

HSS-Synth: Humanities and Social Sciences Data Synthesis for LLMs

Ru Peng^{1*}, Tianyu Zhao², Xijun Gu¹, Zhiting Fan¹,
Haokai Xu¹, Jinyang Zhang^{3,4}, Yawen Zeng¹, Yihong Zhuang²,
Kexin Yang⁴, Junyang Lin⁴, Dayiheng Liu^{4†}, Junbo Zhao^{1†}

¹Zhejiang University

²Inclusion AI, Ant Group

³Peking University

⁴Qwen Team, Alibaba Group

{rupeng, j.zhao}@zju.edu.cn

liudayiheng.ldyh@alibaba-inc.com

Abstract

High-quality, diverse data are vital for large language models (LLMs) but remain scarce and costly. Data synthesis is a viable alternative and succeeds on closed tasks, yet the humanities and social sciences (HSS) are overlooked, and their open-ended nature makes synthesis challenging. Moving beyond prior capability-centric, fragmented attempts, we adopt a subject-centric paradigm, define the first HSS domain system covering 14 mainstream fields, and introduce HSS-Synth—the first data synthesis pipeline for HSS. HSS-Synth comprises: (1) constructing seed document from web corpora via multi-step filtering and text refinement evaluated by a judge; (2) specifying “requirements + persona” to backtranslate seed document into diverse yet faithful instructions with strict Q&A alignment check; and (3) breaking LLM response limits via teacher-forced Answering that fed seed documents during response to anchor semantics, reduce hallucinations, and preserve tone and integrity. HSS-Synth yields 237k high-quality, diverse instruction-tuning samples that outperform 14 leading baselines on 16 benchmarks. The fine-tuned Qwen3-8B-Base set new SOTA and approached official Qwen3-8B, improving both human preference and knowledge capability without performance seesaws. Extensive experiments demonstrate the HSS-Synth’s robustness and transferability. Our code is publicly available at <https://github.com/pengr/HSS-Synth>.

1 Introduction

The capabilities of large language models (LLMs) hinge on high-quality, diverse data (Peng et al., 2025), yet such data remain scarce and costly (Muennighoff et al., 2023; Liu et al., 2024). Consequently, data synthesis has emerged as a

viable counterpart (Huang et al., 2022; Abdin et al., 2024), achieving proven success on closed tasks—mathematics, code, tool use, instruction following, and tabular data (Yu et al., 2024; Austin et al., 2021; Cai et al., 2023; Wang et al., 2023; Zhao et al., 2025). However, the humanities and social sciences (HSS)—the cornerstone of human society and knowledge—have been largely overlooked. HSS, as open-ended tasks lack verifiable answers and demand nuanced judgment, poses significant challenges for data synthesis. Existing synthesis efforts are fragmented and confined to narrow scenarios, such as creative writing, role-playing, dialogue, and long-context (Wang et al., 2024; Ge et al., 2024; Qian et al., 2025; Bai et al., 2024). Departing from above capability-centric, fragmented efforts and driven by LLM deployment needs (Naveed et al., 2025), we adopt a subject-centric paradigm: using the official taxonomy¹, we define **the first HSS domain system with 14 mainstream fields**² and **conduct the first data synthesis work** across HSS domains. A synthetic HSS sample case see Appendix Table 10.

Along this line, we propose **HSS-Synth**, the first data synthesis pipeline for HSS. Figure 1 illustrates its three stages: 1) **Seed document construction**: Given abundant scale and diversity of web corpora, we construct seed documents from the web. To address noise and the scarcity of ideal HSS texts, we integrate source sampling, heuristic filtering, domain classification, and quality rating based on 12 expert-crafted rubrics to select HSS documents that satisfy quality standards. To handle residual noise, redundancy, and weak expression beyond filtering, we apply LLM-based

¹<https://www.topuniversities.com/subject-rankings>

²In this paper, HSS refers to disciplines studying human behavior, cultural expression, social structures, and institutional mechanisms, including Philosophy, Economics, Law, Politics, Sociology, Healthcare, Geography, Education, Sports, Literature, History, Management, Arts, and Psychology.

*This work was done during internships at Qwen Team, Alibaba Group and Inclusion AI, Ant Group.

†Corresponding authors.

text refinement and verify success with a refinement judge, yielding clean seed documents. 2) **Question-answer pair refinement:** Unlike simple instruction backtranslation (Köksal et al., 2023; Li et al., 2024), we adopt multi-attribute instruction backtranslation that specifies both the ‘what to do’ (requirements) and the ‘who to do’ (persona), generating diverse yet faithful instructions from seed documents. Followed by a Q&A alignment check to strictly enforce consistency between instructions and seed documents, further improving instruction fidelity. 3) **Teacher-forced answering:** Even after Q&A pair refinement, LLM-generated answers may exhibit factual gaps, detail lacks, and style drift (Cao et al., 2025; Jiang et al., 2025). We therefore propose Teacher-Forced Answering (TeachForceA): during answering, the seed document is fed alongside the instruction as semantic anchors, providing required information to reduce hallucination, preserving tone and style to enhance human-touch, and leveraging structural and lexical cues to improve completeness and readability. This breaks the ceiling of LLM-generated answers. Using HSS-Synth, we synthesize an instruction-tuning dataset that meets: (i) *sufficient scale* (237k samples); (ii) *high quality* (strong Q&A alignment while surpassing the model’s native answer limits); and (iii) *rich diversity* (broad HSS coverage with content diversity inherited from web sources and varied personas).

In our experiments, we fairly compare 14 leading data-synthesis baselines across 16 mainstream benchmarks spanning eight core LLM capabilities. Qwen3-8B-Base fine-tuned on HSS-Synth surpasses all strong baselines and sets a new SOTA record, with overall performance closest to the official instruction-tuned Qwen3-8B. Using high-quality, diverse HSS data, HSS-Synth markedly improves human-preference alignment and, to a certain extent, enhances knowledge capabilities, effectively avoiding the “performance seesaw” (knowledge capabilities gains accompanied by human preference drops). We conduct ablation studies to show HSS-Synth’s core components are indispensable; HSS-Synth transfers well across model architectures and sizes (Llama3.1-8B, Qwen2.5-14B); and different LLM capabilities require different synthetic data scales for convergence. In-depth analysis of the HSS-Synth dataset highlights its advantages in sentence length, lexical richness, and semantic diversity. Comparing three answer types or instructions (seed document, reverse, teacher-

forced) shows that teacher-forced is the optimal type; seed documents are unsuitable as answers; and reverse answers underperform due to reuse and insufficient grounding. Finally, we cross-validate the introduced quality rubric using both human judgments and proprietary LLM evaluations, and empirically demonstrate the necessity of text refinement. Our contributions are as follows:

- HSS-Synth is the first to define a 14-field HSS domain system via a subject-centered paradigm and perform data synthesis across these domains.
- The techniques are transferable across model architectures and broader data-synthesis tasks.
- HSS-Synth outperforms 14 strong baselines and sets new SOTA on 16 benchmarks, improving both human preference and knowledge capabilities without a “performance seesaw.”

2 Related Work

Humanities and Social Sciences for LLM are central to human society and knowledge, so a strong HSS capability is essential for deploying LLMs. Yet HSS has been underemphasized in LLM research: although ValueByte-AI (2025) compiles a reading list for “LLM in Social Science”, it omits data synthesis. As typical open-domain tasks, HSS problems lack verifiable answers and often require nuanced human judgment, making data collection and synthesis difficult. Despite existing efforts, these work remains fragmented and limited to a few areas: *writing* (simple story generation (Eldan and Li, 2023), *creative writing* (Wang et al., 2024)); *dialogue systems* (task-oriented dialogue for e-commerce (Qian et al., 2025), few-shot dialogue summarization (Lu et al., 2025), multi-turn multi-topic dialogue (Lee et al., 2025)); *role-playing* (billion-scale persona construction (Ge et al., 2024)); and *long context* (ultra-long content generation (Bai et al., 2024)). To address this limitation, unlike the above ability-centric work, we adopt a subject-centric, application-oriented way: drawing 14 mainstream HSS domains from an official taxonomy, we build a comprehensive HSS taxonomy for LLMs and, for the first time, systematically study data synthesis in these domains.

Synthesizing Instruction Tuning Data has evolved and falls into three categories: (1) *Human-crafted methods*, e.g., WildChat (Zhao et al., 2024), which generate multi-turn human-GPT conver-

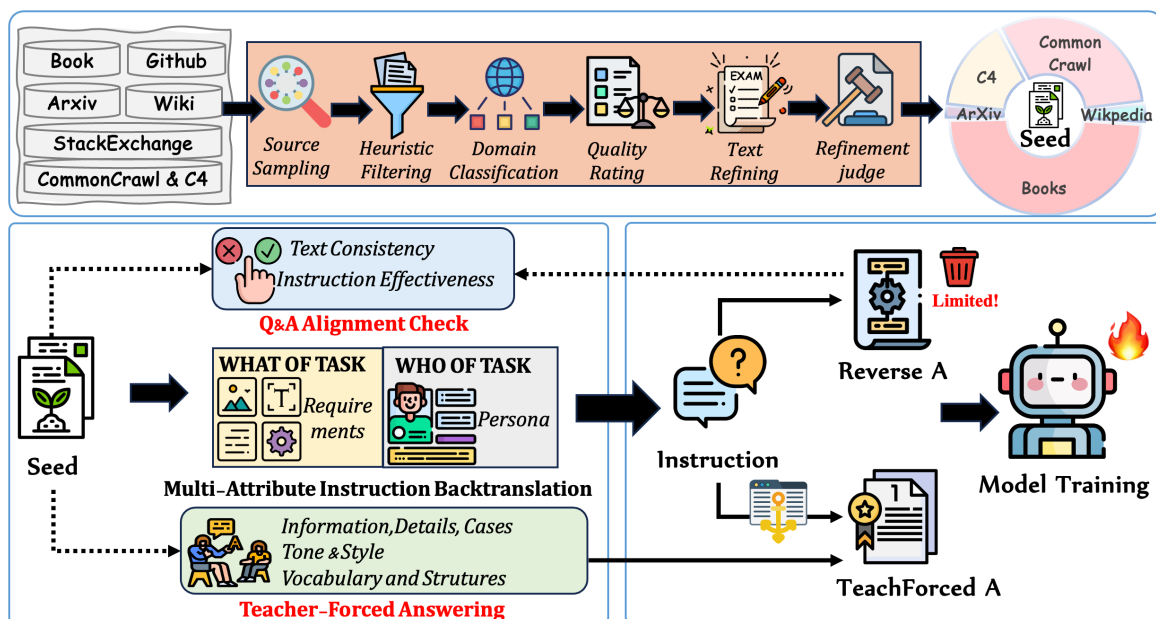


Figure 1: Overview of the **HSS-Synth** pipeline: (1) Seed document construction: sample, filter, classify, rate, and refine web texts under judge into clean HSS seeds; (2) Question-Answer Pair Refinement—backtranslate instructions from seed with multiple attributes, specify the “what” and “who,” and enforce Q&A alignment checks to yield diverse, faithful Q&A pairs; (3) Teacher-forced Answering—anchor answers on seeds to better response beyond the LLM’s inherent limits. Finally, we train models on the resulting instruction-tuning pairs.

sations; (2) *Semi-automated methods*, including Evol-Instruct (Xu et al., 2024a) and SynthQuestions (Zhu et al., 2025), which expand manually annotated instructions via few-shot prompting; Nemotron-CC-HQ (Su et al., 2024) that extracts knowledge and QA pairs; Bonito (Nayak et al., 2024) that remixes task templates to create meta-templates for conditional generation; LongWriter (Bai et al., 2024), which uses agent-based planning for writing; Wrap (Maini et al., 2024) and MGACorpus (Hao et al., 2025), both rephrasing documents, with the latter focusing on genre and audience; LongForm (Köksal et al., 2023) and Back-translation (Li et al., 2023a) that leverage instruction inversion; and WebR (Maini et al., 2024), which combines document rephrasing and instruction inversion; (3) *Fully automated methods*, e.g., Magpie (Xu et al., 2024b), which prompts LLMs with chat templates; Cosmopedia-v2 (Ben et al., 2024), which conducts web-based rewriting on specific topics; and Condor (Cao et al., 2025), which expands world knowledge through self-refinement. Additionally, *mixed datasets*, e.g., OpenHermes 2.5 (Teknium, 2023) aggregate multiple open-source sources. Building on these advances, our semi-automated HSS-Synth incorporates novel mechanisms such as QA consistency checking and teacher-forced answering for HSS data synthesis.

3 HSS-Synth

This section presents HSS-Synth in Figure 1, a three-stage pipeline for generating high-quality, diverse HSS datasets, consisting of: (1) seed document construction, (2) question-answer pair refinement, and (3) teacher-forced answering.

3.1 Seed Document Construction

Given the abundant scale and diversity of web corpora (Yue et al., 2024; Jiang et al., 2025), we construct seed documents from the 627B-token Slimpajama corpus (Soboleva et al., 2023). To address noisy web data and the scarcity of ideal HSS texts, we adopt a multi-step workflow: We begin with **source sampling**, exclude HSS-irrelevant sources (StackExchange, GitHub), retain all documents from high-quality sources (Books, ArXiv, and Wikipedia), and randomly sample 10% from web-crawled sources (C4, CommonCrawl) to balance source composition. Next, we apply **heuristic filtering** via the FineWeb toolkit³, integrating fast-Text language classifier (Joulin et al., 2016), Gopher/C4/FineWeb quality filters (Rae et al., 2021; Raffel et al., 2020; Penedo et al., 2024), and Min-Hash deduplication (Broder, 1997) to eliminate noisy and redundant documents. Then, we use an LLM for **domain classification** to identify texts

³<https://github.com/huggingface/datatrove/blob/main/examples/fineweb.py>

that fall within HSS domains. Subsequently, we assess HSS texts using 12 expert-crafted quality rubrics organized into three tiers: *readability* (grammar, coherence, content accuracy, domain relevance), *applicability* (tone & expression, knowledge depth, vocabulary richness, genre focus), and *human touch* (thematic depth, emotionality, literary diversity, humanities creativity). The validity of these rubrics is empirically supported in Section 4.6. We prompt the LLM for **quality rating** and retain documents with $\text{readability} = 5$, $\text{applicability} \geq 4$, and $\text{human touch} \geq 3$. However, residual noise, redundancy, or weak expression may remain, so we further **refine texts** with an LLM to achieve: i) content cleaning (removing crawl artifacts and irrelevant redundancy); ii) content fidelity (preserving core information, structure, and “human-touch”); iii) expression optimization (improving coherence, fluency, and accuracy). Finally, a **refinement judge** verifies that refined documents satisfy these objectives and are suitable as clean seed documents. See Appendix A for details.

3.2 Question-Answer Pair Refinement

Multi-attribute Instruction Backtranslation

Unlike prior instruction backtranslation work (Köksal et al., 2023; Li et al., 2024), we generate instructions from seed documents with attached multiple attributes. We first specify the “what” of the task: concrete, actionable requirements that guide the LLM to generate instructions that faithfully reproduce the seed document. Specifically, the instruction must (i) state the document’s domain, genre, and length; (ii) summarize the document’s core content and key points; (iii) describe the document’s structure and narrative voice; and (iv) remain concise, non-redundant, and avoid direct references such as “source document”. We then define the “who” of the task—the persona the instruction should adopt to enhance diversity. To this end, we introduce persona settings (Ge et al., 2024) that specify stance, mindset, and tone, and uniformly use second-person phrasing. Notably, personas are co-generated with the instructions, requiring no pre-defined setup and thus offering greater flexibility. This multi-attribute backtranslation enables us to produce instructions that are faithful to the original documents while diverse in content.

Question-Answer Alignment Check Although instruction backtranslation is effective, it may introduce alignment bias—i.e., the reverse instruction (*question*) is misaligned with the seed doc-

ument (*seed answer*)—which violates the strict Q&A alignment required for instruction tuning, a critical issue overlooked by prior work. To address this, we propose a Q&A alignment check: an LLM first answers the reverse instruction to obtain a *reverse answer*, then we measure the consistency between the seed answer and the reverse answer to verify whether the reverse instruction can faithfully reproduce the seed document. The checker yields binary decisions with rationales along two dimensions: (1) textual consistency—alignment of core content and key points; and (2) instruction effectiveness—covering essential information without explicit references. We retain only Q&A pairs that pass both checks, further improving the fidelity of reverse instructions.

3.3 Teacher-Forced Answering

Even after question-answer pair refinement yields usable reversed instruction-answer pairs, reverse answers generated solely by an LLM may exhibit factual gaps, sparse detail, and style drift (Cao et al., 2025; Jiang et al., 2025). Hence, we propose teacher-forced answering (TeachForceA), inspired by teacher forcing—feeding external ground-truth inputs into the model at each step (Lamb et al., 2016). Specifically, when answering an instruction, the seed document (i.e., information source) is fed into the LLM alongside the instruction as a semantic anchor. The seed document need not outperform the reverse answer; it only needs to (1) cover the information, details, and examples required by the instruction; (2) convey the original tone and style; and (3) supply key vocabulary and sentence structures in certain passages. We also explicitly require the LLM that, in case of conflict, the faithful answer to the instruction prevails, preserving strict Q&A consistency. By confining information integration to the controlled set “reversed instruction + seed document,” TeachForceA (i) exploits undistorted facts from the seed document to reduce hallucinations; (ii) transfers the original tone and style to enhance human touch; and (iii) uses structural and lexical cues to improve completeness and readability—thereby surpassing the ceiling of traditional LLM-distilled answers.

4 Experiment

4.1 Experimental Setup

Baselines We compare the instruction tuning (IT) dataset generated by HSS-Synth with 14 leading open-source IT datasets, all containing 230k

Selected Method	Synthesis Model	Human Preference				Knowledge based				AVG
		Writing	Emotion	Social	Instruct	World	Commonsense	LongCtx	Comprehension	
Qwen3-8B (Yang et al., 2025)	–	64.52	49.20	66.63	56.13	74.90	76.49	47.65	35.22	60.70
Qwen3-8B-Base (Yang et al., 2025)	–	31.95	25.10	61.91	31.18	77.07	76.89	46.24	32.76	50.73
WildChat (Zhao et al., 2024)	ChatGPT	32.37	40.52	63.25	38.96	77.01	79.70	45.50	35.97	51.66
OpenHermes 2.5 (Teknium, 2023)	Mixed	41.04	38.24	69.16	35.77	77.16	79.02	50.93	30.71	52.75
SynthQuestions (Zhu et al., 2025)	Llama3-70B-Instruct	40.55	37.34	68.63	35.08	77.01	79.04	51.29	37.90	53.36
Bonito (Nayak et al., 2024)	Mistral-7B	22.89	3.42	72.94	20.05	75.80	79.05	45.76	34.89	44.35
Evol Instruct (Xu et al., 2024a)	ChatGPT	29.38	31.86	62.40	39.16	77.28	80.21	49.91	37.48	50.96
LongWriter (Bai et al., 2024)	GPT-4o	39.82	39.91	38.78	27.44	77.01	80.39	44.07	15.83	45.41
Magpie (Xu et al., 2024b)	Mixed	42.59	40.17	68.54	42.90	77.50	80.06	48.74	36.3	54.59
Nemotron-CC-HQ (Su et al., 2024)	Mistral-NeMo-12B-Ins.	29.29	28.03	58.98	31.40	76.54	79.95	48.72	26.19	47.39
MGACorpus (Hao et al., 2025)	3.3B MoE	36.02	36.04	66.43	36.75	76.89	77.91	51.51	31.79	51.67
Cosmopedia-v2 (Ben et al., 2024)	Mixtral-8x7B-Instruct	36.61	30.67	65.39	35.10	77.41	79.95	48.45	32.84	50.71
Condor (Cao et al., 2025)	Qwen2.5-72B	39.49	44.66	65.73	27.31	77.28	80.07	43.31	24.16	50.25
WebR-Pro (Jiang et al., 2025)	GPT-4o-mini	49.50	45.36	66.55	43.25	77.09	79.93	46.35	35.92	55.49
WRAP (Maini et al., 2024)	Qwen-30B-A3B	21.45	14.63	55.90	19.90	76.74	78.53	45.43	27.35	42.50
LongForm (Köksal et al., 2023)	Qwen-30B-A3B	21.18	6.30	35.29	10.89	76.85	79.23	40.72	30.13	37.57
HSS-Synth (Ours)										
-w/ MA-IBT	Qwen3-30B-A3B	56.71	46.02	58.97	38.67	77.17	78.37	28.37	25.83	51.26
-w/ MA-IBT & QAC	Qwen3-30B-A3B	58.86	51.28	65.86	38.67	77.24	77.77	47.24	24.86	55.22
-w/ MA-IBT & QAC & TeachForcedA	Qwen3-30B-A3B	60.01	51.31	67.67	47.70	76.81	77.86	48.71	32.62	57.84

Table 1: 16 Benchmark results covering 8 core capabilities for Qwen3-8B-Base trained on instruction-tuning data synthesized by different methods. HSS-synth surpasses all baselines (SOTA), nearly matches the official instruct-tuned Qwen3-8B, and improves both human preference and knowledge without performance trade-offs. Here, the abbreviation MA-IBT, QAC denotes multi-attribute instruction backtranslation, Q&A alignment check.

instruction–response pairs⁴, grouped as follows: (1) Human-crafted, e.g., **WildChat** (Zhao et al., 2024); (2) Mixed data, such as **OpenHermes 2.5** (Teknium, 2023); (3) Semi-automated synthetic data, including **Bonito** (Nayak et al., 2024), **Evol-Instruct** (Xu et al., 2024a), **Nemotron-CC-HQ** (Su et al., 2024), **LongWriter** (Bai et al., 2024), **MGACorpus** (Hao et al., 2025), **WebR** (Jiang et al., 2025), and **SynthQuestions** (Zhu et al., 2025) (with LongWriter focusing on writing tasks); (4) Fully automated synthetic data, such as **Magpie** (Xu et al., 2024b), **Cosmopedia-v2** (Ben et al., 2024), and **Condor** (Cao et al., 2025). To ensure fair comparison, we reproduce two representative methods—document rephrasing (**Wrap** (Maini et al., 2024)) and instruction inversion without persona (**LongForm** (Köksal et al., 2023))—using the same seed data and synthesis models as ours.

Data Synthesis Settings Throughout our data synthesis pipeline from large-scale web corpora, we consistently used Qwen3-30B-A3B (Yang et al., 2025), chosen because it offers both strong performance (see Table 1) and cost efficiency (deployable on a single H800 GPU), with the default generation configuration (temperature=0.6, top-k=20, top-p=0.95). After filtering out improper JSON formats and removing <thinking> content, we obtained 230k instruction–answer pairs for fine-tuning. The data synthesis process consumed 2,400 GPU hours.

⁴We use all data if a dataset has fewer than 230k pairs.

Model Training Settings For instruction tuning, we trained Qwen3-8B-base, Qwen2.5-14B-base (Qwen Team, 2024), and LLaMA3.1-8B-base (Dubey et al., 2024) on various datasets, using the same hyperparameters for fair comparison. Full training details are provided in Appendix D.

Evaluation Benchmarks and Metrics We evaluate fine-tuned models on 16 mainstream benchmarks spanning 8 core LLM capabilities, grouped into *human-preference* and *knowledge-based* tasks. For each benchmark, we report results under its default setting (0-shot unless specified), along with the metric and judge model when applicable. **Writing Skill:** WritingBench (rubric score, Claude-3.7-Sonnet) (Wu et al., 2025), CreativeWriting-v3 (rubric score, Claude-3.7-Sonnet), and Judgemark-v2 (Judgemark score) (Paech, 2023); **Emotional Perception:** BuzzBench (rubric score, Claude-3.5-Sonnet-v2) and EQ-Bench3 (rubric score, Claude-3.7-Sonnet) (Paech, 2025); **Social Interaction:** Social-IQA (accuracy) (Sap et al., 2019), IQuiz_EQ (accuracy) (Chen et al., 2024b); **Instruction Following:** IFEval (prompt-level strict accuracy) (Zhou et al., 2023), Collie (aggregate accuracy) (Yao et al., 2023); **World Knowledge:** MMLU (5-shot, accuracy; we also report result on the HSS subset) (Hendrycks et al., 2020); **Commonsense Reasoning:** HellaSwag (normalized accuracy) (Zellers et al., 2019), StoryCloze (accuracy) (Mostafazadeh et al., 2016); **Long-Context:** CoQA (F1) (Reddy et al., 2019), Gov-Report (ROUGE-L) (Huang et al., 2021); **Reading**

Selected Method	Human Preference			Knowledge based				AVG	
	Writing	Emotion	Social	Instruct	World	Commonsense	LongCtx		Comprehension
Qwen2.5-14B (Qwen Team, 2024)	23.94	31.17	68.87	32.26	80.45	82.06	51.33	37.46	49.70
Wrap (Maini et al., 2024)	20.60	22.58	62.76	19.32	80.06	81.97	46.36	19.20	44.11
LongForm (Köksal et al., 2023)	15.09	9.49	60.61	25.30	79.89	80.22	47.30	28.91	43.35
HSS-Synth (Ours)	57.50	50.95	65.59	46.70	80.22	80.79	49.73	33.52	57.00
Llama3.1-8B (Dubey et al., 2024)	9.17	8.41	35.91	17.95	66.93	79.97	43.11	14.04	34.44
Wrap (Maini et al., 2024)	8.14	3.47	38.11	6.33	65.89	80.86	37.72	27.02	32.19
LongForm (Köksal et al., 2023)	18.46	5.58	41.71	11.81	66.77	80.40	40.86	30.45	36.95
HSS-Synth (Ours)	39.95	46.45	50.46	24.71	66.14	79.26	46.63	28.37	47.75

Table 2: Performance comparison for different methods across model architectures (Llama-3.1-8B) and scales (Qwen2.5-14B), where HSS-Synth achieves the best cross-model results.

Comprehension: NarrativeQA (accuracy, Llama-3.3-70B-Instruct) (Kočíský et al., 2018), XSum (ROUGE-L) (Narayan et al., 2018). To ensure reliable evaluation, we perform a standard 13-gram decontamination analysis and find zero overlap between the synthesized data and the benchmark test sets.

4.2 Main Results

We report benchmark results of Qwen3-8B-Base fine-tuned on different instruction-tuning datasets in Table 1, from which we draw the insights below:

Synthetic baselines are strong Most synthetic-data baselines outperform the base model. Notably, WebR-Pro—synthesized from web corpora—stands out, validating web-based synthesis as a promising path, consistent with the design intent of HSS-Synth.

Synthesize model capability and data domain coverage matter Bonito relies solely on Mistral-7B for synthesis and underperforms, confirming that data quality hinges on the synthesis model’s capability. In contrast, using only Qwen3-30B-A3B as our synthesis model already matches many strong baselines, evidencing the effectiveness of HSS-Synth. Moreover, LongWriter focuses narrowly on writing-domain data and is therefore limited to most tasks, whereas HSS-Synth benefits from broad coverage across HSS domains.

Advanced text refinement and multi-attribute instruction backtranslation Under the same settings, WRAP and LongForm apply only basic document rephrasing and instruction backtranslation and thus lag behind, highlighting the advantage of our stronger text refinement and multi-attribute instruction backtranslation techniques.

HSS-Synth attains SOTA without a performance seesaw Qwen3-8B-base fine-tuned on HSS-Synth achieves the best average across 16

mainstream benchmarks and is overall closest to the official instruction-tuned Qwen3-8B. Specifically, HSS-Synth delivers large gains on human-preference evaluations; on knowledge-oriented benchmarks, improvements in instruction following are most pronounced, while other capabilities are on par with or superior to the official base/instruct models. This indicates that high-quality HSS data not only strengthens human preference but also, to some extent, enhances knowledge-intensive performance. Crucially, HSS-Synth balances preference and knowledge, effectively avoiding the “performance seesaw,” wherein gains on knowledge capabilities are accompanied by drops in human preference (Dubois et al., 2024).

4.3 Ablation Study

Core Components Table 1 reports the performance of HSS-Synth under different component settings: (1) **w/ MA-IBT**: compared to LongForm (i.e., without MA-IBT), multi-attribute instruction backtranslation clarifies task attributes (what, who), yielding faithful and diverse instructions, so synthetic data becomes usable and surpasses the base model. (2) **w/ MA-IBT & QAC**: adding Q&A alignment checks further improves backtranslated instruction faithfulness, significantly boosting data quality to lead performance gains. (3) **w/ MA-IBT & QAC & TeachForcedA**: teacher-forced answering overcomes LLM’s inherent response constraints, achieving the best performance. Overall, these components are validated as essential and indispensable to the pipeline.

Model Structures and Sizes Table 2 compares, across model architectures and scales, models trained on HSS-Synth synthetic data, the official Qwen3-base, and two faithfully reproduced baselines (Wrap, LongForm). HSS-Synth consistently outperforms the official base and both baselines,

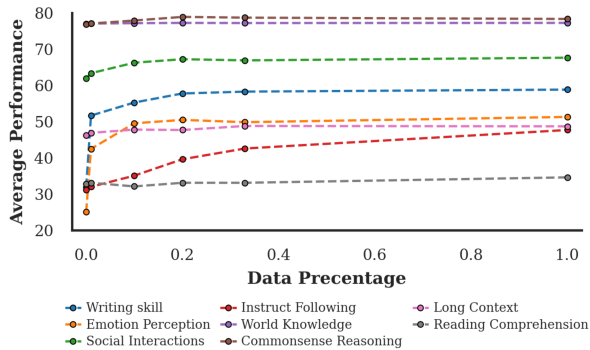


Figure 2: Impact of HSS-Synth data scale on 8 LLM capabilities. Knowledge skills converge at $\approx 10\%$, human-preference at $\approx 20\%$, and instruction following at $\approx 33\%$ of the data (all within tens of thousands of samples). Notably, with no observed cross-skill trade-offs.

with its advantage most pronounced in human-preference evaluations; on knowledge-based benchmarks, it shows the smallest gap to Qwen2.5-14B while clearly outperforming Llama-3.1-8B, demonstrating its strong cross-model applicability.

Synthetic Data Scales Figure 2 illustrates the impact of HSS-Synth data scaling on different LLM capabilities. We plot scaling curves for eight skills and estimate their convergence data requirements. As the data ratio increases, knowledge-based abilities—except instruction following—exhibit diminishing returns around 10% ($\approx 23k$ samples), reflecting that SFT primarily strengthens instruction following rather than injecting new factual knowledge. Human-preference skills plateau at around 20% ($\approx 46k$ samples), indicating that the base model already encodes relevant concepts and requires limited data for style alignment. Instruction following benefit from diversity up to about 33% ($\approx 76k$ samples) before saturating. Overall, for strong base models like Qwen3-8B, convergence typically occurs within tens of thousands of samples, with marginal gains beyond. Notably, we observe no cross-skill trade-offs: human preference gains do not harm knowledge skills and even slightly improve them.

4.4 Data Analysis

Data Length and Lexical Richness Table 3 summarizes average input/output token counts and measures of textual lexical diversity (MTLD) (McCarthy and Jarvis, 2010) for instruction-tuning datasets. Tokens are counted with the Qwen3 tokenizer. MTLD equals total tokens divided by the number of factors, where a factor forms when the type/token ratio (TTR) falls below a thresh-

Selected Method	Avg. Input Len	Avg. Output Len	MTLD
WildChat	400	422	70.78
OpenHermes 2.5	136	214	53.65
Nemotron_CC_HQ	3,025	401	94.82
MGACorpus	1,330	779	92.45
CosmopediaV2	332	687	158.40
Bonito	2,634	134	89.64
Evlo-instruct	123	355	56.63
WebR Pro	464	460	82.51
SynthQuestions	83	762	67.28
LongWriter	282	4,589	124.52
WRAP	172	487	62.95
LongForm	120	2,696	108.41
Magpie	33	440	70.63
Condor	70	876	97.55
HSS-Synth (Ours)	253	1,635	127.50

Table 3: Average input/output length and measure of textual lexical diversity (MTLD) of various instruction-tuning datasets.

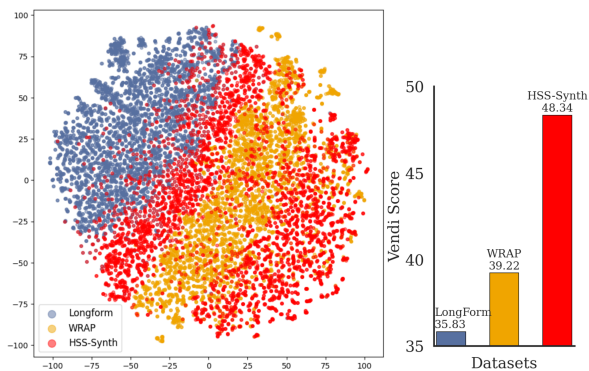


Figure 3: t-SNE visualization (left) and Vendi scores (right) of sentence embeddings for LongForm, HSS-Synth, and WRAP.

old; higher MTLD indicates greater lexical diversity. HSS-Synth back-translates instructions from web-corpus seed documents, yielding input lengths comparable to web-based WebR-Pro and WildChat. Benefiting from TeachForcedA, HSS-Synth averages over 1.6k tokens per answer, meeting the length needs of HSS domains. Encouragingly, while providing long outputs, HSS-Synth attains the second-highest MTLD, only behind the synthetic dataset Cosmopedia2 designed for pre-training, highlighting the rich lexical signals it offers for model fine-tuning.

Semantic diversity of Dataset We assess the semantic diversity of instruction-answer pairs produced by three data synthesis methods. From each dataset, we randomly sample 10k instances, feed them into gte-Qwen2-7B-instruct (Li et al., 2023b) to compute sentence embeddings, and project them to a 2D space using t-SNE. In the left panel of Figure 3, HSS-Synth exhibits a wider spread, while LongForm and WRAP are more clustered. We also report the Vendi Score (Dan Friedman and Ding, 2023), which constructs a Gram matrix from

	Rubric Score	Avg. Sent Len	MTLD
Seed Doc	50.75	2,144	114.56
Reverse A	54.29	2,621	123.50
TeachForced A	54.89	1,660	127.93

Table 4: Quality rubric score, average sentence length, and MTLT for three answers to the reverse instruction.

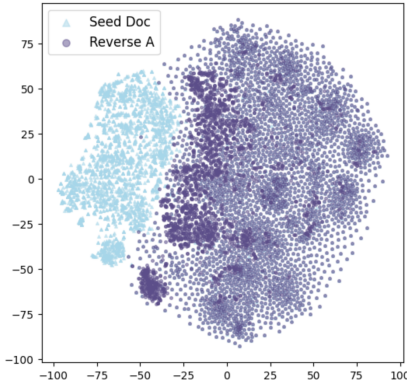


Figure 4: t-SNE plot of sentence embeddings for reverse answers and seed documents.

embedding similarities and measures diversity via the von Neumann entropy of its eigenvalues. In the right panel of Figure 3, HSS-Synth attains the highest Vendi Score among all synthetic datasets. Both quantitative and qualitative results indicate that, under the same seeds and synthesis model, HSS-Synth produces more semantically diverse data, providing richer training signals and thereby improving model generalization.

4.5 Evaluation of Different Answers

Rubric scores, sentence length, lexical diversity, and semantic embeddings Table 4 reports quality rubric scores, average sentence length, and measure of textual lexical diversity (MTLD) for three types of answers to the same instruction:

1) Teacher-forced answers achieve the highest rubric score and MTLT while being the shortest sentence length, indicating the best performance with the most concise phrasing. Together with component ablations, this further supports teacher-forced as the optimal answer type for instructions.

2) Reverse answers are used as the reference baseline because they are generated directly for the instruction by the synthesis model, inherently ensuring Q&A consistency. They also show high rubric and MTLT scores and perform well under the w/ MA-IBT & QAC ablations.

3) Seed documents are not selected as answers for two reasons: (i) lower quality—their rubric and MTLT scores are clearly below the other two; (ii) distribution shift—t-SNE in Figure 4 shows sen-

	4-gram Overlap	LCS Length	Copy Ratio
Reverse A	25.11	31.35	0.42
TeachForced A	27.35	64.77	0.78

Table 5: Reuse metrics from seed documents for reverse and teacher-forced answers.

tence embeddings of seed documents (blue) and reverse answers (red) forming two distinct clusters with an average cosine similarity of only 0.20 ± 0.05 , indicating substantial divergence. This indicates a distribution shift between seed documents and reverse answers, and using seed documents as answers may induce Q&A mismatch; instead, they are better suited as factual sources, consistent with recent findings (Zhu et al., 2025; Jiang et al., 2025).

Inference log-likelihood heatmaps We train two models with (instruction, seed document) and (instruction, reverse answer) pairs, and visualize inference token-level log-likelihood on the validation set in the Appendix’s heatmaps 5,6. Darker colour denotes tokens that the model fails to predict confidently. The model trained on reverse answers is predominantly green, suggesting the answer pattern is easy to learn, whereas the model trained on seed answers exhibits frequent red tokens, indicating harder learning and potential errors.

Reuse metrics from seed documents To quantify how much reverse and teacher-forced answers are extracted from seed documents, we randomly sampled 10k instances and computed three reuse metrics: i) average 4-gram overlap, ii)- longest common substring (LCS) length, iii)- length-normalised copy ratio = $2 \times \text{LCS} / |\text{answer}|$. As shown in Table 4, teacher-forced answers outperform reverse answers on all token- and substring-level reuse metrics, indicating they anchor on seed documents to extract original statements rather than generating answers from scratch like reverse answers. This strong textual grounding also explains its superior performance.

4.6 Validating Quality Rubric Effectiveness

Here we validated the expert-crafted HSS quality rubric on 400+ randomized seed texts spanning diverse sources, lengths, domains, and quality tiers. Twenty uninvolved raters independently scored each item on a five-point scale, and Qwen3-30B-A3B and GPT-4.1 also applied the rubric. The synthesis model matched majority-vote human labels in 86% of cases and showed 91% agreement with the GPT-4.1. Inter-rater agreement, summarized by Cohen’s kappa (McHugh, 2012), indicated

strong consistency among human raters. Overall, cross-checking human judgments and LLM outputs provides strong evidence of the rubric’s validity.

5 Conclusion

In this paper, we introduce **HSS-Synth**, the first method for synthesizing humanities and social sciences data for LLMs. Our three-stage pipeline successfully generated over 230k high-quality, diverse samples across 14 HSS domains, significantly improving model fine-tuning performance. Comprehensive evaluations, ablation studies, and in-depth analyses validate its effectiveness. This work not only alleviates the scarcity of HSS data but also advances the field of open-domain data synthesis.

Limitations

HSS-Synth’s three-stage pipeline depends on the synthesis model and may introduce LLM bias, a common limitation of current data synthesis methods. The synthetic dataset currently covers 14 mainstream HSS domains (e.g., education, economics, law) but misses subfields such as anthropology and religious studies, and its domain distribution shows a pronounced long tail and imbalance; further work should improve domain coverage and balance. For commonsense reasoning, long-context understanding, and reading comprehension, model fine-tuning yields only marginal gains, indicating that most knowledge is injected during pre-training rather than SFT; we therefore plan to explore pre-training data synthesis and the associated scaling laws to more effectively strengthen model capabilities.

Ethical Considerations

All authors attest that this work strictly adheres to the ACL Code of Ethics. Below, we summarize the ethical considerations specific to this study. First, we report the limitations of this work and ensure that it does not pose any potential risks. Second, all code, data, and models used are open source, released under the Apache-2.0 license for academic research, and contain no discrimination, personally identifiable information, or other harmful content. Furthermore, we provide detailed reporting of our computational experiments and the procedures involving human annotators. Finally, we used AI assistants solely to support the language of the paper—paraphrasing or polishing the authors’ original content.

Acknowledgements

This work was mainly supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62402424.

References

- Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.
- Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and 1 others. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.
- Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. Longwriter: Unleashing 10,000+ word generation from long context llms. *arXiv preprint arXiv:2408.07055*.
- Loubna Ben, Anton Lozhkov, Guilherme Penedo, Thomas Wolf, and Leandro von Werra. 2024. *Cosmopedia*.
- Andrei Z Broder. 1997. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, pages 21–29. IEEE.
- Tianle Cai, Xuezhi Wang, Tengyu Ma, Xinyun Chen, and Denny Zhou. 2023. Large language models as tool makers. *arXiv preprint arXiv:2305.17126*.
- Maosong Cao, Taolin Zhang, Mo Li, Chuyu Zhang, Yunxin Liu, Haodong Duan, Songyang Zhang, and Kai Chen. 2025. Condor: Enhance llm alignment with knowledge-driven data synthesis and refinement. *arXiv preprint arXiv:2501.12273*.
- Yongrui Chen, Haiyun Jiang, Xinting Huang, Shuming Shi, and Guilin Qi. 2024a. *DoG-instruct: Towards premium instruction-tuning data via text-grounded instruction wrapping*. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4125–4135, Mexico City, Mexico. Association for Computational Linguistics.
- Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and 1 others. 2024b. Tombench: Benchmarking theory of mind in large language models. *arXiv preprint arXiv:2402.15052*.
- Dan Dan Friedman and Adji Bousso Dieng. 2023. The vendi score: A diversity evaluation metric for machine learning. *Transactions on machine learning research*.

- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.
- Ronen Eldan and Yuanzhi Li. 2023. Tinstories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*.
- Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.
- Xintong Hao, Ruijie Zhu, Ge Zhang, Ke Shen, and Chenggang Li. 2025. Reformulation for pretraining data augmentation. *arXiv preprint arXiv:2502.04235*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.
- Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*.
- Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436.
- Yuxin Jiang, Yufei Wang, Chuhan Wu, Xinyi Dai, Yan Xu, Weinan Gan, Yasheng Wang, Xin Jiang, Lifeng Shang, Ruiming Tang, and 1 others. 2025. Instruction-tuning data synthesis from scratch via web reconstruction. *arXiv preprint arXiv:2504.15573*.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, H erve J egou, and Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.
- Tom as Ko cisk y, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, G abor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.
- Abdullatif K oksal, Timo Schick, Anna Korhonen, and Hinrich Sch utze. 2023. Longform: Effective instruction tuning with reverse instructions. *arXiv preprint arXiv:2304.08460*.
- Alex M Lamb, Anirudh Goyal ALIAS PARTH GOYAL, Ying Zhang, Saizheng Zhang, Aaron C Courville, and Yoshua Bengio. 2016. Professor forcing: A new algorithm for training recurrent networks. *Advances in neural information processing systems*, 29.
- Jing Yang Lee, Hamed Bonab, Nasser Zalmout, Ming Zeng, Sanket Lokegaonkar, Colin Lockard, Binxuan Huang, Ritesh Sarkhel, and Haodong Wang. 2025. Doctalk: Scalable graph-based dialogue synthesis for enhancing llm conversational capabilities. *arXiv preprint arXiv:2507.05750*.
- Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Omer Levy, Luke Zettlemoyer, Jason Weston, and Mike Lewis. 2023a. Self-alignment with instruction back-translation. *arXiv preprint arXiv:2308.06259*.
- Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Omer Levy, Luke Zettlemoyer, Jason E Weston, and Mike Lewis. 2024. Self-alignment with instruction back-translation. In *The Twelfth International Conference on Learning Representations*.
- Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023b. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.
- Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinneng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. 2024. Best practices and lessons learned on synthetic data. In *First Conference on Language Modeling*.
- Yen-Ju Lu, Ting-Yao Hu, Hema Swetha Koppula, Hadi Pouransari, Jen-Hao Rick Chang, Yin Xia, Xiang Kong, Qi Zhu, Simon Wang, Oncel Tuzel, and 1 others. 2025. Mutual reinforcement of llm dialogue synthesis and summarization capabilities for few-shot dialogue summarization. *arXiv preprint arXiv:2502.17328*.
- Pratyush Maini, Skyler Seto, Richard Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. 2024. Rephrasing the web: A recipe for compute and data-efficient language modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14044–14072, Bangkok, Thailand. Association for Computational Linguistics.
- Philip M McCarthy and Scott Jarvis. 2010. Mtl-d, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.
- Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.
- Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories. *arXiv preprint arXiv:1604.01696*.

- Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. 2023. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36:50358–50376.
- Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*.
- Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2025. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72.
- Nihal V Nayak, Yiyang Nan, Avi Trost, and Stephen H Bach. 2024. Learning to generate instruction tuning datasets for zero-shot task adaptation. *arXiv preprint arXiv:2402.18334*.
- Samuel J Paech. 2023. Eq-bench: An emotional intelligence benchmark for large language models. *arXiv preprint arXiv:2312.06281*.
- Samuel J. Paech. 2025. Eq-bench 3: Emotional intelligence benchmark. <https://github.com/EQ-bench/eqbench3>. Commit <hash> or release <tag>.
- Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin A Raffel, Leandro Von Werra, Thomas Wolf, and 1 others. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849.
- Ru Peng, Kexin Yang, Yawen Zeng, Junyang Lin, Dayiheng Liu, and Junbo Zhao. 2025. Dataman: Data manager for pre-training large language models. In *The Thirteenth International Conference on Learning Representations*.
- Kun Qian, Maximillian Chen, Siyan Li, Arpit Sharma, and Zhou Yu. 2025. Bottom-up synthesis of knowledge-grounded task-oriented dialogues with iteratively self-refined prompts. *arXiv preprint arXiv:2504.14375*.
- Qwen Team. 2024. [Qwen2.5: A party of foundation models](#).
- Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, and 1 others. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.
- Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.
- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.
- Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. 2023. [SlimPajama: A 627B token cleaned and deduplicated version of RedPajama](https://cerebras.ai/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-RedPajama). <https://cerebras.ai/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-redpajama>.
- Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Nemotron-cc: Transforming common crawl into a refined long-horizon pretraining dataset. *arXiv preprint arXiv:2412.02595*.
- Teknum. 2023. [Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants](#).
- ValueByte-AI. 2025. [Awesome-llm-in-social-science](https://github.com/ValueByte-AI/Awesome-LLM-in-Social-Science). <https://github.com/ValueByte-AI/Awesome-LLM-in-Social-Science>.
- Tiannan Wang, Jiamin Chen, Qingrui Jia, Shuai Wang, Ruoyu Fang, Huilin Wang, Zhaowei Gao, Chunzhao Xie, Chuou Xu, Jihong Dai, and 1 others. 2024. Weaver: Foundation models for creative writing. *arXiv preprint arXiv:2401.17268*.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. [Self-instruct: Aligning language models with self-generated instructions](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.
- Maurice Weber, Dan Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, and 1 others. 2024. Redpajama: an open dataset for training large language models. *Advances in neural information processing systems*, 37:116462–116492.
- Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Zijia Wang, Ji Zhang, Mengyue Wu, Qin Jin, and 1 others. 2025. Writing-bench: A comprehensive benchmark for generative writing. *arXiv preprint arXiv:2503.05244*.

- Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024a. WizardLM: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations*.
- Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024b. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *arXiv preprint arXiv:2406.08464*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Shunyu Yao, Howard Chen, Austin W Hanjie, Runzhe Yang, and Karthik Narasimhan. 2023. Collie: Systematic construction of constrained text generation tasks. *arXiv preprint arXiv:2307.08689*.
- Longhui Yu, Weisen Jiang, Han Shi, Jincheng YU, Zhengying Liu, Yu Zhang, James Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024. [Metamath: Bootstrap your own mathematical questions for large language models](#). In *The Twelfth International Conference on Learning Representations*.
- Xiang Yue, Tuney Zheng, Ge Zhang, and Wenhui Chen. 2024. Mammoth2: Scaling instructions from the web. *Advances in Neural Information Processing Systems*.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.
- Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. [Wildchat: 1m chatGPT interaction logs in the wild](#). In *The Twelfth International Conference on Learning Representations*.
- Zilong Zhao, Robert Birke, and Lydia Y Chen. 2025. Tabula: Harnessing language models for tabular data synthesis. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 247–259. Springer.
- Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyang Luo. 2024. [LlamaFactory: Unified efficient fine-tuning of 100+ language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.
- Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.
- Chiwei Zhu, Benfeng Xu, Xiaorui Wang, and Zhen-dong Mao. 2025. From real to synthetic: Synthesizing millions of diversified and complicated user instructions with attributed grounding. *arXiv preprint arXiv:2506.03968*.

Appendix

A Detail Process of Seed Document Construction

Web Corpus We start from the 627B tokens SlimPajama corpus (Soboleva et al., 2023)—a cleaned, deduplicated RedPajama derivative (Weber et al., 2024), comprising 590M documents from seven sources: StackExchange, GitHub, CommonCrawl, C4, Books, arXiv, and Wikipedia.

Tokenization and Chunking The web corpus is then tokenized with the Qwen3 tokenizer and, following Abdin et al. (2024); Xu et al. (2024b), split into segments of 200–4k tokens to preserve the common length distribution.

Source Sampling Following Chen et al. (2024a), we discard sources that lack HSS content (StackExchange, GitHub). Among the remaining five sources—CommonCrawl, C4, Books, arXiv, and Wikipedia—we retain all Books, arXiv, and Wikipedia documents due to their typical high quality, and randomly sample 10% from the crawler-based C4 and CommonCrawl to balance source composition. This step yields an initial corpus of 30M documents.

Heuristic Filtering Next, we process the initial corpus with the Fineweb toolkit⁵, which sequentially applies the following heuristic filters:

1. a fastText language classifier to filter out non-English text, using the lid.176⁶ backend model with a confidence threshold of 0.9 (Joulin et al., 2016);
2. GopherRepetitionFilter, GopherQualityFilter (Rae et al., 2021), C4QualityFilter (Raffel et al., 2020), and FineWebQualityFilter (Penedo et al., 2024) to remove heuristically low-quality text.
3. a MinHash deduplicator (Broder, 1997) to de-duplicate documents, configured with n_grams=5, num_buckets=14, and hashes_per_bucket=8 to balance recall and false positives.

Domain Classification Following the official QS subject taxonomy⁷, we partition HSS into 14 core domains: philosophy, economics, law, politics, sociology, healthcare, geography, education, sports,

⁵<https://github.com/huggingface/datatrove/blob/main/examples/fineweb.py>

⁶<https://dl.fbaipublicfiles.com/fasttext/supervised-models/lid.176.bin>

⁷<https://www.topuniversities.com/subject-rankings>

literature, history, management, arts, and psychology. We then use the Qwen3-30B-A3B model to identify documents that fall into these HSS domains, yielding 2.09 million documents.

Quality Rating We introduce 12 expert-crafted quality rubrics for LLM rating of HSS texts. Each rubric is rate 1–5 (definitions in Figure 8,9,10) and grouped into three weighted tiers:

- Readability: grammar, coherence, content accuracy, domain relevance (weight 0.5 each, maximum total 10).
- Applicability: tone& expression, knowledge depth, vocabulary richness, genre focus (weight 1.0 each; maximum total 20).
- Human-touch: thematic Depth, emotionality, literary diversity, humanities creativity (weight 1.5 each; maximum total 30).

The overall rubric score ranges from 0 to 60. Based on tier-specific minimum requirements, we partition the corpus into four quality bands (see Table 6). We set the seed-document threshold at a rubric score of 45, as this balances corpus quality—perfect readability (≥ 5), good applicability (≥ 4), and moderate human-touch (≥ 3)—with scale. A higher threshold would substantially reduce data volume and domain coverage, which is detrimental to subsequent instruction backtranslation. After this stage, 1.07 million documents remain.

Text Refinement Even after multi-step processing, residual noise, redundancy, or weak expression may persist and cannot be remedied by filtering alone. We therefore employ LLM-based text refinement to achieve: i) content cleaning (removing crawl artifacts and irrelevant redundancy); ii) content fidelity (preserving core information, structure, and “human-touch”); and iii) expression optimization (improving coherence, fluency, and accuracy).

Refinement Judge We use an LLM to judge whether a refined document meets the three goals above, and accordingly decide whether it qualifies as a curated seed document.

B Necessity of Text Refinement

Since data filtering alone is insufficient to produce clean seed documents, we introduce LLM-based text refinement. To validate its necessity, we randomly sample 1k instances to compare Q&A alignment before (Filtered-Only) and after text refinement (Filtered+Refined), and further evaluate quality rubric scores on 10k instances.

Quality Band	Rubric Score Range	Tier-specific Min. Thresholds	#Docs (M)	#Tokens (B)	Avg. Tok. Len (K)
Excellent	55–60	Read. \geq 5, Applic. \geq 5, H-Touch \geq 4	0.29	1.10	3.75
Seed	45–54	Read. \geq 5, Applic. \geq 4, H-Touch \geq 3	1.07	3.15	2.94
Usable	31–44	Read. \geq 3, Applic. \geq 3, H-Touch \geq 2	12.14	12.53	1.03
Unusable	\leq 30 or any Readability \leq 2	—	1.50	1.62	1.08

Table 6: The corpus quality bands and their statistics: Rubric Score Range denotes the total rubric score interval for each band; Tier-specific Min. Requirements list the minimum thresholds for the rubric scores across three weighted tiers (Readability, Applicability and Human-touch); #Docs is the number of documents (in thousands); #Tokens is the total token count (in billions); Avg. Tok. Len is the average document length (in thousands of tokens).

Setting	Q&A Alignment	Score Range	
		45–54	55–60
Filtered-Only	938 (93.8%)	7917 (79.2%)	2083 (20.8%)
Filtered+Refined	959 (95.9%)	7758 (77.6%)	2242 (22.4%)
Δ Gain	+21 (2.1%)	-159 (-1.6%)	+159 (1.6%)

Table 7: Comparison of Q&A alignment and quality rubric score distributions before (Filtered-Only) and after text refinement (Filtered+Refined). Q&A Alignment denotes the documents that pass GPT-4.1’s Q&A consistency check; Score ranges 45–54 and 55–60 indicate the documents whose quality rubric scores fall within these bands.

- **Better Q&A alignment:** We backtranslate instructions from seed documents in the Filtered-only and Filtered+Refined sets, and use GPT-4.1 to assess Q&A alignment. The Filtered+Refined set exhibits a markedly lower rejection rate, reducing 21 Q&A misaligned instances in the 1k samples, indicating that text refinement improves the fidelity of instruction backtranslation.
- **Better Quality distribution:** Table 7 reports the distribution of quality rubric scores between Filtered-only and Filtered+Refined. After text refinement, the proportion of excellent documents (score range of 55–60) increases substantially, adding 159 such documents in the 10k samples.

C Domain Distribution of HSS-Synth Dataset

To demonstrate the domain diversity of the instruction-tuning dataset synthesized by HSS-Synth, we present statistics on its domain distribution in Table 8, with a total of 237,340 samples.

D Model Training Details

Our model training is conducted on the Llama-factory⁸ (Zheng et al., 2024) using 64x NVIDIA H800 GPUs. We trained the Qwen3-8B-Base model for three epochs, each taking approximately 1 hour, totaling about 3 hours. The checkpoint with the lowest validation loss was selected as

⁸<https://github.com/hiyouga/LLaMA-Factory>

Domain	#Docs	Proportion (%)
History	56,241	23.70
Arts	35,390	14.91
Politics	28,966	12.20
Literature	27,664	11.66
Philosophy	15,204	6.41
Sociology	13,191	5.56
Economics	13,123	5.53
Law	9,217	3.88
Sports	8,548	3.60
Psychology	8,183	3.45
Management	7,662	3.23
Education	7,641	3.22
Healthcare	5,129	2.16
Geography	1,181	0.50
Total	237,340	100.00

Table 8: Domain distribution of the HSS-Synth dataset.

the final model. Detailed training hyperparameters are listed in Table 9.

E Case Study

We provide a case from the HSS-Synth synthesis pipeline in Table 10, comprising the seed document, reverse instruction, reverse answer, and teacher-forced answer. This case shows that, unlike closed tasks with verified answers, HSS questions (i.e., reverse instruction) admit open-ended responses, such as both the reverse and teacher-forced answers. The reverse instruc-

Hyperparameter	Value
Batch size	512
Learning rate	$7e^{-6}$
Maximum length	5120
Epochs	3
Scheduler	Cosine
Warmup ratio	0.03
Weight decay	0.1
Adam β_2	0.95
Precision	BF16
Random seed	42

Table 9: Training hyperparameters used for Qwen3-8B-Base, Qwen2.5-14B, and Llama3.1-8B.

tion presents a richly specified persona and faithfully reflects the content of the seed document. Compared with the reverse answer (433 tokens), the teacher-forced answer (644 tokens) satisfies each requirement of the instruction more precisely—covering the essay core information (four skills of the Heart IQ model: emotional connection, embodiment, love, and authenticity), the essay organization, and the length constraints (413 tokens)—demonstrating TeachForceA’s strengths in supplying requisite information to reduce hallucination and leveraging structural and lexical cues to improve completeness and readability. Moreover, it preserves the tone and style of the seed document, yielding a more human-touch quality.

F Prompt Template

This section summarizes the prompt templates used at each stage. Figure 7 shows the prompt template for domain classification to classify seed documents into HSS domains. Figures 8, 9, and 10 show the prompt template for quality rating that assesses HSS documents using expert-crafted rubrics. Figure 11 shows the prompt template for text refinement to polish seed documents. Figures 11 and 12 show the prompt template for refinement judgement that determines whether text refinement is successful. Figure 13 shows the prompt template for multi-attribute instruction backtranslation, which backtranslates instructions from seed documents while attaching multiple attributes. Figure 14 shows the prompt template for the question-answer alignment check used to verify that the reverse instructions faithfully reproduce the seed document. Figure 15 shows the prompt template for teacher-forced answering, which produces answers anchored to seed documents that exceed the synthesis model’s response limits.

Text Type	Text
Seed Document	<p>The Heart IQ Model consists of four skills that are developed together. A good analogy is a four legged stool. If each skill represents a leg on the stool, you can appreciate the need for all four legs to be of equal length or the stool becomes unstable and thus unusable. In Heart IQ, we've found a unique way to awaken these skills in our students through an organic, highly experiential approach.</p> <p>When we speak about opening the Heart, we are not referring to the pump that pushes blood around your body! Instead, we are talking about FEELING MORE. Feeling more connected to source, feeling more embodied, feeling more love from those you care about and feeling more power and authenticity in your expression to others.</p> <p>When your heart is closed, you will find it difficult to access your unique guidance system. Instead of hearing the subtle messages of your inner wisdom, you will only hear the loud noise of your conditioned mind. When you learn to listen to your heart (through stillness and a process we call 'Tracking'), you will develop the ability to tell the difference between your authentic longings and those that were given to you by your family, education and culture. Through practice, your true purpose and life's direction is revealed.</p> <p>To know your deepest heart's longing is only part of the journey. What's required next is to follow your heart and take inspired action. In Heart IQ 2122 you will develop the willingness, courage and trust required to take fearless action so that you can manifest the life of your dreams.</p> <p>One of the most rewarding aspects of learning Heart IQ 2122 is how it transforms your relationships with others. Learning to communicate from the heart is a vital life skill that serves all relationships, whether it's with your partner, your customers or your children. Most of us were never taught the critical steps to healthy heart centred communication and conflict resolution. With practice you will develop the skill to express yourself in a way that really lands in the hearts and minds of others while having the tools to resolve differences and create harmony with those you love.</p>
Reverse Instruction	<p>You are a compassionate guide specializing in emotional intelligence, dedicated to helping readers unlock their inner wisdom. Your stance prioritizes holistic self-discovery and authentic living. You approach the topic with empathetic curiosity, balancing practical advice with spiritual insight. Your tone is reassuring and empowering, fostering a sense of possibility while maintaining intellectual rigor.</p> <p>Write a Psychology self-help essay explaining the Heart IQ Model's four interdependent skills using the four-legged stool analogy. Emphasize the concept of 'opening the heart' as emotional connection, embodiment, love, and authenticity. Include the process of 'Tracking' to distinguish authentic desires, the importance of inspired action for manifesting goals, and the role of heart-centered communication in relationships. Organize the essay with a problem-solution structure, starting with the consequences of a closed heart, followed by the model's methodology, and concluding with relational benefits. The final text must be approximately 413 tokens (around 2130 characters), matching the length of the original document.</p>
Reverse Answer	<p>The Heart IQ Model: A Four-Legged Stool for Emotional Resilience</p> <p>A closed heart marked by emotional detachment, fear, or self-censorship leaves us isolated, reactive, and disconnected from our true selves. Without emotional openness, we struggle to form meaningful relationships, pursue authentic goals, or navigate life's challenges with grace. This imbalance weakens our capacity for empathy, creativity, and joy, trapping us in cycles of dissatisfaction. The Heart IQ Model offers a transformative framework to rebuild emotional stability by cultivating four interdependent skills, much like a four-legged stool: opening the heart, tracking desires, inspired action, and heart-centered communication.</p> <p>1. Opening the Heart: The Foundation of Emotional Connection</p> <p>The first leg of the stool is opening the heart, which involves embracing emotional vulnerability, embodiment (physical and emotional presence), love, and authenticity. A closed heart numbs us to our needs and others' experiences, but opening it allows us to feel deeply, connect authentically, and align with our values. This process begins with self-compassion, acknowledging pain without judgment and allowing space for healing. When we open our hearts, we become more attuned to our inner wisdom, creating a foundation for the other skills.</p> <p>2. Tracking: Distinguishing Authentic Desires</p> <p>Tracking is the practice of observing thoughts, emotions, and bodily sensations to discern what truly resonates with our authentic self. In a world of external noise, this skill helps us differentiate between societal expectations and inner calling. For example, a person might track their energy levels after a social interaction to identify whether it aligns with their need for connection or drains their spirit. Tracking is not about suppression but about cultivating awareness, enabling us to make choices that honor our deepest desires.</p> <p>3. Inspired Action: Manifesting Goals with Purpose</p> <p>The third leg, inspired action, bridges intention and manifestation. When we act from a place of heart-centered clarity, our efforts feel aligned with our values, fostering resilience and joy. This is not about forced productivity but about responding to life's callings with curiosity and courage. For instance, someone passionate about art might begin a creative project not out of obligation but because it feels like a natural extension of their being. Inspired action transforms abstract dreams into lived experiences, reinforcing our connection to purpose.</p> <p>4. Heart-Centered Communication: Building Authentic Relationships</p> <p>The final leg, heart-centered communication, emphasizes presence, honesty, and empathy in interactions. It requires listening without judgment, speaking from a place of vulnerability, and honoring the emotional needs of others. This skill strengthens relationships by fostering trust and mutual growth. When we communicate from the heart, we create space for others to do the same, building communities rooted in compassion and understanding.</p> <p>The Interdependence of the Four Legs</p> <p>Like a stool, the Heart IQ Model thrives when all four legs are strong. A closed heart hinders tracking, as fear clouds self-awareness. Without tracking, inspired action lacks direction. And without heart-centered communication, even the most authentic goals remain unshared. Together, these skills form a dynamic system that supports emotional resilience, authenticity, and connection.</p> <p>Relational Benefits: A Heart-Open World</p> <p>When we nurture these skills, relationships deepen. We become more patient, empathetic, and present, fostering environments where others feel safe to open their hearts too. This creates a ripple effect: individual healing expands into collective well-being.</p> <p>In closing, the Heart IQ Model invites us to rebuild our emotional lives not through force, but through gentle, consistent practice. By opening our hearts, tracking our truths, acting with purpose, and communicating with authenticity, we cultivate a life of meaning, connection, and inner peace. The stool stands firm when all legs are nurtured, reminding us that emotional intelligence is not a destination but a journey of continual growth.</p>
Teacher-forced Answer	<p>When the heart is closed, the soul's compass falters. A rigid, guarded heart muffles the quiet whispers of intuition, leaving individuals adrift in a sea of external validation and conditioned fears. Without emotional connection, embodiment, love, and authenticity, life becomes a series of fragmented attempts to fill an unspoken void. Relationships fray, goals feel hollow, and the self dissolves into a shadow of what it might be. This is the cost of a heart unopened: a dissonance between inner truth and outer expression, where authenticity is buried under the weight of societal expectations.</p> <p>The Heart IQ Model offers a pathway to reclamation, anchored in the metaphor of a four-legged stool. Each leg—emotional connection, embodiment, love, and authenticity—must be equally strong to sustain balance. Emotional connection is the foundation, grounding us in the present through mindful presence. Embodiment, the second leg, bridges the mind and body, allowing us to feel rather than merely think. Love, the third, expands our capacity to give and receive without condition, while authenticity, the fourth, ensures our actions align with our deepest values. Together, they form a stable platform for holistic growth.</p> <p>To unlock this balance, the practice of Tracking becomes essential. This is not mere introspection but a deliberate act of stillness, where one distinguishes between the clamor of external influences and the quiet resonance of authentic desire. By sitting with discomfort and curiosity, we learn to recognize the difference between cultural scripts and the soul's true yearning. This clarity paves the way for inspired action—not driven by fear or obligation, but by a vibrant alignment with purpose. When we act from this space, goals transform from distant ideals into lived realities, fueled by courage and trust.</p> <p>Yet the true alchemy of Heart IQ lies in its relational impact. Heart-centered communication, rooted in empathy and vulnerability, dissolves barriers and fosters profound connection. It is not about perfection but presence: listening without judgment, speaking with integrity, and resolving conflict as a collaborative dance rather than a battle. Such practices mend fractured relationships, creating spaces where love can flourish unencumbered by fear.</p> <p>To open the heart is to embrace life's full spectrum: its joys, its sorrows, its infinite possibilities. It is a journey of becoming, not a destination, where each step deepens our capacity to live authentically and connect meaningfully. In this light, the Heart IQ Model is not merely a tool but a testament to the human spirit's resilience—a reminder that when we dare to open, we unlock the wisdom that has always dwelled within.</p>

Table 10: A case from the HSS-Synth synthesis pipeline, comprising the seed document, reverse instruction, reverse answer, and teacher-forced answer.

Prompt Template for Domain Classification

You are an expert in text domain classification. Your task is to classify the given text into **one** of the fourteen predefined domain types, or determine if it falls outside these domains. Accuracy in identifying non-relevant texts is as important as correctly classifying relevant texts.

[Text]: {text}

Predefined Domain Types: [Philosophy, Economics, Law, Politics, Sociology, Health and Nursing, Geography, Education, Sports, Literature, History, Management, Arts, Psychology]

[Instruction]

1. Carefully analyze the main subject matter, methodology, and terminology of the text.
2. Determine if the text's **primary focus** clearly and substantially aligns with **one** of the domains listed above.
3. **If a strong match is found:** Select the **single best-fitting domain** from the list.
4. **Crucially: If the text does not clearly and primarily belong to any of the 14 listed domains** (e.g., it is about natural sciences, technology, engineering, mathematics, a general news report without relevant domain focus, creative writing without clear thematic analysis, a personal blog, technical documentation, etc.), you **MUST** output "None" in the "Domain Types" field. **Do NOT force a classification** into the listed domains if the fit is poor or ambiguous.
5. Rate your confidence in the assigned domain (or the "None" classification) on a scale of 1 (lowest confidence, very uncertain or weak fit) to 5 (highest confidence, very clear and strong fit/non-fit). A low confidence score (1-2) for a chosen domain might indicate it's a borderline case.
6. Output the result **strictly** in the specified JSON format. Generate only the JSON object, with no additional text before or after it.

[Output Format]

```
{  
  "Domain Types": "<Selected Domain or None>",  
  "Confidence": <1-5>  
}
```

Figure 7: Prompt template for domain classification to classify documents into HSS domains.

Prompt Template for Quality Rating (Part I)

You are a professional document quality reviewer; if the text warrants it, do not hesitate to use the full range of scores, including the lowest and highest scores. Evaluate the document based on the following criteria, each with a maximum score of 5 points. Please provide a brief explanation for each criterion and output the final result in JSON format directly, where the score represents the score for each item, and the reason provides a simple explanation. Besides the genre, which is given separately.

- **Grammar, Punctuation, and Spelling:** Assess whether there are noticeable grammatical errors, improper punctuation use, or spelling mistakes in the article. Also, check if there's any garbled text or illogical multilingual mixing.
 - 5: No errors at all.
 - 4: Minor errors only.
 - 3: Some noticeable errors, but understandable.
 - 2: Many errors affecting reading experience.
 - 1: Errors are pervasive and severely impact text quality, or the text includes extensive meaningless repetition, gibberish, or a mix of illogical language, rendering it largely unreadable.
- **Logical Coherence and Fluency:** Examine whether the text has a good logical structure and if the transitions between paragraphs are smooth.
 - 5: Very clear logic and fluent.
 - 4: Mostly logical, some abrupt parts.
 - 3: Overall acceptable but with a few noticeable gaps.
 - 2: Logic is relatively disorganized, making it difficult to follow the author's line of thought, or there is noticeable redundancy or unnecessary repetition that interferes with reading fluency.
 - 1: Completely lacks organization, or due to extreme, meaningless repetition, internal contradictions, or a complete structural breakdown, the text's logic completely fails, rendering it incomprehensible.
- **Content Accuracy:** Ensure information is accurate and tightly centered around the topic.
 - 5: All information is very accurate and closely centered on the main idea.
 - 4: Mostly correct with only a few details needing correction.
 - 3: Some content deviates from the topic or contains small errors.
 - 2: Many important factual errors.
 - 1: Grossly inaccurate.
- **Domain Relevance:** Assess whether the text's content, terminology, concepts, and focus are closely related to the specific academic field it claims or implies (e.g., history, sociology, literature, philosophy, etc.). Predefined Domain Types: [Philosophy, Economics, Law, Politics, Sociology, Health and Nursing, Geography, Education, Sports, Literature, History, Management, Arts, Psychology] The domain of the document is **{Domain}**.
 - 5: Highly relevant. The text fully focuses on the core issues, concepts, and terminology of the specified field, clearly belonging to that field.
 - 4: Good relevance. The main focus is appropriate, but it may contain some marginal content or slight deviations.
 - 3: Moderately relevant. The text touches on the field but may include more unrelated content, or the focus is too broad, lacking specific connections to the field.
 - 2: Weak relevance. The connection to the field is weak or tenuous, with most content seeming unrelated to the topic.
 - 1: Almost non-relevant. The text content has little or no connection to the specified field.
- **Literary Diversity:** Evaluate the diversity and innovation in rhetorical devices (metaphors, personification, exaggeration, etc.), expression methods (narrative, exposition, argumentation, etc.), article structure (transition, flashback, etc.), narrative perspective, and person.
 - 5: The work shows high diversity and innovation in various aspects, making the text rich and vivid.
 - 4: Good diversity in most areas with some slight monotony or traditionalism.
 - 3: Some diversity but overall rather plain, lacking enough innovation.
 - 2: Poor diversity, and both rhetorical devices and expression methods are monotonous, with traditional structure and perspective.
 - 1: Almost no diversity, rhetorical devices and expression methods are very monotonous, and the structure and perspective are dull.

(Continue in the next page)

Figure 8: Prompt template for quality rating that assesses HSS documents using expert-crafted rubrics (Part I).

Prompt Template for Quality Rating (Part II)

- **Vocabulary Richness:** Check if the article uses varied vocabulary instead of repeatedly using the same words.
 - 5: Extremely diverse vocabulary.
 - 4: Mostly rich vocabulary.
 - 3: Basically sufficient but somewhat monotonous.
 - 2: Quite limited.
 - 1: Extremely limited vocabulary; heavy reliance on a very restricted set of words, or features extreme, meaningless repetition of words/phrases, completely sacrificing expressive richness.
- **Knowledge Depth and Breadth:** Measure the degree of expertise and detail in the background information provided.
 - 5: Provides a wealth of in-depth and detailed related knowledge.
 - 4: Quite comprehensive information but lacks depth.
 - 3: Basic introduction is relatively sufficient.
 - 2: Overly brief.
 - 1: Almost no additional information.
- **Theme Exploration Depth:** Evaluate whether the work deeply explores its core issue rather than skimming the surface.
 - 5: Profound insights and thought-provoking.
 - 4: Some depth but could still go further.
 - 3: Remains at a superficial discussion.
 - 2: Very superficial treatment.
 - 1: Did not touch the essence at all.
- **Humanities Creativity:** Different from the creativity of scientific and engineering works, the creativity of humanities works is about observing the world from a unique perspective, designing plots and themes innovatively, and using creative expression techniques to convey thoughts and emotions, thereby inspiring readers or audiences to think and feel, bringing new value to human culture and the spiritual world.
 - 5: Highly creative. The work exhibits unprecedented novelty in view, plot, and achieves a high level of innovation in artistic style and expression, having a profound impact on the cultural field;
 - 4: Quite creative. Exhibits a high degree of originality and uniqueness in content or form, effectively capturing readers' attention and provoking deeper thoughts;
 - 3: Basic creativity. Can propose interesting view or relatively novel forms from more common perspectives but has not reached the level of being refreshingly new;
 - 2: Some attempts but not very successful. Though there have been some innovative attempts, the overall impression remains quite ordinary, leaving no deep impression;
 - 1: Almost no creativity. Content is outdated, lacks novelty, and the expression technique is mediocre, unable to arouse readers' interest or thought.
- **Genre Recognition:** Identify the genre of the document, such as poetry, fiction, essay, advertisement, news report, public document, official document, literature, academic paper, report, etc.

Note: The scores for “Tone and Expression”, “Emotional Expression”, and “Genre Focus” must relate to the genre.
- **Tone and Expression:**
 - 5: Perfectly matches the required style for the genre. For advertisements, essays, novels, plays, and screenplays, speeches, etc.: Uses appropriate tone words, colloquial or humorous expressions. For news reports, official documents, literature, academic papers, reports, etc.: Uses precise wording, maintaining written expression as much as possible rational and objective.
 - 4: Mostly meets the requirements with only slight deviations from the expected style.
 - 3: Expression method basically meets genre needs but with some inconsistencies.
 - 2: Significantly deviates from the expected style but still partially acceptable.
 - 1: Almost entirely does not meet the expression style required by the genre.

(Continue in the next page)

Figure 9: Prompt template for quality rating that assesses HSS documents using expert-crafted rubrics (Part II).

Prompt Template for Quality Rating (Part III)

- **Emotional Expression:**

- 5: Emotionally rich and appropriate, perfectly matching the emotional depth required by the genre. For advertisements, essays, novels, plays, and screenplays, speeches, etc.: Emotion is abundant and appropriate. For news reports, official documents, literature, academic papers, reports, etc.: Uses precise wording to minimize personal emotional expression.
- 4: Emotionally appropriate with minor shortcomings or excesses.
- 3: Medium level, neither particularly outstanding nor notably flawed.
- 2: Weak emotional expression, or overly emotional for certain genres.
- 1: Extremely lacking in appropriate emotional color, or severely improper emotional expression.

- **Genre Focus:**

- 5: The document is highly consistent with the genre's focus in structure, language, and overall presentation, fully meeting the requirements of its genre. Announcements, notices, contracts need to emphasize rigor and professionalism, especially in standardized format, ensuring accuracy and ease of understanding. Reports, plans, and summaries require clarity, detailed content, and some format norms to reflect their seriousness. Regulations demand clarity and explicit statement for ease of compliance. Speeches pursue language appeal and expressiveness to resonate with the audience, and language expression needs to be colloquial with a certain diversity (writing techniques, rhetorical diversity, structure diversity, vocabulary diversity, etc.), but does not require the same level of diversity as pure literary creation. Literary creations such as novels, poetry, and essays focus more on emotional expression and aesthetic beauty, attracting readers through rich plot settings and graceful language style and requiring ample diversity (writing techniques, rhetorical diversity, structure diversity, etc.). Biographies and historical novels need to ensure both story appeal and factual accuracy with necessary diversity (writing techniques, rhetorical diversity, structure diversity, etc.). Sci-fi and fantasy rely on novel and unique world settings to spark readers' imagination, needing to abide by certain scientific laws and requiring necessary diversity (writing techniques, rhetorical diversity, structure diversity, etc.). Travelogues and essays record personal experiences or insights, reflecting personal color, and require necessary diversity (writing techniques, rhetorical diversity, structure diversity, etc.).
- 4: Generally good performance but slightly lacking in some details.
- 3: Basically meets requirements but with clear room for improvement.
- 2: Fails to meet expected standards in multiple key areas.
- 1: Severely violates the basic rules or conventions of the genre.

The domain of the document is {Domain}.

[Input Text]: {text}

Example Output:

```
{
  "Grammar, Punctuation, and Spelling": {"score": <1-5>, "reason": "<detailed explanation>"},
  "Logical Coherence and Fluency": {"score": <1-5>, "reason": "<detailed explanation>"},
  "Content Accuracy": {"score": <1-5>, "reason": "<detailed explanation>"},
  "Domain Relevance": {"score": <1-5>, "reason": "<detailed explanation>"},
  "Literary Diversity": {"score": <1-5>, "reason": "<detailed explanation>"},
  "Vocabulary Richness": {"score": <1-5>, "reason": "<detailed explanation>"},
  "Knowledge Depth and Breadth": {"score": <1-5>, "reason": "<detailed explanation>"},
  "Theme Exploration Depth": {"score": <1-5>, "reason": "<detailed explanation>"},
  "Humanities Creativity": {"score": <1-5>, "reason": "<detailed explanation>"},
  "Genre": "",
  "Tone and Expression": {"score": <1-5>, "reason": "<detailed explanation>"},
  "Emotional Expression": {"score": <1-5>, "reason": "<detailed explanation>"},
  "Genre Focus": {"score": <1-5>, "reason": "<detailed explanation>"
}
```

Figure 10: Prompt template for quality rating that assesses HSS documents using expert-crafted rubrics (Part III).

Prompt Template for Text Refinement

You are an expert in text refinement and content cleansing, specializing in preparing “answer” texts for high-quality datasets. Your core task is to: thoroughly refine and clean the raw text provided below, to maximally eliminate all “crawler traces” and any redundant information unrelated to the article’s core content, while **absolutely ensuring 100% completeness and accuracy of the original text’s core content, structure, meaning, author’s tone, and “human touch,”** enabling it to serve directly as a reference answer for Training.

[Specific Operational Requirements]

1. Thoroughly Clear Useless Information (Crawler Traces & Redundancy):

- **Must remove, including but not limited to:**
 - Webpage navigation bars (headers, footers, sidebars, menus, breadcrumb navigation)
 - Book tables of contents (unless the content itself is part of a TOC structure and must be retained)
 - Publisher information, copyright notices, disclaimers
 - Advertisements, promotional material, related article recommendations, social sharing buttons/prompts
 - Subscription prompts, user comment sections, donation requests
 - Excessive blank lines, residual broken HTML tags, URL lists
 - Repeated text blocks (e.g. duplicated paragraphs or sentences due to crawling errors)
 - Any peripheral, fragmented text clearly unrelated to the article’s core subject.
- **Judging Criteria:** If the absence of a certain text segment does not affect the article’s logical coherence, the completeness of its core information, or the reader’s understanding of the main topic, it can be considered useless information.

2. Core Content & Structure Retention (Of Paramount Importance):

- **Ensure that all key information, arguments, facts, data, explanations, examples, and logical deductions from the original text are completely intact.** No form of summarization, generalization, paraphrasing, or information loss is permitted.
- **Maintain the original narrative structure:** Retain original paragraph divisions, heading levels (e.g. H1, H2, etc.), and list formats (if any, e.g. ordered/unordered lists) as much as possible, unless these structures themselves constitute crawler traces.
- **Retain the original “human touch”:** Preserve the author’s original tone, style, and textual fluidity, avoiding any machine-generated or overly rigid sound.

3. Text Coherence and Fluency:

- The refined text must be a self-contained, complete, and logically coherent whole.
- After removing useless information, transitions in the text should be natural and smooth, with no abruptness or disconnections.
- Ensure the text remains grammatically correct and clearly expressed after redundancy removal.

4. No Repetition, High Precision:

- The final text should not contain any duplicated sentences or paragraphs.
- The quality of the generated text must meet the extremely high standards of an SFT answer, meaning accuracy, completeness, and usability approaching 100%.

5. Robustness:

- If the input raw text is already very clean and does not contain any of the elements listed above for removal, please return its content directly encapsulated in the JSON format, without any modifications.

[Input Text]: **{text}**

[Output Format]

```
{  
  "text": "refined text"  
}
```

Figure 11: Prompt template for text refinement to polish seed documents.

Prompt Template for Refinement Judge

You are a text-quality assessment expert specializing in evaluating answer texts. Your task is to carefully compare the original document and the refined document, evaluating whether the refinement meets the requirements for a high-quality answer.

[Original Document (Seed Document)]: **{text}**

[Refined Document]: **{refined text}**

[Evaluation Task]: Please evaluate the performance of the “Refined Document” relative to the “Original Document” based on the following criteria:

1. **Redundant Information Removal**: Has the refined text thoroughly removed all crawler traces and redundant information (e.g., web navigation, advertisements, copyright notices, excessive blank lines, duplicate content, fragmented text not central to the core topic, etc.)?
2. **Core Content and Structure Integrity**: Has the refined text 100% preserved all core content, key information, data, arguments, narrative structure, and the author’s tone and “human touch” from the original text, without any form of summarization, generalization, rephrasing, or information loss?
3. **Text Coherence and Quality**: Is the refined text natural and fluent, without abruptness or breaks, grammatically correct, and does it meet the extremely high standards of accuracy, completeness, and usability required for an SFT answer?
4. **Robustness (if original text is clean)**: If the “Original Document” itself was already very clean, did the refined text remain completely unchanged, without unnecessary modifications?

[Scoring (1-5)]

- 1: Very poor – completely fails to meet requirements, severe issues, unusable as an SFT answer.
- 2: Poor – multiple issues, unsuitable as an SFT answer.
- 3: Fair – meets some requirements, but still has significant room for improvement.
- 4: Good – largely meets requirements, with only minor, negligible issues.
- 5: Excellent – perfectly meets all requirements, can directly serve as a high-quality SFT answer.

[Output Format]

```
{  
  "Reason": "<detailed explanation>",&br/>  "Score": <1-5>  
}
```

Figure 12: Prompt template for a refinement judge that determines whether text refinement is successful.

Prompt Template for Multi-attribute Instruction Backtranslation

You are an expert in instructions inversion. Your task is to generate a “reverse synthesis instruction” for a given text.

GUIDELINES FOR INSTRUCTION GENERATION

Part 1: The Core Instruction

This section outlines “what” to the task – the specific, actionable requirements for generating the instruction.

1. **Objective:** The generated instruction should explicitly guide the writer to create a piece of writing that is highly similar to the given text.
2. **Domain:** The instruction must clearly specify the text’s **Domain** as **{domain}** and **Genre** as **{genre}**.
3. **Core Content & Key Points:**
 - Summarize the central topic and the main thesis/argument of the text.
 - List the essential supporting points, key data, primary examples, or narrative events that **must** be included to preserve the original’s substance.
 - Ensure all essential elements—such as key figures, significant events, central concepts, and defining terms—must be retained with exactness.
4. **Structure & Narrative Voice:**
 - Describe the text’s organizational structure (e.g., problem-solution, chronological, thematic).
 - Specify the **Narrative Voice** (e.g., first-person reflective, third-person objective narrator).
5. **Length Constraint:** Strictly enforce the word count. State clearly: “The final text must be approximately **{token_len}** tokens (around **{char_len}** characters), matching the length of the original document.”
6. **Concise Without Repetition:** The instruction sentence itself must be concise and avoid unnecessary repetition.
7. **No Source Reference:** The instruction must not mention, imply, or hint to the “source document”, “original text”, etc.

Part 2: The Persona

This section defines the “how”—the persona the writer should adopt. It’s a role-playing guide to capture the original author’s spirit. Frame this entire section in the second person (“You are...”).

1. **Stance:** Based on the text’s content, describe the author’s likely underlying stance or values when discussing the topic.
2. **Mindset & Tone:** Based on the text’s overall emotional color and undertones, describe the author’s probable underlying mindset and emotional tone.

INSTRUCTION GENERATION TASK

Please generate the “reverse synthesis instruction” for the following document: **{text}**

FINAL OUTPUT FORMAT

Provide the output in a single JSON object with two keys: “Instruction” and “Persona”.

```
{  
  "Instruction": "(The core task instruction generated according to Part 1 guidelines)",  
  "Persona": "(The persona generated according to Part 2 guidelines, written in the second person.)"  
}
```

Figure 13: Prompt template for multi-attribute instruction backtranslation, which backtranslates instructions from seed documents while attaching multiple attributes.

Prompt Template for Question-Answer Alignment Check

You are an expert in the quality evaluation of QA pairs. Your task is to evaluate the quality of a reverse instruction. Evaluate whether the reverse instruction can guide a model to generate reverse text similar to the original text. According to the Instruction Effectiveness in the Evaluation Criteria, and using Textual Consistency as a supplementary reference, determine whether the reverse instruction is “qualified” or “unqualified”. Your final output is a binary conclusion: **1 means qualified, 0 means unqualified**.

[Input]

REVERSE INSTRUCTION: The instruction to be evaluated.

ORIGINAL TEXT: The human-written reference (“correct answer”).

REVERSE TEXT: The actual output generated by the model based on the reverse instruction.

<REVERSE INSTRUCTION> {reverse_instructions} </REVERSE INSTRUCTION>

<ORIGINAL TEXT> {text} </ORIGINAL TEXT>

<REVERSE TEXT> {generated_text} </REVERSE TEXT>

[Evaluation Criteria]

Please judge whether the reverse instruction is qualified strictly according to the following standards.

- **1 (Qualified)**

Textual Consistency: The core content of the reverse text is strictly consistent with the original text.

Instruction Effectiveness: The reverse instruction should include the core content from the original text. The reverse instruction must NOT contain words or phrases such as “original” or “the above content” that directly refer to the original text.

- **0 (Unqualified)**

Textual Inconsistency: The core content of the reverse text is inconsistent with the original text. The key characters, key events or key messages are inconsistent.

Instruction Ineffectiveness: The reverse instruction lacks the core content and key characters from the original text, or contains incorrect or fabricated information. The reverse instruction contains words or phrases such as “original” or “the above content” that directly refer to the original text.

[Output Format]

Return a JSON object with two keys:

```
{
  "reason": "A explanation of why you chose the score",
  "score": <1 | 0>
}
```

Figure 14: Prompt template for question-answer alignment check used to verify that the reverse instructions faithfully reproduce the seed document.

Prompt Template for Teacher Forced Answering

Your task is to generate a **brand new document**. This new document must simultaneously achieve two core objectives:

1. Content – 100% QA Consistency

Strictly, completely, and precisely respond to all requirements of the **[Reverse Instruction]**, which is a precise instruction that defines the core requirements, theme, and scope of the content to be generated. The final output must strictly adhere to this instruction, ensuring **100% Question Answer (QA) consistency**. This is the highest priority requirement.

2. Style & Depth

Learn from and reproduce, to the greatest extent possible, the strengths of the **[Human-Written Reference Document]**, including but not limited to:

- Tone & Style
- Vocabulary & Sentence Structure
- Information Architecture
- Richness of Detail & Examples

Constraints and Rules

1. **Instruction Priority** – If there is any conflict in content between the **[Reverse Instruction]** and the **[Human-Written Reference Document]**, the **[Reverse Instruction]** must take precedence.
2. **Information Integration** – You may borrow details, data, or examples from the **[Human-Written Reference Document]** that support the **[Reverse Instruction]**, but you must not introduce new information that falls outside the scope of or is irrelevant to the **[Reverse Instruction]**.
3. **Beyond Model Limitations** – Your goal is for the generated document's quality to surpass the machine-templated content typically produced by simply using the **[Reverse Instruction]**, reaching or even exceeding the sophistication of the **[Human-Written Reference Document]**, while maintaining perfect QA consistency.

Here are two input materials:

- **[Reverse Instruction]: {Reverse Instruction}**
- **[Human-Written Reference Document]: {Seed document}**

Generation Requirement

Please generate the text according to the above requirements.

Figure 15: Prompt template for teacher-forced answering, which produces answers anchored to seed documents that exceed the synthesis model's response limits.