPQA	HLE	Health Bench	PubMed	MedQA
Qwen3-8B-Base	45.97	28.06	10.36	70.06	65.90	51.45
Qwen3-8B-Instruct	60.39	37.16	7.21	87.70	65.70	57.09
ChatDoctor	16.48	21.22	6.31	_	85.40	21.91
MedQA	13.45	34.61	9.91	30.46	71.70	53.41
Medical-o1	60.51	27.26	8.11	70.14	73.90	58.75
Medical-R1-Distill	57.33	35.28	4.95	80.38	73.40	<u>57.81</u>
ReasonMed	55.00	28.01	14.41	51.76	79.80	23.56
Ours	56.97	38.28	10.81	74.82	80.70	58.75

Baselines In the *humanities and social sciences*, our baselines comprise Qwen3-8B-Instruct (without additional fine-tuning) and models fine-tuned on eight SFT datasets. The datasets span three categories: (i) human-authored—WildChat (Zhao et al., 2024) and OpenHermes (Teknium, 2023); (ii) semi-automatically synthesized—MGACorpus (Hao et al., 2025), WebR Pro (Jiang et al., 2025), Bonito (Nayak et al., 2024), and SynthQuestions (Zhu et al., 2025); and (iii) fully automated synthetic—Conder (Cao et al., 2025). In the *medical and health* domain, our baselines include Qwen3-8B-Instruct and models fine-tuned on five SFT datasets: the real-world clinical dialogue corpus ChatDoctor (Li et al., 2023), the multiple-choice dataset MedQA (Singhal et al., 2025), and three LLM-distilled medical reasoning datasets—Medical-o1 (Chen et al., 2024), Medical-R1-Distill (Chen et al., 2024), and ReasonMed (Sun et al., 2025).

Implementation Details For influence estimation, we first warm up the target model using 10% of the synthetic data generated by the initialized prompter and generator. The resulting model provides reference gradients for computing influence scores, which are subsequently incorporated into the RL stage. For reinforcement learning, we adopt GRPO (Guo et al., 2025) as the optimization algorithm, where the reward function integrates the influence score with a format-consistency component. Training is performed with a batch size of 256, learning rate of 1×10^{-6} , rollout temperature 1.5, rollout size n=5, and one epoch of updates. λ is set to 0.1. Processing approximately 20K samples takes about 10 hours. All experiments are conducted on $8\times H200$ GPUs. Additional implementation details are provided in the Appendix D.

3.2 MAIN RESULTS

Superiority over Human-Annotation and Synthetic-Data Baselines. We present results for the two domains in Table 1. Our rubric-guided synthesis consistently enhances the target base model across both Humanities & Social Sciences and Medical domains, surpassing strong open SFT baselines and its instruct model. On HSS tasks, it rivals or even exceeds deliberate-reasoning teachers,

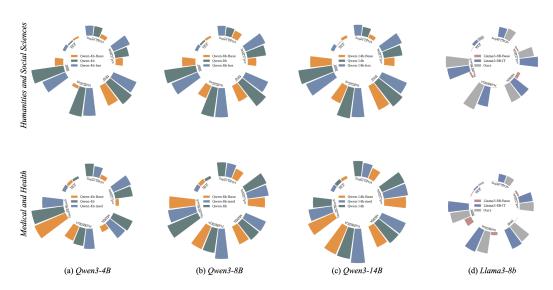


Figure 4: The performance of our framework across different target model scales and families.

achieving a +27.2% relative gain on HLE (0.0785 vs. 0.0570), thereby demonstrating that structured, group-aware data synthesis can effectively distill reasoning ability *without* test-time reasoning. In the medical domain, our method narrows the gap on MMLU_PRO and SUPER_GPQA while decisively outperforming baselines on HLE and HEALTHBENCH. The most significant improvements appear on tasks aligned with domain rubrics, underscoring the strength of our prompter-centric RL framework in producing high-quality, domain-grounded supervision. Overall, the approach consistently upgrades the same 8B backbone beyond widely used SFT corpora, often matching or surpassing inference-time "thinking" models on reasoning-centric metrics.

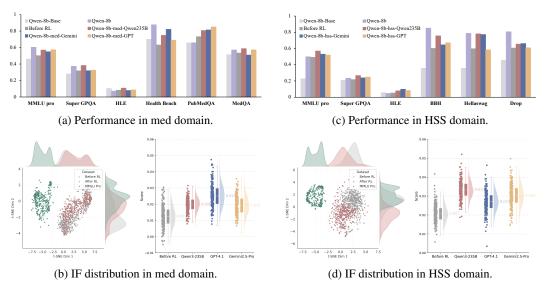
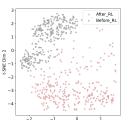


Figure 5: The performance of our framework across different generators.

Generalization across Target Model Scales and Families. To test whether our training recipe is target-agnostic, we replace the backbone with *Qwen3*- $\{4B, 8B, 14B\}$ and *Llama3*-8B and evaluate on two domain suites—*Humanities & Social Sciences* and *Medical & Health*, with results shown in Figure 4. Across all four backbones, the plots exhibit consistent improvements for **Ours** than for the *Base* and *Instruct* variants on most tasks within both domain suites. The gains persist when moving from small to larger models (4B \rightarrow 14B), while remaining substantial at the smallest scale,



379

385

386

387

388 389

390 391

392

393

394

395

396 397

398

399

400

401

402 403

404

405

406

407

408

409

410

411 412

413

414

415

416

417

418

419

420

421

422

423

424

425 426

427

428

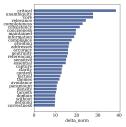
429

430

431







- (a) Distributions of the embeddings of rubrics.
- without our method.
- with our method.
- (b) Word cloud of rubrics (c) Word cloud of rubrics (d) New high-frequency terms in Figure 6c.

Figure 6: Effects of the Proposed Method on Rubrics

suggesting that our method does not rely on model capacity to be effective. Moreover, the improvements transfer from the Qwen3 family to Llama3 with comparable margins, despite architectural and pretraining differences. Taken together, these results demonstrate that the proposed method confers family- and scale-robust benefits, supporting the claim that our method yields broadly applicable performance gains for diverse target LLMs.

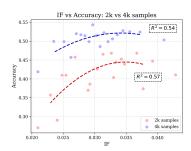
Robustness to the Choice of Generator. We ablate the synthetic-data *generator* by swapping Qwen3–235B, GPT-4.1, and Gemini-2.5-Pro while keeping the target model fixed. The results are provided in Figure 5a and 5c. Across both domains-Medical & Health and Humanities & Social Sciences—models trained with our method consistently outperform the base model on nearly all benchmarks, indicating that the gains are not tied to any single generator.

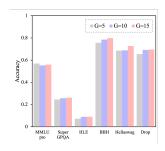
3.3 Analysis

Consistent Enhancement of the Influence Distribution in Synthetic Data. We further visualize the distribution of influence scores produced by our method, as shown in Figure 5d and Figure 5b. The left panel shows that our approach aligns the distribution of synthetic data influence scores more closely with that of the validation data. The right panel demonstrates that, even when substituting different generators, our method consistently shifts the distribution toward higher average influence scores. Taken together, these results indicate that the proposed reward signal effectively selects and amplifies high-utility synthetic examples in a generator-agnostic fashion.

Effects of the Proposed Method on Rubrics. We examine how the prompter-centric RL reshapes the learned rubrics. Figure 6a visualizes t-SNE embeddings of rubric texts in the generator space before and after RL: post-RL rubrics occupy a broader, more structured region, whereas pre-RL rubrics cluster in a narrow band, indicating increased coverage and diversity aligned with the seed data distribution. From our case analyses, we observe that the post-RL rubrics are more tightly grounded in the source documents, which in turn explains their broader coverage and clearer structure in the embedding space. We hypothesize that such more detailed rubrics better steer the generator to synthesize data that conforms to the specification and task requirements. The word clouds in Figure 6b-6c reveal a qualitative shift from generic imperatives (e.g., focus, align, short) to domainand quality-oriented attributes (e.g., critical, clarity, completeness, parsability, information density, logical soundness). Figure 6d further quantifies these changes by highlighting newly introduced high-frequency terms and their normalized frequency deltas, confirming that our method promotes specific, actionable criteria rather than vague instructions. Detailed case studies of the rubrics and question-answer pairs are provided in Appendix 5 and 6.

Predictive Relationship Between Influence Scores and Downstream Accuracy. We investigate whether per-example Influence Scores (IF), computed on the synthetic training pool, are predictive of downstream supervised fine-tuning (SFT) performance. Specifically, we repeatedly sample random subsets of size 2K or 4K from the generated synthetic dataset, compute the aggregate IF of each subset, and fine-tune the model exclusively on that subset. Figure 7 shows that higher-IF subsets consistently yield higher test accuracy under both training budgets. Quadratic regressions (dashed curves) summarize the trend with strong goodness of fit ($R^2 = 0.57$ for 2K and $R^2 = 0.54$





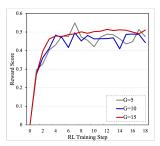


Figure 7: Influence score vs. accuracy.

Figure 8: Accuracy across different group sizes.

Figure 9: Training curves across different group sizes.

for 4K), revealing a clear positive correlation with mild saturation at the high-IF end. These results demonstrate that IF serves as a reliable proxy for the utility of synthetic examples, and that prioritizing higher-IF data can improve downstream accuracy even with limited training budgets.

Effects of the Rubrics on Synthetic Data. We ablate the number of rubric rollouts per seed, G, to assess how exploration affects synthetic data quality. Figure 8 shows RL training curves for $G! \in !5, 10, 15$: larger G reaches higher reward and exhibits reduced variance, suggesting that broader rubric exploration stabilizes IF-driven optimization. Figure 9 reports downstream accuracy on multiple benchmarks: increasing N consistently improves performance. Together, these results support that diverse, targeted rubrics—enabled by higher rollout counts—both stabilize training and amplify the utility of synthetic data.

4 RELATED WORKS

Synthetic data is increasingly used to reduce annotation costs for instruction tuning. Document-to-SFT methods convert readily available corpora into dialogue pairs: WebR (Jiang et al., 2025) formulates webpages as dual-view reconstruction; MAmmoTH2 (Yue et al., 2024) scales via web recall to refinement; Bonito (Nayak et al., 2024) maps unlabeled text and task attributes to instruction–response via meta-template conditional generators. A complementary line iteratively improves synthesized data: Evol-Instruct (Xu et al., 2024) evolves seeds to raise difficulty; Condor (Cao et al., 2025) builds a World Knowledge Tree with self-reflection refinement; Li et al. (2024a) use failure-inducing exploration; Montessori-Instruct (Li et al., 2024b) contrasts beneficial vs. non-beneficial samples with DPO. Compared to these, our approach replaces heuristic filters with a training-signal—aligned objective, learns rubrics from model feedback rather than expert priors (enabling cross-domain transfer), and uses influence-based analysis to expose the mismatch between embedding similarity and training influence. Elevating the target model's own training signal turns rubric design into a learnable component and provides a principled route to scalable SFT in data-scarce, high-stakes domains.

5 CONCLUSION

We introduced a synthesis-training framework that replaces heuristic rubric design with *model-impact supervision*. At its core is an optimizer-aware, gradient-based influence estimator that quantifies each synthetic QA pair's contribution to the target model's objective, revealing a marked mismatch between embedding similarity and training impact and motivating gradient-aligned selection. Leveraging this signal, we cast rubric construction as a learnable policy within a prompter-centric RL loop, optimized with GRPO objectives under verifiable rewards that blend lightweight validity checks with influence scores. This closes the synthesis-training feedback cycle and grounds rubric learning in downstream impact rather than surface heuristics. Empirically, our approach consistently improves performance across domains, model families, and generators, rivaling or surpassing standard SFT baselines. Influence scores correlate strongly with realized accuracy under fixed budgets, validating them as a reliable proxy for synthetic data quality and highlighting the portability of our prompter-centric optimization framework.

ETHICS STATEMENT

All authors attest that they have read and will abide by the ICLR Code of Ethics. Our study develops an influence-aligned, prompter-centric framework for rubric-guided synthetic data generation and evaluates it on knowledge-intensive domains. Below we summarize ethical considerations specific to this work.

Human subjects and IRB. This work does not involve human participants, user studies, or the collection of personal data. No IRB approval was required.

Data sources, licensing, and privacy. Seed documents are drawn from publicly available sources in the Humanities Social Sciences and Medical domains; preprocessing removes boilerplate and noisy content, and the generator is instructed to paraphrase rather than reproduce lengthy excerpts verbatim. We release only prompts/rubrics and synthesized QA pairs, not the original documents. We screen out personally identifiable information (PII) and sensitive attributes to the best of our ability and respect the licenses and terms of use of all sources. We will honor takedown requests for inadvertently included material.

Safety, misuse, and domain constraints. Although our method targets data quality, synthetic outputs in medical or other high-stakes settings may be misinterpreted as professional advice. Our models and datasets are research artifacts and *not* intended for deployment in clinical, legal, financial, or safety-critical contexts. The rubric pipeline includes format/safety checks (e.g., refusal of unsafe instructions) and excludes harmful topics during synthesis. We recommend downstream users apply additional content filters and human review.

Use of external APIs and compliance. When large closed models are used as generators or judges, all interactions comply with provider terms of service; no user data or proprietary content is transmitted beyond the required prompts. We report which systems are used and how (teacher, judge) in the paper.

Conflicts of interest and sponsorship. The authors declare no conflicts of interest related to this submission. No external sponsor influenced the problem formulation, experiments, or reporting.

Limitations. Influence estimates depend on the chosen validation distribution and require a short warmup; these choices may bias rewards and entail extra compute. We discuss these limitations and propose mitigations and future directions in the paper.

REPRODUCTION CHECKLIST

1. General Paper Structure

- 1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) yes
- 1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) yes
- 1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) yes

2. Dataset Usage

- 2.1. Does this paper rely on one or more datasets? (yes/no) yes
- If yes, please address the following points:
 - 2.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) yes
 - 2.3. All novel datasets introduced in this paper are included in a data appendix (yes/par-tial/no/NA) yes

- 2.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA) yes
- 2.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) yes
- 2.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) yes
- 2.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing (yes/partial/no/NA) yes

3. Computational Experiments

- 3.1. Does this paper include computational experiments? (yes/no) yes
- If yes, please address the following points:
 - 3.2. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA) yes
 - 3.3. Any code required for pre-processing data is included in the appendix (yes/partial/no) ves
 - 3.4. All source code required for conducting and analyzing the experiments is included in a code appendix (yes/partial/no) yes
 - 3.5. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no) yes
 - 3.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no) yes
 - 3.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/partial/no/NA) yes
 - 3.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no) yes
 - 3.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no) yes
 - 3.10. This paper states the number of algorithm runs used to compute each reported result (yes/no) yes
 - 3.11. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no) yes
 - 3.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no) yes
 - 3.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA) yes

REFERENCES

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.
- Kingma DP Ba J Adam et al. A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 1412(6), 2014.
- Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.
- Maosong Cao, Taolin Zhang, Mo Li, Chuyu Zhang, Yunxin Liu, Haodong Duan, Songyang Zhang, and Kai Chen. Condor: Enhance llm alignment with knowledge-driven data synthesis and refinement. *arXiv preprint arXiv:2501.12273*, 2025.
- Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-o1, towards medical complex reasoning with llms, 2024. URL https://arxiv.org/abs/2412.18925.
- Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. Meditron-70b: Scaling medical pretraining for large language models. *arXiv* preprint *arXiv*:2311.16079, 2023.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv* preprint arXiv:2507.06261, 2025.
- Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, et al. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *arXiv preprint arXiv:2502.14739*, 2025.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.
- Yuxian Gu, Li Dong, Hongning Wang, Yaru Hao, Qingxiu Dong, Furu Wei, and Minlie Huang. Data selection via optimal control for language models. *arXiv preprint arXiv:2410.07064*, 2024.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- Xintong Hao, Ke Shen, and Chenggang Li. Maga: Massive genre-audience reformulation to pre-training corpus expansion. *arXiv e-prints*, pp. arXiv–2502, 2025.
- Yuxin Jiang, Yufei Wang, Chuhan Wu, Xinyi Dai, Yan Xu, Weinan Gan, Yasheng Wang, Xin Jiang, Lifeng Shang, Ruiming Tang, et al. Instruction-tuning data synthesis from scratch via web reconstruction. *arXiv preprint arXiv:2504.15573*, 2025.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, 2019.
- Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, Sanket Vaibhav Mehta, Lalit K Jain, Virginia Aglietti, Disha Jindal, Peter Chen, et al. Big-bench extra hard. *arXiv preprint arXiv:2502.19187*, 2025.
- Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. PMLR, 2017.

- Qintong Li, Jiahui Gao, Sheng Wang, Renjie Pi, Xueliang Zhao, Chuan Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Forewarned is forearmed: Leveraging llms for data synthesis through failure-inducing exploration. *arXiv* preprint arXiv:2410.16736, 2024a.
 - Xiaochuan Li, Zichun Yu, and Chenyan Xiong. Montessori-instruct: Generate influential training data tailored for student learning. *arXiv preprint arXiv:2410.14208*, 2024b.
 - Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6), 2023.
 - Kiwan Maeng, Alexei Colin, and Brandon Lucia. Alpaca: Intermittent execution without checkpoints. *Proceedings of the ACM on Programming Languages*, 1(OOPSLA):1–30, 2017.
 - Nihal V Nayak, Yiyang Nan, Avi Trost, and Stephen H Bach. Learning to generate instruction tuning datasets for zero-shot task adaptation. *arXiv preprint arXiv:2402.18334*, 2024.
 - Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin A Raffel, Leandro Von Werra, Thomas Wolf, et al. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024.
 - Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. arXiv preprint arXiv:2304.03277, 2023.
 - Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.
 - Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33: 19920–19930, 2020.
 - Kashun Shum, Yuzhen Huang, Hongjian Zou, Qi Ding, Yixuan Liao, Xiaoxin Chen, Qian Liu, and Junxian He. Predictive data selection: The data that predicts is the data that teaches. *arXiv* preprint arXiv:2503.00808, 2025.
 - Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, 2025.
 - Yu Sun, Xingyu Qian, Weiwen Xu, Hao Zhang, Chenghao Xiao, Long Li, Yu Rong, Wenbing Huang, Qifeng Bai, and Tingyang Xu. Reasonmed: A 370k multi-agent generated dataset for advancing medical reasoning. *arXiv preprint arXiv:2506.09513*, 2025.
 - Teknium. Openhermes 2.5: An open dataset of synthetic data for generalist Ilm assistants, 2023. URL https://huggingface.co/datasets/teknium/OpenHermes-2.5.
 - Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.
 - Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multitask language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024.
 - Alexander Wettig, Aatmik Gupta, Saumya Malik, and Danqi Chen. Qurating: Selecting high-quality data for training language models. *arXiv preprint arXiv:2402.09739*, 2024.
 - Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. Less: Selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333*, 2024.
 - Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy S Liang. Data selection for language models via importance resampling. *Advances in Neural Information Processing Systems*, 36: 34201–34227, 2023.

- Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. Wizardlm: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations*, 2024.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.
- Zichun Yu, Spandan Das, and Chenyan Xiong. Mates: Model-aware data selection for efficient pretraining with data influence models. *Advances in Neural Information Processing Systems*, 37: 108735–108759, 2024.
- Xiang Yue, Tianyu Zheng, Ge Zhang, and Wenhu Chen. Mammoth2: Scaling instructions from the web. *Advances in Neural Information Processing Systems*, 37:90629–90660, 2024.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.
- Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470*, 2024.
- Chiwei Zhu, Benfeng Xu, Xiaorui Wang, and Zhendong Mao. From real to synthetic: Synthesizing millions of diversified and complicated user instructions with attributed grounding. *arXiv* preprint arXiv:2506.03968, 2025.

A THE USE OF LARGE LANGUAGE MODELS

In the preparation of this manuscript, GPT-5 was employed solely for language refinement purposes, including grammar checking and stylistic polishing. All conceptualization, methodological design, experiments, analyses, and interpretations were conducted by the authors. The LLM was not used to generate novel content, ideas, or results, but only to assist in improving clarity, readability, and consistency of expression. Specifically includes: 1 INTRODUCTION; 2 METHOD; 4 RELATED WORKS; 5 CONCLUSION.

B INFLUENCE ESTIMATION

Influence estimation Koh & Liang (2017); Pruthi et al. (2020) aim to quantify how an individual training example perturbs both the learned parameters and the loss at a target (validation/test) point. In the large-language-model (LLM) regime, computing exact influences is prohibitively expensive; consequently, prior work Pruthi et al. (2020); Xia et al. (2024) adopts a first-order approximation to the training dynamics to estimate the influence of a training datapoint on held-out performance.

Per-step influence. Let θ^t denote the model parameters at step t and $\ell(\cdot; \theta^t)$ the loss. A first-order Taylor expansion of the loss on a validation example z' gives

$$\ell(z'; \theta^{t+1}) \approx \ell(z'; \theta^t) + \langle \nabla_{\theta} \ell(z'; \theta^t), \theta^{t+1} - \theta^t \rangle. \tag{4}$$

Assuming SGD with batch size 1 and learning rate η_t , if z is used at step t then $\theta^{t+1} - \theta^t = -\eta_t \nabla_{\theta} \ell(z; \theta^t)$, and equation 4 yields the per-step influence

$$\ell(z'; \theta^{t+1}) - \ell(z'; \theta^t) \approx -\eta_t \left\langle \nabla_{\theta} \ell(z; \theta^t), \nabla_{\theta} \ell(z'; \theta^t) \right\rangle. \tag{5}$$

(For mini-batches, replace $\nabla_{\theta} \ell(z; \theta^t)$ with the batch gradient.)

Trajectory influence (SGD). Aggregating equation 5 over epochs results in a trajectory-level score:

$$\operatorname{Inf}_{\operatorname{SGD}}(z, z') = \sum_{i=1}^{N} \bar{\eta}_{i} \langle \nabla_{\theta} \ell(z'; \theta_{i}), \nabla_{\theta} \ell(z; \theta_{i}) \rangle, \tag{6}$$

where $\bar{\eta}_i$ is the average learning rate in epoch i (out of N total epochs) and θ_i is the checkpoint after epoch i.

Extension to Adam. Instruction tuning typically uses Adam (Adam et al., 2014). Let the per-coordinate update direction be

$$\Gamma(z, \theta^t) \triangleq \frac{\hat{\boldsymbol{m}}^{t+1}}{\sqrt{\hat{\boldsymbol{v}}^{t+1}} + \epsilon}, \quad \theta^{t+1} - \theta^t = -\eta_t \Gamma(z, \theta^t), \tag{7}$$

with elementwise operations and

$$\begin{split} & \boldsymbol{m}^{t+1} = \beta_1 \boldsymbol{m}^t + (1 - \beta_1) \, \nabla_{\theta} \ell(z; \theta^t), \\ & \boldsymbol{v}^{t+1} = \beta_2 \boldsymbol{v}^t + (1 - \beta_2) \, \nabla_{\theta} \ell(z; \theta^t)^{\odot 2}, \\ & \hat{\boldsymbol{m}}^{t+1} = \boldsymbol{m}^{t+1} / (1 - \beta_1^{t+1}), \quad \hat{\boldsymbol{v}}^{t+1} = \boldsymbol{v}^{t+1} / (1 - \beta_2^{t+1}). \end{split}$$

A first-order expansion analogous to equation 5 gives the stepwise effect $-\eta_t \langle \nabla_{\theta} \ell(z'; \theta^t), \Gamma(z, \theta^t) \rangle$. Because the magnitude of Γ varies with the adaptive moments, we use a scale-robust cosine similarity when accumulating across epochs, and define

$$\operatorname{Inf}_{\operatorname{Adam}}(z, z') = \sum_{i=1}^{N} \bar{\eta}_{i} \cos(\nabla_{\theta} \ell(z'; \theta_{i}), \Gamma(z, \theta_{i})). \tag{8}$$

Computing Γ requires the moment statistics (m, v), which depend on past gradients; following our procedure in §3.1.3, we obtain them from a short warmup run before scoring.

C RELATED WORKS: SYNTHETIC DATA GENERATION

With the continued improvement of LLM capabilities, synthetic data has been widely used to alleviate the high annotation costs in instruction tuning and dialogue scenarios, many studies convert readily available document corpora into SFT-style dialogue samples using LLMs. WebR (Jiang et al., 2025) treats web pages as a dual-view reconstruction target: first, page content is converted into structured instructions and corresponding responses; second, the page is regarded as the response to infer latent instructions, after which the initial responses are refined under evidence constraints. This enables the construction of high-quality instruction-tuning corpora without human intervention or seed data. MAmmoTH2 (Yue et al., 2024) proposes a scalable pipeline comprising web recall, extraction of instruction-response pairs, decontamination, and refinement. Bonito (Nayak et al., 2024) maps unlabeled text and task attributes into instruction-response samples via meta-template-driven conditional task generators, turning domain documents into trainable data and enabling zero-shot cross-task and cross-domain adaptation. A complementary line of work focuses on iteratively improving the synthesized data. Evol-Instruct (Xu et al., 2024) evolves seed instructions in depth and breadth to progressively increase task difficulty and generate matched responses, illustrating the potential of self-evolving synthesis. Condor (Cao et al., 2025) constructs a World Knowledge Tree to ensure topic coverage and difficulty stratification and introduces a self-reflection refinement mechanism that conducts structured self-evaluation of candidate answers along strengths, weaknesses, and suggestions before rewriting, yielding a two-stage knowledge-driven and self-reflective synthesis framework. Li et al. (2024a) adopts failure-inducing exploration by feeding cases that the student model still fails on back to the teacher to synthesize targeted hard examples. Montessori-Instruct (Li et al., 2024b) forms pairs of beneficial and non-beneficial data and applies Direct Preference Optimization to train the teacher's preference toward examples that are most instructive for downstream performance.

Current data selection strategies can be broadly divided into two categories:(1) Static criteria (human-authored rubrics / LLM-as-Judge): Emphasize interpretability and control. Representative approaches include: QuRating (Wettig et al., 2024), which uses an LLM to score texts along dimensions such as quality, expertise, factuality/knowledge, and educational value to guide sample selection; DSIR (Xie et al., 2023), which applies importance resampling to large-scale pretraining subset selection; and FineWeb (Penedo et al., 2024), which performs large-scale deduplication and filtering and uses distributional visualizations and ablations to validate how data quality affects performance, thereby providing a tooling stack for curating high-quality general-purpose corpora. (2) Dynamic feedback (model-/task-driven): Directly selects samples using signals for what is most instructive for the model. Representative approaches include: LESS (Xia et al., 2024), which selects instruction samples via influence estimates and low-rank gradient similarity; PreSelect (Shum et al., 2025), which uses predictability as a lightweight scoring function; PDS (Gu et al., 2024), which formulates data selection as an optimal-control problem to characterize selection—training dynamics; and MATES (Yu et al., 2024), which learns a data-influence model that adapts over training to match the most effective samples for the current stage.

D EXPERIMENT DETAILS

1 SOURCE OF TRAINING DATASET

Our seed documents for synthesizing data in the humanities and social sciences were collected from books. The list of book titles is as follows:

A Black Jurist in a Slave Society Antonio Pereira Rebouças and the Trials of Brazilian Citizenship; A History of Intellectual Property in 50 Objects; A History of Modern Africa 1800 to the Present Wiley Blackwell Concise History of the Modern World; A Taste of Irrationality Sample Chapters From Predictably Irrational and Upside of Irrationality; A Case for Sometimes Tubefeeding Patients in Persistent Vegetative State; Adult Attachment A Concise Introduction to Theory and Research; Advanced Programming in the UNIX Environment 3rd Edition; Advanced Signal Integrity for Highspeed Digital Designs; Advocating for Self Womens Decisions Concerning Contraception; African History A Very Short Introduction; Algorithm Design Monograph; American Criminal Justice Policy An Evaluation Approach to Increasing Accountability and Effectiveness; American Grace How Religion Divides and Unites Us; An Analysis of Ernest Gellners Nations and

865

866

867

868

870

871

872

873

874

875

876

877

878

879

880

883

884

885

887

889

890

891

892

893

894

895

896

897

899

900

901

902

903

904

905

906

907

908

909

910

911

912

914

915

916

Nationalism; An Analysis of Marcel Mauss The Gift The Form and Reason for Exchange in Archaic Societies; An Analysis of Robert A Dahls Who Governs Democracy and Power in an American City; An Analysis of Robert E Lucas Jrs Why Doesnt Capital Flow from Rich to Poor Countries; An Organizers Tale Speeches; Assessments in Forensic Practice A Handbook; Bargaining with the Devil When to Negotiate When to Fight; Basic Political Writings; Bertrand Russell and the Nature of Propositions A History and Defence of the Multiple Relation Theory of Judgement; Beyond Reason Using Emotions as You Negotiate; Blackness in Britain; Challenging Behaviour in Dementia A Personcentred Approach; Changing Minds The Art and Science of Changing Our Own and Other Peoples Minds; Child Soldiers A Reference Handbook; Children and Play Understanding Childrens Worlds; Cite Them Right The Essential Referencing Guide 12th Edition; Cognitive Behavioral Therapy for PTSD A Case Formulation Approach; Cognitive Behavioral Treatment for Generalized Anxiety Disorder From Science to Practice; Collaborative Intelligence Using Teams to Solve Hard Problems; Collective Action and Exchange A Gametheoretic Approach to Contemporary Political Economy; Companion Encyclopedia of Medicine in the Twentieth Century; Computer Architecture A Quantitative Approach; Concise Guide to APA Style The Official APA Style Guide for Students; Confronting The Internets Dark Side Moral And Social Responsibility On The Free Highway; Contemporary Debates in Cognitive Science Contemporary Debates in Philosophy; Curing Their Ills Colonial Power and African Illness; Dependency and Development in Latin America; Descriptive Physical Oceanography Sixth Edition An Introduction; Development Microeconomics; Dignity and Daily Bread New Forms of Economic Organizing among Poor Women in the Third World and the First; Divine Hiddenness and Human Reason; Economics and Policy Issues in Climate Change; Economics of Regulation and Antitrust Fifth Edition The MIT Press; Effective Treatments for PTSD Third Edition Practice Guidelines from the International Society for Traumatic Stress Studies; European Legal History A Cultural and Political Perspective; Euthanasia Ethics and the Law from Conflict to Compromise; Exchange Rate Regimes Fixed Flexible or Something in Between; Experiences of Depression A Study in Phenomenology; Factfulness Ten Reasons Were Wrong About the World and Why Things Are Better Than You Think; Fatal Invention How Science Politics and Big Business Recreate Race in the Twentyfirst Century; Feeding the World An Economic History of Agriculture 1800-2000; Feminism Literature and Rape Narratives Violence and Violation; Feminism Unmodified Discourses on Life and Law; Feminist Judgments From Theory to Practice; Fiasco The American Military Adventure in Iraq 2003 to 2005; Formulation in Psychology and Psychotherapy Making Sense of Peoples Problems; Free Innovation; From the Great Recession to Labour Market Recovery Issues Evidence and Policy Options; Fundamentals of Developmental Psychology; Geographies of Development An Introduction to Development Studies; Getting It Wrong How Canadians Forgot Their Past and Imperilled Confederation; HBRs 10 Must Reads on Managing Strategy; Hackers Delight; Helping People Win at Work A Business Philosophy Called Dont Mark by Paper Help Me Get an A; Her Place at the Table A Womans Guide to Negotiating Five Key Challenges to Leadership Success; Hindu Worldviews Theories of Self Ritual and Reality; How To Do Your Research Project 3rd Edition; In Command of History Churchill Fighting and Writing the Second World War; Infertility and Patriarchy The Cultural Politics of Gender and Family Life in Egypt; Influence New and Expanded The Psychology of Persuasion; Innovation and Entrepreneurship Practice and Principles; Institutions Institutional Change and Economic Performance; Introduction to Econometrics Global Edition; Introduction to Meta-Analysis; Introductory Econometrics A Modern Approach MindTap Course List; Islam and Global Dialogue Religious Pluralism and the Pursuit of Peace; John P Kotter on What Leaders Really Do; Judgment How Winning Leaders Make Great Calls; Just Health Meeting Health Needs Fairly; Key Concepts in Geography; Law Legitimacy and the Rationing of Health Care A Contextual and Comparative Perspective; Leadership on the Line Staying Alive Through the Dangers of Leading; Lean In Women Work and the Will to Lead; Leviathan and the Air Pump Hobbes Boyle and the Experimental Life; Linear Algebra Done Right; Macleods Clinical Examination Ebook; Making Babies Is There a Right to Have Children; Making Democracy Work Civic Traditions in Modern Italy; Manias Panics and Crashes A History of Financial Crises; Marshall and the Marshallian Heritage Essays in Honour of Tiziano Raffaelli; Mastering Leadership A Vital Resource for Health Care Organizations; Math with Bad Drawings Illuminating the Ideas That Shape Our Reality; Microeconomics 9th Global Edition; Modeling Foundations of Economic Property Rights Theory An Axiomatic Analysis of Economic Agreements Studies in Economic Theory 23; Modern Classics Making of the English Working Class Penguin Modern Classics; Modern Environmentalism An Introduction; Money Well Spent A Strategic Plan for Smart Philanthropy; Mostly Harmless Econometrics; National Insecurity American Leadership in an Age of Fear; New Museum Theory and Practice An Introduction; Notes on

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

957

958

959

960

961

962

963

964

965

966

967

968

969

970

971

the Theory of Choice; Oil Palm A Global History; One Economics Many Recipes Globalization Institutions and Economic Growth; Operating Systems In Depth Design and Programming; Ordinary and Partial Differential Equations With Special Functions Fourier Series and Boundary Value Problems; Organizational Trust A Reader; Paul Samuelson on the History of Economic Analysis Selected Essays; Personality and Intelligence at Work Exploring and Explaining Individual Differences at Work; Philosophical Issues in Psychiatry Explanation Phenomenology and Nosology; Philosophy of Psychology Contemporary Readings; Presidential Leadership and the Creation of the American Era; Psychiatry as Cognitive Neuroscience Philosophical Perspectives; Psychological Assessment and Therapy with Older Adults; Psychology The Science of Mind Behaviour; Publication Manual of the American Psychological Association The Official Guide to APA Style; Raising Cando Kids Giving Children the Tools to Thrive in a Fastchanging World; Randomized Algorithms; Real Estate Development Principles and Process; Reconstructing Educational Psychology; Religion and Politics in the United States; Religion as Make Believe A Theory of Belief Imagination and Group Identity; Religion Violence Memory and Place; Research Methods and Statistics in Psychology 7th Edition; Rethinking Informed Consent in Bioethics; Reveille for Radicals; Rewriting the Soul Multiple Personality and the Sciences of Memory; Risk Ambiguity and Decision; Rule of Experts Egypt Technopolitics Modernity; Science and the Practice of Medicine in the Nineteenth Century; Seeing Like a State How Certain Schemes to Improve the Human Condition Have Failed; Selfcare Science Nursing Theory and Evidencebased Practice; Senior Leadership Teams What It Takes to Make Them Great; Slave Country American Expansion and the Origins of the Deep South; Small Differences That Matter Labor Markets and Income Maintenance in Canada and the United States; Small States in World Markets Industrial Policy in Europe; Smart Choices A Practical Guide to Making Better Life Decisions; Social Research; State Responsibility The General Part; Strategic Leadership and Management in Nonprofit Organizations Theory and Practice; Strategic Management for Nonprofit Organizations Theory and Cases; Strategy Safari A Guided Tour Through the Wilds of Strategic Management; Studies in Tudor and Stuart Politics and Government Papers and Reviews 1946-1972 Vol 1; Supervision in the Helping Professions Supervision in Context; The 5 Elements of Effective Thinking; The American Political Economy Macroeconomics and Electoral Politics; The Cambridge Companion to Shakespeare Cambridge Companions to Literature; The Cambridge Companion to Victorian and Edwardian Theatre Cambridge Companions to Literature; The Cambridge Handbook of Personality Psychology; The Economy of the Word Language History and Economics; The Far Enemy Why Jihad Went Global Second Edition; The Greatest Benefit to Mankind A Medical History of Humanity from Antiquity to the Present; The IBS Elimination Diet and Cookbook The Low FODMAP Plan for Eating Well and Feeling Great; The Little Black Book of Neuropsychology A Syndrome Based Approach; The Lucifer Effect Understanding How Good People Turn Evil; The Moral Economy of the Peasant Rebellion and Subsistence in Southeast Asia; The Mystery of Capital Why Capitalism Triumphs in the West and Fails Everywhere Else; The Politics of Life Itself Biomedicine Power and Subjectivity in the Twenty First Century; The Postcolonial Studies Reader; The Probabilistic Method; The Protestant Ethic and the Spirit of Capitalism; The Quest Energy Security and the Remaking of the Modern World; The Routledge Companion to Islamic Philosophy Routledge Philosophy Companions; The Routledge Handbook of Memory Activism; The SAGE Handbook of Social Geographies; The Social History of English Seamen 1485–1649; The United Nations and Changing World Politics Revised and Updated with a New Introduction; The Western Medical Tradition 1800-2000; The Wiley Blackwell Handbook of Individual Differences; The Architecture of the Mind Massive Modularity and the Flexibility of Thought; The Balanced Scorecard Translating Strategy into Action; The Blank Slate The Modern Denial of Human Nature; The Emergence of Industrial America Strategic Factors in American Economic Growth since 1870; The Female Frontier A Comparative View of Women on the Prairie and the Plains; The Filter Bubble What the Internet is Hiding from You; The Fortune at the Bottom of the Pyramid Eradicating Poverty Through Profits; The Global Interior Mineral Frontiers and American Power; The Missing Peace The Inside Story of the Fight for Middle East Peace; The Political Determinants of Health; The Prize The Epic Quest for Oil Money Power; The Psychological Complex Social Regulation and the Psychology of the Individual; The Slave Ship A Human History; To Vote or Not to Vote The Merits and Limits of Rational Choice Theory; Treatment of Obsessive Compulsive Disorder; Unified Growth Theory; Varieties of Capitalism The Institutional Foundations of Comparative Advantage; West from Appointant The Reconstruction of America After the Civil War; Why David Sometimes Wins Leadership Organization and Strategy in the California Farm Worker Movement; Why Nations Fail The Origins of Power Prosperity and Poverty; Why We Disagree About Climate Change Understanding Controversy Inaction and Opportunity; Wind Energy Renewable Energy and the Environment; Working Bodies

Interactive Service Employment and Workplace Identities; Yangzi Waters Transforming the Water Regime of the Jianghan Plain in Late Imperial China China Studies 44; The Courage to Create

2 IMPLEMENTATION DETAILS

 We conduct all experiments on 8×H200 GPUs. The framework consists of three stages: warm-up, reinforcement learning (RL), and final SFT training.

Warm-up. At the beginning of training, the initialized prompter synthesizes a set of *rubrics*, which guide GENERATOR to produce an initial batch of SFT data. We then warm up the target model qwen3-8b-base by fine-tuning it with LoRA adapters and the Adam optimizer, using 10% of the synthetic data with batch size 16, learning rate $1! \times !10^{-5}$, and training for one epoch. The resulting model provides loss gradients as reference signals for influence score estimation. Each training example is assigned an *influence score*, which is normalized within the batch and used as the reward signal to update the prompter via reinforcement learning (RL).

Reinforcement Learning. During RL, a generator is deployed on $8 \times H200$ GPUs to synthesize QA pairs conditioned on seed documents and prompter rollouts. Each synthetic instance is scored by comparing the update direction (from Adam on the generated sample) with validation gradients of the target model. The similarity is then normalized via min-max scaling to [0,1], forming the primary reward signal. We adopt GRPO as the training algorithm, where the overall reward combines the influence score with format-consistency rewards. Training uses batch size 256, learning rate 1×10^{-6} , rollout temperature 1.5, rollout size n=5, and one epoch of updates. Processing $\sim 20 \text{K}$ samples requires roughly 10 hours.

Final SFT. After RL, the refined prompter generates an extended synthetic dataset, which is then used to conduct supervised fine-tuning (SFT) on <code>qwen3-8b-base</code> as the target model. This ensures that the model benefits both from high-quality rubric-guided data and from the RL-enhanced prompter optimization.

3 Instructions for Prompter

System Prompt for Prompter to Generate a Rubrics

You are an expert in the {domain} field. Your core task is to instruct the "Generator" LLM to produce a single, high-quality Question–Answering (QA) data point based on the provided {domain} content.

Figure 10: Instruction template for generating a QA pair.

Prompt Template for Prompter to analyze quality dimensions

There is a given document: {document}

Objective:

1026

1027 1028

1029 1030

1031

1032

1033 1034 1035

1039

1041 1042

1043

1045 1046

1047

1048

1049

1050

1051

1052

1056

1058

1062

1063 1064

1067

1068

1069

1070

1071

1075

1077 1078 1079 Guide the model to generate corresponding questions and answers based on the document provided above. This QA data will be used to fine-tune a downstream Target LLM, aiming to significantly enhance its **comprehension**, **analysis**, **and creative capabilities** in {domain} tasks.

Instructions:

- 3.1 Randomly select one question type from the following set: [Multiple choice, Fill-in-the-blank, Short answer, Essay].
- 3.2 Based on the characteristics of the document, analyze which quality dimensions should be prioritized for the question and answer under the selected type (with the ultimate goal of enabling the model to perform well in {domain} downstream tasks).
 - The analysis may include, but is not limited to: the document's topic, style, purpose, complexity level, target audience, linguistic style, and knowledge density. Broad quality dimensions (e.g., scope, clarity, creativity, relevance, reasoning depth, factual accuracy) can be used, but their importance and descriptions should be adapted to match the specific characteristics of the document.
 - At least three additional dimensions beyond the above examples should be considered such as completeness, format compliance, tone alignment, safety, neutrality, information density, instructional precision, or adaptability to ambiguous inputs.
 - When defining and prioritizing quality dimensions, you may draw on examples such as:
 - **Prompt-related**: clarity, unambiguity, executability, conciseness, focus, language and grammar correctness, format appropriateness, domain relevance, bias avoidance, temporal validity, contextual coherence.
 - **Response-related**: correctness, completeness, relevance, logical soundness, conciseness, fluency, credibility of information sources, creativity, information value density, neutrality, safety.
 - **Prompt–Response alignment**: consistency, logical linkage, coverage of all sub-questions, appropriate disambiguation, handling of potentially incorrect prompts, compliance with constraints.
 - **Technical/trainability aspects**: parsability, consistent encoding, absence of PII, diversity across samples, low noise, length suitability, copyright compliance.
 - Avoid using fixed templates or rigid numerical scales. Do not provide specific content of the Q-A; instead, focus on evaluating which dimensions matter most and why.
- 3.3 The standards must be actionable to ensure the "Generator" can accurately understand and synthesize the required high-quality data.

Output Format:

```
{"Prompt-related": ...,

"Response-related": ...,

"Prompt-Response alignment": ...,

"Technical/trainability aspects": ...}
```

Figure 11: Prompt template for analyzing quality dimensions before QA generation.

4 Instructions for Generator

System Prompt for Generator to Generate a QA pair

Prompt Template for Generator to generate QA

You are an expert in the field of {domain}. I will now provide you with a document. Based on this document, you need to synthesize **ONLY 1 pair** of high-quality question—answer (QA) data, which will be used for downstream task training.

Figure 12: Instruction template for generating a QA pair.

```
There is a given document: {doc}
## Specific requirements for generating the QA data:
     4.1 Generate ONLY 1 QA pair.
     4.2 Question Quality Standard: {Prompt-related}.
     4.3 Answer Quality Standard: {Response-related}.
    4.4 Prompt-Response alignment: {Prompt-Response alignment}.
    4.5 Technical/trainability aspects: {Technical/trainability aspects}.
     4.6 For multiple-choice questions, it is necessary to generate a "think step by step" reasoning
         process in the answer, ultimately leading to the conclusion.
     4.7 All generated questions and answers must be closely related to the given content. However,
         do not indicate that you can see the content or use expressions like "according to the docu-
         ment." Do not mention the document at all.
     4.8 The output format must be JSONL, with the specific structure as follows:
            "messages": [
               {"role": "system", "content": "You are a helpful assistent."},
               { "role": "user", "content": Question },
               {"role": "assistant", "content": Answer}
         }
```

Figure 13: Prompt template for generating QA data in the HSS and Medical domain.

```
1134
              EXAMPLES OF RUBRICS
1135
1136
1137
1138
1139
1140
1141
                 There is a given document: {doc}
1142
                 1. Generate ONLY 1 QA pair.
1143
                 2. Question Quality Standard: {
1144
                       "Clarity": "Questions must be clear and unambiguous, explicitly referring to specific categories
1145
                       of medieval hospitals, such as small hostels/hospices, large civic hospitals, or hospitals
1146
                       established by crusading orders, to ensure precise alignment with the structured descriptions in
1147
                       the document.",
                       "Domain Relevance": "Questions should focus on core themes such as religious character, the
1148
                       process of medicalization, social service recipients, and drivers of expansion, to maintain high
1149
                       relevance to the Humanities and Social Sciences domain.",
1150
                       "Format Appropriateness": "For short-answer questions, prompts should specify the required
1151
                       response format, such as briefly explaining the main function or expansion reason of a
1152
                       particular type of hospital, to facilitate evaluation."
1153
                 3. Answer Quality Standard: {
1154
                       "Correctness": "Answers must accurately reflect the definitions in the document, for example
1155
                       correctly describing how a type of game uses power, to ensure factual reliability."
1156
                       "Completeness": "Responses should address key aspects of the game type, including its
1157
                       objective, primary participants, and operating mechanisms.",
1158
                       "Relevance": "Answers must avoid introducing unrelated theories or cases that do not pertain to
                       organizational political games, so as to maintain focus."
1159
                       },
1160
                 4. Prompt-Response alignment: {
1161
                       "Consistency": "The answer must remain consistent with the question's focus, such as clearly
1162
                       differentiating the roles and objectives of two game types when making comparisons.".
1163
                       "Format Matching": "For multiple-choice questions, the answer should correspond to the
1164
                       provided options and avoid adding new, unlisted information.",
                       "Handling Ambiguous Inputs": "If the question is vague, the answer should explicitly reference
1165
                       the name of a specific game type from the document to ensure clarity."
1166
1167
                 5. Technical/trainability aspects: {
1168
                       "Parsability": "QA pairs must use structured, machine-readable formats (such as JSON) to
1169
                       facilitate automated processing and training.",
                       "Copyright Compliance": "Answers should not reproduce the text verbatim but instead
1170
                       paraphrase in original language while preserving meaning.",
1171
                       "Length Suitability": "Responses should be concise, ideally within 200 tokens, to prevent
1172
                       truncation during model training.'
1173
1174
                 6. For multiple-choice questions, it is necessary to generate a "think step by step" reasoning process in
1175
                 the answer, ultimately leading to the correct conclusion.
                 7. All generated questions and answers must be closely related to the given content. However, do not
1176
                 indicate that you can see the content or use expressions like "based on the document." Do not mention
1177
                 the document at all.
1178
                 8. The output format must be JSONL, with the following structure:
1179
1180
                             "messages": [{"role": "system", "content": "you are a helpful assistent."},
                                   {"role": "user", "content": Question},
1181
                                   {"role": "assistant", "content": Answer}]
1182
1183
1184
```

Figure 14: An example of generated rubrics for data synthesis.

1188 1189 1190 1191 There is a given document: {doc} 1192 1193 1. Generate ONLY 1 QA pair. 1194 2. Question Quality Standard: Questions should have high clarity and strong domain relevance, focusing directly on the core concepts described in the source document about medieval hospitals 1195 — including their religious foundations, structural variations, medicalization process, and historical 1196 expansion. For example, they might compare differences between small hostels/hospices and large 1197 civic hospitals, analyze the role of the Crusades in hospital network expansion, or examine how 1198 institutions such as Sta Maria Nuova evolved to serve both indigent and higher-class patients. Avoid 1199 vague wording that could obscure distinctions between civic and religious institutions or the nature of services provided. Each question should clearly set the theme and scope and, when applicable, identify relevant historical categories. For short-answer formats, prompts may indicate the expected 1201 response structure (e.g., "Summarize how Sta Maria Nuova transitioned from a hospice to a medicalized hospital in two sentences"). Ideally, questions should require synthesis across the entire text rather than retrieval of a single sentence. 3. Answer Quality Standard: Answers should accurately reflect the text's historical details, 1205 distinguishing between different hospital types (religious vs civic, small hospices vs large urban hospitals), and highlighting contextual factors such as the influence of Crusading orders and nonmilitary brotherhoods. Responses must be logically coherent, factually correct, and address all 1207 relevant aspects mentioned in the question, including chronology, institutional roles, and socio-1208 economic reach. Information density can be enhanced by providing key examples from the text 1209 (e.g., St Leonard's in York, St Bartholomew's, Paris Hôtel Dieu, Sta Maria Nuova's 1210 medicalization). Answers must rely on a comprehensive reading of the entire passage, not just isolated lines, and preserve the distinctions made in the text. Clear differentiation between similar 1211 cases will help maintain alignment with the text's historical framework. 1212 4. Prompt-Response alignment: Strong alignment is shown when the answer directly ties to the 1213 question's focus and cites concrete examples from the text that illustrate structural, functional, and 1214 historical contrasts. If the prompt requests a comparison of hospital types or factors driving 1215 expansion, the answer should deliver that comparison with historical examples (e.g., Crusader-built 1216 hospitals vs. Order of the Holy Spirit network). Completeness involves covering all subcomponents in the question (such as stakeholders, mechanisms, intended outcomes) and presenting them in the 1217 expected structure and tone. Matching quality should consider whether the depth, detail selection, 1218 and contextual fit align with the question's intent. Logical consistency must be maintained, 1219 avoiding conflation between distinct institutions or time periods. 1220 5. Technical/trainability aspects: Paraphrasing should replace verbatim copying to maintain meaning while improving linguistic flexibility. Responses should avoid excessive length to stay within model context limits. A neutral, academic tone should be maintained, factually describing hospital foundations, structures, and development without subjective evaluation. These standards guide the generation of high-quality question-answer pairs that are clearly aligned with the theme 1224 of medieval hospitals, historically accurate, and evaluable in training contexts. 1225 6. For multiple-choice questions, it is necessary to generate a "think step by step" reasoning process 1226 in the answer, ultimately leading to the correct conclusion. 1227 7. All generated questions and answers must be closely related to the given content. However, do not indicate that you can see the content or use expressions like "based on the document." Do not 1228 mention the document at all 1229 8. The output format must be JSONL, with the following structure: 1230 1231 "messages": [{"role": "system", "content": "you are a helpful assistent."}, 1232 {"role": "user", "content": Question}, 1233 {"role": "assistant", "content": Answer}]

Figure 15: An example of generated rubrics for data synthesis.

6 EXAMPLES OF SYNTHETIC QA PAIRS

1242

1247 1248

1249

1250

1251 1252 1253

1255

1259

1262 1263 1264

1270

1278

1279

1280

1281

1282

1283

1284

1285

1286

1293 1294 1295 **System prompt:** You are a helpful assistant.

Question: Briefly compare the differences between large medieval civic hospitals and small hostels/hospices in terms of religious background, medical staff allocation, and drivers of expansion.

Answer: Large civic hospitals were typically established in urban centers, with a strong religious background, but supported by city economies and municipal funding. They gradually incorporated professional doctors, pharmacists, and auxiliary staff, achieving a high degree of medicalization, and catered to the poor while also providing private rooms for upper-class patients. In contrast, small hostels or hospices were mostly tied to local religious charity traditions, limited in scale, and had almost no resident medical staff. Their expansion was driven mainly by charitable donations from local religious orders or parishes, rather than systematic public funding.

Figure 16: An example of synthetic QA pairs.

System prompt: You are a helpful assistant.

Question: Analyze how the religious character and medicalization process of medieval hospitals were manifested in different types of institutions (such as small hostels/hospices, large civic hospitals, and hospitals established by crusading orders), and compare their differences in medical staff allocation, social service recipients, and drivers of expansion.

Answer: Medieval hospitals generally possessed a strong religious character. Small hostels or hospices were essentially attached to charitable religious traditions, primarily serving pilgrims with lodging and basic care in their early stages, and lacked resident medical personnel. In contrast, large civic hospitals such as Paris's Hôtel Dieu and Florence's Santa Maria Nuova gradually introduced dedicated doctors and pharmacists, signaling significant progress in medicalization. These institutions not only treated the poor but also provided private rooms reserved for upper-class patients, reflecting diversity in their social service recipients. Hospitals established by crusading orders—such as the Knights Hospitaller, the Knights Templar, and the Teutonic Knights—were mostly located in Mediterranean and German-speaking regions, primarily serving the needs of military forces and local populations along their routes, with expansion driven by military campaigns and religious missions rather than urban public service demands. The growth of small religious hostels was constrained by funding and local religious resources, while large civic hospitals relied on urban populations and municipal support, and the network of crusader hospitals was sustained by the transregional organizational strength of the orders. Overall, these three types of institutions were similar in their pervasive Christian ethos but differed sharply in the scale of medical staff deployment, the breadth of their service recipients, and the factors driving their expansion.

Figure 17: An example of synthetic QA pairs.

EXAMPLES OF SEED DATA

Title / Chapter: HOSPITALS

Medieval hospitals were religious foundations through and through. Those planted in the West had originally been small and mainly for pilgrims; their late medieval successors were often more impressive. St Leonard's in York had 225 sick and poor in 1287; still larger were the civic hospitals of Milan, Siena and Paris. In Florence alone, a city of some 30,000 inhabitants, there were over thirty foundations by the fifteenth century. Some had only ten beds, others hundreds. In England hospitals and almshouses totalled almost five hundred by 1400, though few were of any size or significance. London's St Bartholomew's dates from 1123 and St Thomas's from around 1215. At Bury St Edmunds six hospitals were endowed between I150 and I260 to cater for lepers, pilgrims, the infirm and the aged.\n\nSmall hospitals were essentially hostels or hospices lacking resident medical assistance, but physicians were in attendance by 1231 at the Paris Hôtel Dieu, next to Notre Dame, and Sta Maria Nuova in Florence was gradually medicalized: from twelve beds in 1288 for 'the sick and the poor', this 'first hospital among Christians', as one Florentine patriot called it, expanded by 1500 to a medical staff of ten doctors, a pharmacist and several assistants, including female surgeons. Although catering largely for the indigent, it had eight private rooms 'reserved for the sick of the higher classes'. Within hospital walls the Christian ethos was all-pervasive.\n\nIn hospital expansion the Crusades played their part, since crusading orders such as the Knights of St John of Jerusalem (later the Knights of Malta), the Knights Templar, and the Teutonic Knights built hospitals throughout the Mediterranean and German-speaking lands. By the fourteenth century nonmilitary brotherhoods, such as the Order of the Holy Spirit, were also running infirmaries from Alsace to Poland, while the Order of St John of God appeared in Spain in the sixteenth century, building insane asylums and putting up about 200 hospitals in the New World.

Figure 18: An example of seed data.