DomAINS - DOMain Adapted INStructions

Anonymous ACL submission

Abstract

Domain-specific large language models (LLMs) demonstrate strong domain expertise by training on large-scale, domain-aligned instruction data. However, manually constructing such datasets is resource-intensive due to the need for expert annotators. A promising alternative is to use LLMs to synthesize training data. While existing frameworks effectively generate general instruction datasets, generating domain-specific instruction datasets presents the following main challenges: the data must (1) be strongly aligned with the target domain, (2) exhibit high in-domain diversity, and (3) be factually grounded on domain-specific knowledge. In this paper, we present **DomAINS**, a three-stage framework to generate instruction datasets for any target domain using only a domain name and a brief description. DomAINS constructs a tree of domain-relevant keywords to increase in-domain diversity, retrieves factually grounded domain articles from Bing, and prompts an LLM to generate domain-aligned instruction data based on the retrieved articles. Our evaluation across nine domains shows that models tuned on DomAINS-generated dataset achieve 60–95% win rate over those trained on datasets from existing synthetic frameworks for general domains, demonstrating the effectiveness of our approach.

Prompt: Generate instruction–response pairs for domain = 'zoology'.

(1) Domain Misalignment

Instruction: How does python behave when it encounters a bug in the wild?
Response: When Python encounters a bug at *runtime*, it raises error ...*e.g., IndexError* ... *try/except blocks* ...."

(2) Factual Incorrectness

Instruction: Describe the mating behavior of clownfish.
Response: Clownfish build ... *nests of seaweed* and lay *thousands* of eggs every hour, ...

(3) Redundant Samples

Instruction: How is the mating behavior of clownfish?
Response: Clownfish builds ... lays *thousands* of eggs

Figure 1: Challenges observed during domain-specific instruction dataset generation: (1) Domain Misalignment: "python" is treated as coding language rather than a snake; (2) Factual Incorrectness: seaweed nests (erroneous) and thousands (exaggeration) due to lack of domain knowledge; (3) Redundant Samples: trivial paraphrases yield identical outputs, adding no new information

1 Introduction

General-purpose LLMs (OpenAI, 2022, 2023a; Grattafiori et al., 2024; Anthropic, 2024) have demonstrated proficiency across diverse tasks but often wane in specialized domains (Ling et al., 2023). Such shortcomings stem from insufficient domain-specific knowledge, leading to overly generalized or inaccurate responses. For instance, ambiguity in domain-specific jargon can result in significant errors: when queried *"What does PEP stand for?"*, a financial expert may expect *"Politically Exposed Person"*, a Coding expert would anticipate *"Python Enhanced Proposal"* and a virologist would consider *"Post-Exposure Prophylaxis."* Incorrect assumption of PEP could derail a model's entire reasoning chain, exacerbating the risk of deploying such models.

To equip language models with domain expertise, researchers either pre-train (Wu et al., 2023; Wang et al., 2024) or instruction fine-tune (IFT) (Wang et al., 2023; Yue et al., 2023; Zhang et al., 2023; Cui et al., 2023) domain-specific models. Despite its effectiveness, training domain-specific models hinges on the availability of high-quality, domain-aligned datasets, which are often proprietary or scarce. A trivial approach is to manually curate instruction datasets (Wang et al., 2022c; Bach et al., 2022; Conover et al., 2023; Vila-Suero and Aranda, 2023) by recruiting domain experts, but it's inherently time-intensive and costly.

Synthetic instruction dataset generation frameworks (Wang et al., 2022b; Xu et al., 2024b,a; Köksal et al., 2023; Yehudai et al., 2024; Ge et al., 2024; Gupta et al., 2023) offer a promising alternative to abate data curation costs, but they inherently do not focus on domain-aligned instruction dataset generation. We observe three key challenges to achieve our goal, as presented in Table 1. First, **weak domain alignment**. SOTA frameworks struggle to curtail generation to the targeted domain, necessitating post-hoc filtering to retain domain-specific samples, often leaving low (or even zero) relevant data samples. A recourse is modifying the prompts to steer generation over the domain keyword, which helps with domain-alignment but leads to the second issue, **high sample redundancy & low in-domain diversity** (as witnessed in Table 1) – yielding repetitive, ambiguous, or overly generic instruction samples. Third, **low factual grounding**. Since the underlying generic models inherently lack domain knowledge, the instruction–response pairs generated exhibit weak semantic correlation and factual inconsistencies, culminating in noisy datasets. While employing post-filtration steps (Wang et al., 2022b; Xu et al., 2024b; Gupta et al., 2023) can mitigate the aforementioned quality concerns, they often compromise scalability and diversity.

To bridge these gaps, we introduce **DomAINS** (**DOM**ain **A**dapted **INS**tructions), a 3-stage framework that, given a domain keyword (e.g., *history*) and a brief description (1–2 lines), automatically generates a domain-aligned instruction dataset for fine-tuning **domain-expert models**. Our key contributions: **Multi-level Tree Expansion** strategy, where we initialize a tree with the user-fed domain keyword and expand it iteratively by adding suitable domain-relevant subtopic words, aiming to capture a domain expert's breadth of knowledge, bolstering in-domain diversity. **Domain-relevant Grounded Generation** by sourcing real-world, domain-relevant, text-rich articles via Bing to anchor instruction-response pairs in factual content, which also aids in maintaining strong instruction-response correlations and reducing hallucinations, thus improving overall dataset quality. Together, these strategies administer **Domain Aligned Generation** by generating samples highly relevant to the target domain, as seen in Figure 5.

We generate 9 domain-specific datasets (music, astronomy, history, Mesopotamia, agriculture, virology, Leukemia, art, fish), employing DomAINS, each comprising 100K samples – using only 1 A6000 GPU (Llama-3.1-8B-Instruct). Our qualitative results indicate that DomAINS effectively produces **strongly domain aligned, intrinsically diverse, high quality factually grounded** instruction datasets. Moreover, we witness significant performance improvements (**60-95% Win-Rate (RC)**) in Llama-3.1-8B-Instruct when tuned on DomAINS compared to our baseline SOTA-tuned variants((Xu et al., 2024b; Köksal et al., 2023; Xu et al., 2024a)) across all 9 domains.

2 DomAINS - Proposed Framework

Overview. In this section, we detail the components of our proposed framework - **DomAINS**, an end-to-end framework that requires only a domain keyword and a brief contextual description (domain context) to generate a domain-aligned instruction dataset. DomAINS operates in 3 main stages: (1) Subdomain topic words sampling (2) Article retrieval (3) Instruction dataset generation, as illustrated in Figure 2. In Stage 1, the framework garners an eclectic set of domain-relevant subtopic words to ensure in-domain diversity. In Stage 2, it retrieves domain-relevant text-rich Bing articles paramount for content grounding. In Stage 3, query LLM to generate instruction-response pairs grounded to the articles retrieved in Stage 2.

Input: User provides **domain** keyword (e.g., *kayak*) along with a brief description (**domain context**). Here, *"domain"* refers to an industry/business domain (e.g., fish, agriculture) or area of knowledge or expertise within academia (e.g., history, Mesopotamia, astronomy). *Domain* keyword and *domain context* help to disambiguate polysemous terms(e.g., *virus*, *chord*), thereby reinforcing strong domain alignment for dataset generation.

2.1 Subdomain Topic Words Sampling

To capture the intrinsic diversity within a given domain, DomAINS generates a set of **subdomain topic words** that emulates the breadth of knowledge a well-rounded expert possesses. For example, a historian would naturally be familiar with subtopics such as "Renaissance," "ancient India," "Egyptian civilization," and "Viking era."

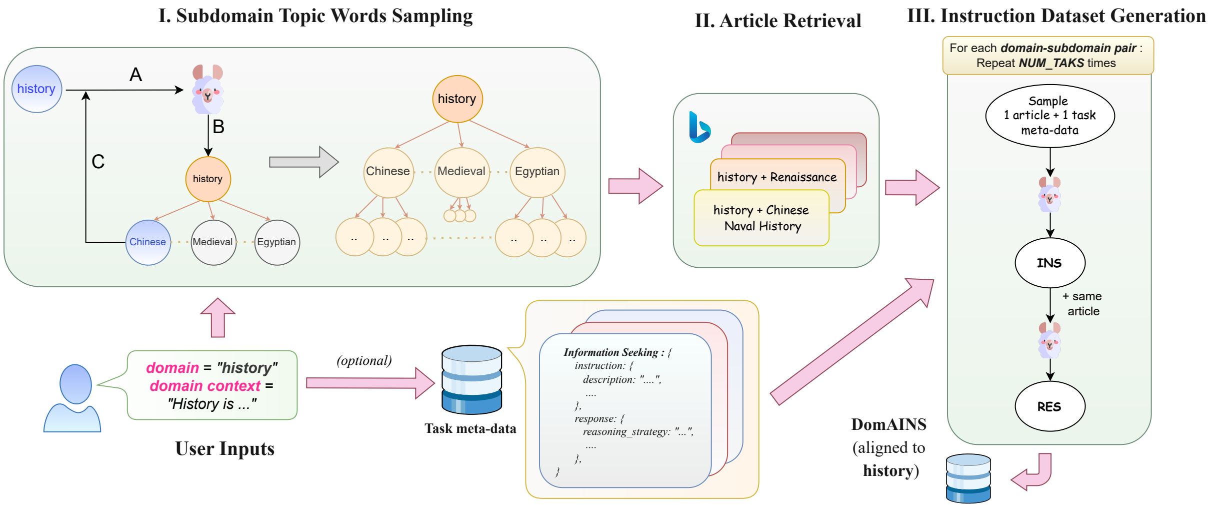


Figure 2: Architectural overview of DomAINS. User first inputs the domain keyword with a brief description. Providing task metadata is optional. User can additionally define tasks specific to the domain. In Stage I, DomAINS generates a list of subdomain topic words Phase A: Prompt LLM to "Generate 50 synonyms for *{domain}*". Phase B: Retain candidate words relevant to *domain* & *domain context* and append to the tree. Phase C: Expand the tree in BFS fashion: "Generate 50 synonyms for *{subdomain}*" until the stopping condition is reached. Then in Stage II, it retrieves articles for each domain–subdomain pair. Stage III: Instruction Dataset Generation. Randomly sample an article and a task to produce grounded instruction-response pairs via an LLM.

Initial experiments with Word2Vec produced subdomain words that lost relevance beyond the top 200-400 words, returning irrelevant words or gibberish symbols (e.g., ####, @ad). In some cases, such as "Mesopotamia," no subdomains were retrieved. We subsequently present **Multi-level Tree Expansion strategy**. First, initialize the *root* of the *subdomain tree* with the domain keyword. Next, LLM[1] is queried to generate a small set of synonyms (e.g., 50) for the root. These candidate words are filtered by computing their cosine similarity against the domain context; retaining candidates above the predefined threshold. The filtered words form the next level of the tree, and the expansion continues in a BFS manner until a sufficiently diverse and comprehensive set of subdomain words is obtained. By our Multi-level Tree Expansion strategy, we gather eclectic domain-relevant set of subtopic words essential for the subsequent article retrieval step. Ablation studies on subdomain sampling and cosine threshold selection are detailed in the Appendix A.

2.2 Article Retrieval

In this step, DomAINS retrieve text-rich English articles for each domain-subdomain pair via Bing Search API[2]. This stage solves 3 major issues: (1) ever-evolving internet serves as a scalable data source; (2) abates parametric knowledge distillation as witnessed in (Xu et al., 2024b; Wang et al., 2022b) (3) Text-rich articles offer a reference for instruction-response pair generation that maintain strong correlations and reduce hallucinations, thereby improving overall dataset quality. Notably, fine-tuning on synthetically modified versions of publicly available datasets (Köksal et al., 2023; Yin et al., 2023a) yields minimal performance gains. On the contrary, unseen raw data sources (e.g. internet) or organizational external KB help produce unprecedented datasets.

2.3 Instruction Dataset Generation

In the final stage, DomAINS generates instruction datasets grounded in retrieved Bing articles and predefined task metadata. Owing to the poor quality of task descriptions (often ambiguous and repetitive) in the existing works, we manually curated 31 distinct task metadata templates[3], covering diverse task types like Advice Seeking, Reading Comprehension, Event ordering, Contrastive Analysis, etc. Users can easily adapt these templates to suit specific requirements, promoting flexibility in task design. Task templates, instruction and response generation prompts can be found in Appendix C.

[1]Llama-3.1-8B-Instruct

[2]Bing Search API (refer Appendix E for search queries)

[3]These tasks were written without any reference to the evaluation dataset used in Section 4

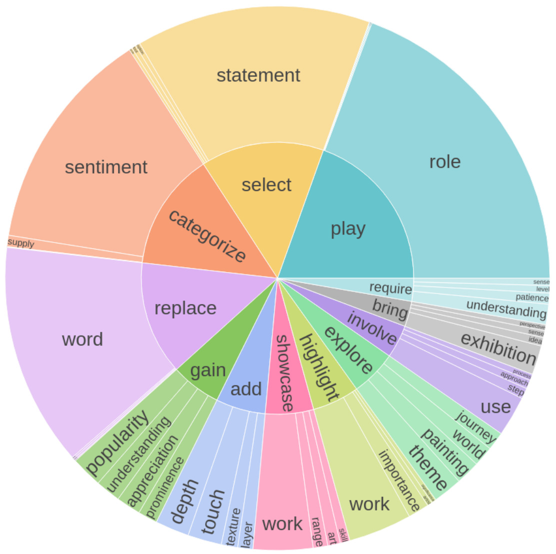


Figure 3: Pie chart demonstrates the top 12 commonly appearing root verbs and their top 4 direct noun objects for domain=**"art"**. Figure clearly arrays relevant verb-noun pairs - *"showcase-work"*, *"explore-theme"*, *"bring-perspective"*, *"add-texture"* pertinent to domain **art**.

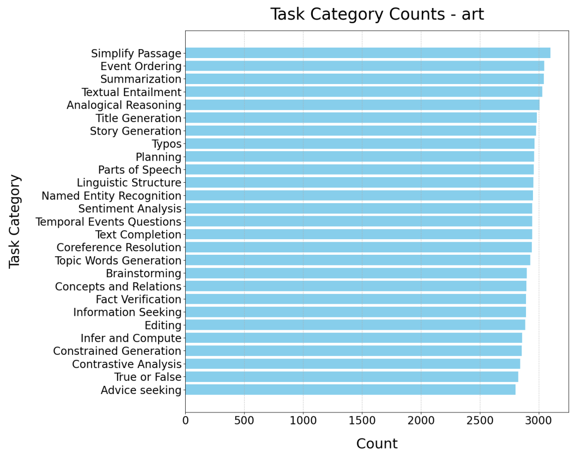


Figure 4: Uniform task coverage for domain=**"art"**

Each generation iteration has 3 steps: **(1) Retrieving Generation Metadata** - Randomly sample a Bing article from the article pool of that domain-subdomain pair. Concurrently, randomly sample a task from the task pool. Random mapping of articles with task-types minimizes duplicates and leads to uniform task coverage as evidenced in Figure 4. **(2) Instruction Generation** - Utilizing the retrieved metadata (Bing article and task configuration), DomAINS populates the *instruction generation template*, and queries the LLM, to generate an instruction. **(3) Response Generation** - Update the *response generation template* with the generated instruction and the same Bing article and query the LLM to generate a respone. Reusing the same Bing article for both instruction and response generation ensures semantic alignment between the two, thereby enhancing coherence and completeness while mitigating hallucinations.

3 Dataset Analysis

In this section, we validate the claims for the DomAINS datasets — namely, **strong domain alignment**, **intrinsic diversity**, and **high-quality factually grounded instruction–response pairs**. We conduct comprehensive analysis comparing our datasets with SOTA datasets, accentuating that DomAINS better meets the demands of domain-specific applications. Refer Appendix F for Instance examples from DomAINS datasets.

3.1 Diversity

Lexical diversity in a dataset is critical for fostering semantic understanding, particularly for domain-adapted models that emulates the breadth of knowledge possessed by domain experts. For example, a *"historian"* would recognize a wide array of relevant topics — such as the *"Renaissance," "medieval India," "Egyptian civilization,"* and *"Viking era"*, covering multiple facets of a domain. Prior approaches (filtering domain-specific samples from generic datasets or modifying prompts to focus on a particular domain), result in repetitive samples (as indicated by low domain unigram and bigram ratios in Appendix B). Our Multi-Level Tree Expansion strategy guarantees broad topic coverage. To further assess diversity, we extracted topic words[4] from both Magpie-Pro and DomAINS datasets and plotted t-SNE plots[5] (Van der Maaten and Hinton, 2008) (refer Figure 5). It is evident that DomAINS offers an extensive coverage of subtopics compared to Magpie-Pro, which clearly depicts overalapping subtopics, particularly in niche domains (Leukemia, Virology). Additionally, where SOTA suffers from skewed task distributions, DomAINS offers balanced task coverage as evidenced in Figure 4. So, DomAINS offers both diverse subtopic and balanced task coverage.

3.2 Domain Alignment

Our primary objective is to ensure strong alignment with the user-defined domain besides maintaining

[4]We first computed n-grams(ranging from 2 to 10) from all the instruction samples. Then computed the cosine similarity w.r.t the domain context and selected the top 3000 sorted n-grams

[5]Computed embeddings via all-mpnet-base-v2

intrinsic diversity, i.e., the dataset should be lexically rich but with domain relevant topics. Figure 5 connotes that the clusters from the DomAINS dataset are centered around the domain keyword, underscoring robust domain alignment. Furthermore, Figure 3 illustrates that each domain dataset befittingly associates domain-relevant nouns with verbs, thereby emphasizing domain specificity. Appendix B and I further details on dataset attributes, statistics, root-verb and task coverage plots.

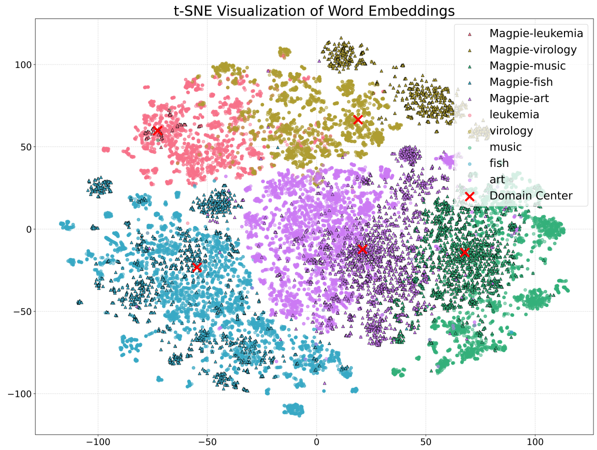


Figure 5: The figure demonstrates intrinsic diversity and strong domain alignment of our generated dataset against Magpie over 5 domains. We extracted subtopics from both datasets and plotted the T-SNE plot. Red crosses represent respective domain keywords. Extended plot in Appendix B.

3.3 Quality

Existing frameworks (Xu et al., 2024a; Köksal et al., 2023) rely on expensive proprietary LLMs (e.g., (OpenAI, 2022, 2023a)) or (Wang et al., 2022c; Xu et al., 2024b; Wang et al., 2022b), require extensive post-filtration to maintain quality. In contrast, DomAINS yields high-quality instruction–response pairs without an explicit post-filtration step. We validate our claim through multi-metric evaluations on both instructions and responses. We use Llama-3.1-8B-Instruct, following similar evaluation protocols as in (Xu et al., 2024b; Liang et al., 2022). Qualitative evaluation prompts are presented in Appendix D.

3.3.1 Instruction Quality

We assess instruction quality using metrics - coherence(Eldan and Li, 2023; Zhang et al., 2024b), ambiguity(Niwa and Iso, 2024), complexity (Li et al., 2023), and completeness(Côrtes, 2024), and report the scores in Table 1. Among SOTA

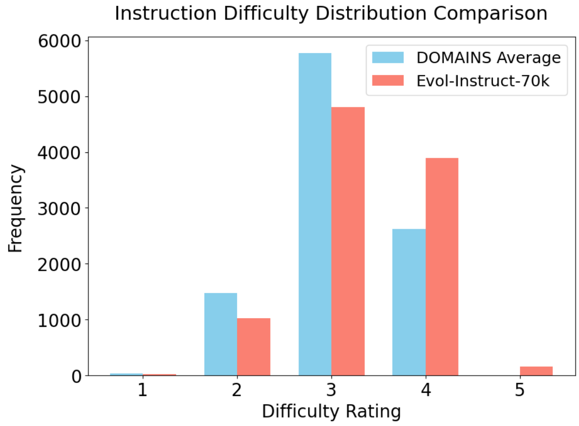


Figure 6: Instruction Difficulty distribution comparision of Evol-Instruct-70k vs DomAINS (averaged across all domains). We can clearly see DomAINS difficulty distribution is more skewed towards the *easy and medium ratings* in contrast to Evol-Instrct-70k which is skewed on the *difficult and very difficult ratings*.

datasets, SuperNI exhibit high coherence and clarity. Evol-Instruct generates instructions with better complexity and completeness. DomAINS produces instructions that are competitive (at-par or even better) across all metrics, even though using a comparably smaller open-source model. Grounded generation to text-rich Bing articles contributes to the factual accuracy, coherence, clarity, and completeness of the instructions. Notably, majority of instructions generated by DomAINS are of *medium* difficulty (Figure 6) so integration with Evol-Instruct can help enhance the complexity of the instruction dataset.

3.3.2 Response Quality

Response quality is assessed on metrics - relevance (Zhou et al., 2024; Wu et al., 2024), coherence, adherence (Qin et al., 2024; Xia et al., 2024), completeness, and effectiveness of CoT reasoning (Amirizaniani et al., 2024). As shown in Table 1, Magpie-Pro has superior results for Completeness, Relevance, and Adherence - primarily due to its extensive post-filtration, retaining only ~10-30%[6] of the entire generated dataset. Such an approach is not an optimal utilization of API calls. SuperNI (manually written) attained the best coherence. Overall, DomAINS performs at-par with the best SOTA in all quality metrics, with far superior

[6] Magpie-Pro uses Llama-3-70B-Instruct for instruction-response generation and then selects top 300k high quality instances from 1M generated samples, whereas Magpie-Air uses Llama-3-8B-Instruct and selects top 300k instances from 3M generated samples

Table 1: Qualitative Analysis. All the scores are computed by Llama-3.1-8B-Instruct and averaged over 10k randomly sampled records. We consider Manually annotated dataset (SuperNI), Synthetic dataset frameworks using (a) Proprietary, expensive LLMs as generators (Evol-Instruct (ChatGPT), Longform (GPT4)) (b) Extensive Post-filtration (Magpie) (c) Both (Self-Instruct (GPT3 and post-filtration). Then we have DomAINS dataset for all 9 domains.

Dataset	**Instruction**				**Response**				
	Co	**Cl**	**Cx**	**Comp**	**Co**	**Comp**	**R**	**A**	**CoT**
SuperNI	4.960	4.739	3.122	4.927	**4.303**	4.158	4.748	4.011	3.207
Self-instruct	4.859	4.732	2.640	4.740	3.793	3.671	4.373	3.307	2.573
Magpie-Pro	4.912	4.729	2.824	4.651	4.076	**4.619**	**4.984**	**4.627**	4.193
Evol-Instruct-70k	4.960	4.537	**3.785**	**4.931**	4.122	4.319	4.925	4.394	3.697
Longform-C	4.842	4.584	2.705	4.774	3.942	3.902	4.860	3.933	3.308
Mesopotamia	4.964	4.828	3.141	4.926	4.110	4.170	4.915	4.508	4.643
Fish	4.959	**4.858**	3.050	4.929	4.202	4.354	4.864	4.357	4.494
History	**4.968**	4.847	3.112	**4.931**	4.123	4.159	4.927	4.507	**4.672**
Art	**4.968**	4.570	3.137	4.925	4.087	4.175	4.916	4.568	4.654
Music	4.955	4.804	3.056	4.928	4.075	4.138	4.905	4.485	4.623
Astronomy	4.954	4.832	3.075	4.903	4.096	4.187	4.915	4.554	4.660
Virology	4.951	4.853	3.190	4.905	4.113	4.194	4.906	4.516	4.635
Agriculture	4.964	4.711	3.161	4.907	4.102	4.186	4.912	4.524	4.664
Leukemia	4.950	**4.858**	3.143	4.908	4.111	4.137	4.905	4.508	4.646

Underline signifies best score among the SOTA models. ***Bold*** *signifies best score overall.* NOTE: We use Co=Coherence; Cl=Clarity; Cx=Complexity; Comp=Completeness; R=Relevance; A=Adherence; CoT=Chain-of-Thought Reasoning Strategy

reasoning capabilities in its responses, aligning with our goal of creating domain-expert models with apt reasoning abilities rather than merely rote memorization (Mitchell and Martin, 1997; Zečević et al., 2023). Yet improving DomAINS for better adherence (intent understanding) remains a promising direction for future works.

4 Performance Analysis

4.1 Evaluation Dataset Setup

Due to the unavailability of domain-specific datasets, we devised a 2-stage retrieval mechanism. First, we compiled a diverse collection of public benchmark datasets (detailed in Appendix H). Next, we retained samples with the "domain keyword" (e.g. *history*), to reduce the sample pool, followed by a context-based filtration, where we used Llama-3.1-8B-Instruct to get the instance embeddings, computed its cosine similarity w.r.t the "domain context" and eliminated entries below the threshold[7]. This step filtered out polysemous samples from the evaluation dataset.

[7]We tested for a range of cosine thresholds [0.4 , 0.9] and empirically found **0.7** produced relevant batches of domain-relevant instances

4.2 Experimental Setup

Baselines: (1) Pure distilled datasets: (Xu et al., 2024b,a) (2) RAG-based distilled dataset: (Köksal et al., 2023).

Implementation details: We generated 9 instruction datasets for domains viz, art, agriculture, fish, music, astrology, virology, Leukemia, history, Mesopotamia, using **Llama-3.1-8B-Instruct** model on a **single A6000 GPU** (~100 hrs). We use **2000** subdomain topic words and retrieve **25** bing articles for each domain-subdomain pair. Finally, we generate **50** tasks per domain-subdomain pair yielding **100,000** samples per domain. We tuned **Llama-3.1-8B-Instruct** using LoRA(Hu et al., 2022) on **2 NVIDIA A100 60GB** GPUs (~33 hrs), releasing **9 DomAINS-tuned models**. We use the same model and hyperparameter configurations for each baseline datasets as DomAINS models. Additional training hyperparameters and configurations are detailed in Appendix G.

Metrics: We report **Win-Rate(WR)** scores of DomAINS-tuned models over baselines, computed using GPT-4o[8].

[8]Open AI GPT-4o

Table 2: Performance comparison of models on our domain-specific evaluation dataset. We report **(WR%)** of DomAINS-tuned models against the baselines across 9 domains. First section consists of vanilla Llama-3.1-8B-Instruct. Second section comprises Llama-3.1-8B-Instruct tuned on SOTA datasets.

Baseline \ DomAINS	**Leukemia**	**Virology**	**Music**	**Astro***	**Meso***	**Hist***	**Fish**	**Agri***	**Art**
Llama-3.1-8B-Instruct	71.13	79.82	79.80	71.18	81.74	52.00	65.05	75.10	68.55
Llama-magpie-pro	72.63	77.09	74.13	65.01	67.27	71.34	50.00	68.18	59.43
Llama-evol-instruct-70k	79.38	80.77	51.22	68.8	74.92	75.56	68.36	76.90	71.41
Llama-longform-C	95.83	95.07	94.26	83.13	95.52	90.51	93.37	92.67	92.88

*Note: We use DomAINS' respective domain-specific models for each domain. All WR scores are reported in %. *We use short-forms for Astronomy, Mesopotamia, History, and Agriculture, respectively.*

5 Discussion

Our experiments reveal several key observations regarding the efficacy of DomAINS. Notably, DomAINS-tuned models achieved **~60–95% WR** improvements across 9 diverse domains (see Table 2), demonstrating its capability to generate datasets that are **strongly domain aligned**, **intrinsically diverse** and **factually grounded** with responses exhibiting **good reasoning capabilities** — critical properties for tuning **domain-expert models**. Our current implementation generates only 100K samples per domain due to technical constraints, we anticipate that increasing the number of subdomain words (enhances intrinsic diversity), task variations (promoting task generalization), and retrieved articles (improving content grounding) would further enhance dataset size and robustness.

Evaluation in this setting remains challenging due to the paucity of standardized domain-specific gold standards. Our curated evaluation dataset serves as a close approximation; however, in-apt samples may still persist, which could obscure the full extent of domain-specific improvements. This underscores the need for frameworks capable of rapidly synthesizing evaluation sets tailored to niche domains, a requirement that DomAINS is well-positioned to address. Moreover, the overfitting of existing LLMs on standard evaluation datasets further emphasizes the importance of developing methods that synthesize datasets with unseen content and diverse tasks to mitigate data contamination, thereby enabling more reliable evaluations for domain-specific benchmarking.

DomAINS offers additional benefits. First, it shows promise for **multi-lingual dataset generation** and **regional adaptation** by employing multilingual LLMs (DeepSeek-AI, 2024; Penedo et al., 2023) and updating Bing API search filters to target region-specific content and languages. Second, while our current implementation relies on Bing articles for grounding, future iterations will incorporate **external knowledge bases (KBs)** to produce even more relevant datasets for specialized domains and businesses. Third, extending the framework to support **multi-turn dialogue generation** is a promising avenue for developing robust **domain-expert chatbots**. Finally, DomAINS-generated data can serve as **high-quality seed data** for Evol-Instruct, to further generate more challenging domain-specific instructions, thereby facilitating a **weak-to-strong generalization** trajectory. Dynamic **domain-specific task adaptation** also remains an open research topic.

We also observe that current frameworks cannot readily generate diverse task formulations tailored to different domains. For instance, medical domain tasks might entail generating diagnostic queries or summarizing patient histories; legal domain tasks could require drafting contract clauses or analyzing case law; and coding tasks may focus on snippet generation and debugging algorithms. A flexible generation pipeline is therefore crucial, allowing **dynamic task adaptation** catering to the varying complexities and nuances of specialized fields.

6 Related Works

Human-Curated Datasets Expert-authored corpora (Bach et al., 2022; Longpre et al., 2023; Wang et al., 2022c; Mishra et al., 2022) offer high semantic fidelity and task relevance, but incur cumbrous labor and time costs, rendering them infeasible, especially for niche domains with limited expert availability. This lack of domain knowledge can lead to weak instruction-following ability (Yin et al., 2023b; Kung and Peng, 2023) and controllability (Zhang et al., 2024a) on specific domains, leading to suboptimal performance.

Pure-Distilled Synthetic Datasets Subsequent works then leveraged LLMs to alleviate reliance on human annotation and directly distilled instruction data from LLMs. (Wang et al., 2022b) and (Honovich et al., 2022) bootstraped a small set of seed prompts and iteratively prompted LLM to generate new instruction-response pairs. (Xu et al., 2024b) introduced a pre-query templated strategy that exploited the autoregressive generation of LLMs to produce instruction dataset. (Xu et al., 2024a) and (Sun et al., 2023) focused on "evolving" the instructions' complexity. (Tran et al., 2024) prompted GPT4 with 3 seed data to generate Biomedical dataset. However, they often fail to enforce strict domain alignment or factual grounding, producing generic, repetitive, or noisy samples that demand heavy post-filtering. Although prior works have proposed improved generation strategies (Ma et al., 2024; Cai et al., 2024) and dataset selection (Kung et al., 2021, 2023) for specific tasks, maintaining domain alignment can still be challenging.

RAG-Based Distilled Datasets RAG was introduced to address the distillation shortcomings. (Ge et al., 2024) and (Yin et al., 2023a) uses structured datasets from HuggingFace; (Köksal et al., 2023) and (Yehudai et al., 2024) draw on large text corpora (e.g., C4, Wikipedia) to supply contextual passages. These methods improve factual consistency compared to pure-distillation, yet fine-tuning on synthetically modified versions of publicly available datasets yields minimal performance gains because the LLMs are already saturated on those datasets.

Mixed and Community-Sourced Datasets Hybrid approaches combine multiple data sources and human involvement to balance scale, diversity, and realism within the dataset. (OpenAI, 2023b) collects real user–assistant conversations; (Feuer and Hegde, 2025) merges public chat logs with distilled synthetic samples; (Gandhi et al., 2024) applies MoE filtering to samples from heterogeneous pools. (Vila-Suero and Aranda, 2023) curates domain-aligned datasets using volunteers and synthetic generation frameworks.

7 Conclusion

In this work, we introduce **DomAINS**, a 3-stage framework that generates strongly domain aligned, intrinsically diverse, and high quality factually grounded instruction datasets with minimal user inputs. Our evaluations across 9 domains demonstrate that DomAINS produces high-quality instructions(coherent, clear, complete) and responses(coherent, relevant to the instruction, adherent to constraints, sound reasoning strategy), even though we employed significantly smaller LLM (Llama-3.1-8B-Instruct) and no extensive post-filtering in our framework. Additionally, DomAINS-tuned Llama-3.1-8B-Instruct achieves **~60–95% WR** improvements over SOTA-tuned Llama-3.1-8B-Instruct, highlighting the efficacy of DomAINS produced datasets. Importantly, seeding Evol-Instruct with DomAINS datasets promises a **weak-to-strong generalization** trajectory, generating complex instruction datasets for highly specialized fields. By lowering the barrier to tune custom domain-expert LLMs, DomAINS paves the way for broader adoption of reliable, context-aware models in both research and industry. Looking forward, we envision extending DomAINS to support dynamic Task Adaptation, multilingual and regional acclimatization, and multi-turn dialogue generation, to further broaden its applicability across diverse domains.

Limitations

While DomAINS is scalable and efficient, it has some limitations. First, our current implementation relies on manually curated task metadata. Integrating AI agents (Hu et al., 2024; Cao and Lee, 2023) capable of automatically determining relevant tasks for a given domain and generating the corresponding metadata would further enhance the framework's Domain Adaptation capabilities. Second, our pipeline directly utilizes Bing articles without a dedicated pre-filtration step. Although grounding instruction–response pairs in factually correct articles reduces hallucinations, its still susceptible to biased or opinionated content, which may lead datasets to inherit such biases. Incorporating either pre-filtration during the content preparation phase or post-filtration of the curated dataset could mitigate these issues as an extra-precautionary step.

Ethical Statement

We conduct all experiments on 1 48GB NVIDIA A6000 GPUs or 2 NVIDIA A100 GPUs with 48 TB disk storage and AMD EPYC 7413 24-Core Processor. Our dataset generation takes 100 GPU hours (A6000) and instruction fine-tuning takes 30 GPU hours (A100). We use OpenAI GPT-4o model for evaluation. We use open-sourced model (Llama-3.1-8B-Instruct) and publically available evaluation datasets from Huggingface for our experiments and will release our code once the paper is accepted. In our work, we generate datasets for some high-stakes niche domains like Leukemia and Virology. Although we intend to release it publicly, we do not guarantee its realiblity for real-world applications.

References

Maryam Amirizaniani, Elias Martin, Maryna Sivachenko, Afra Mashhadi, and Chirag Shah. 2024. Can llms reason like humans? assessing theory of mind reasoning in llms for open-ended questions. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 34–44, New York, NY, USA. Association for Computing Machinery.

Anthropic. 2024. Claude: An ai assistant by anthropic. https://assets.anthropic.com/m/61e7d27f8c8f5919/original/Claude-3-Model-Card.pdf. Accessed: 2025-04-10.

Stephen H Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, et al. 2022. Promptsource: An integrated development environment and repository for natural language prompts. *arXiv preprint arXiv:2202.01279*.

Zefan Cai, Po-Nien Kung, Ashima Suvarna, Mingyu Ma, Hritik Bansal, Baobao Chang, P. Jeffrey Brantingham, Wei Wang, and Nanyun Peng. 2024. Improving event definition following for zero-shot event detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2842–2863, Bangkok, Thailand. Association for Computational Linguistics.

Yue Cao and CS Lee. 2023. Robot behavior-tree-based task generation with large language models. *arXiv preprint arXiv:2302.12927*.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Eduardo Gabriel Côrtes. 2024. Beyond accuracy: completeness and relevance metrics for evaluating long answers.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *CoRR*.

DeepSeek-AI. 2024. Deepseek-v3 technical report.

Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, et al. 2025. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *arXiv preprint arXiv:2502.14739*.

Ronen Eldan and Yuanzhi Li. 2023. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*.

Benjamin Feuer and Chinmay Hegde. 2025. Wildchat-50m: A deep dive into the role of synthetic data in post-training.

Saumya Gandhi, Ritu Gala, Vijay Viswanathan, Tongshuang Wu, and Graham Neubig. 2024. Better synthetic data by retrieving and transforming existing datasets. *arXiv preprint arXiv:2404.14361*.

Jiaxin Ge, Xueying Jia, Vijay Viswanathan, Hongyin Luo, and Graham Neubig. 2024. Training task experts through retrieval based distillation. *arXiv preprint arXiv:2407.05463*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Himanshu Gupta, Kevin Scaria, Ujjwala Anantheswaran, Shreyas Verma, Mihir Parmar, Saurabh Arjun Sawant, Chitta Baral, and Swaroop Mishra. 2023. Targen: Targeted data generation with large language models. *arXiv preprint arXiv:2310.17876*.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2022. Unnatural instructions: Tuning language models with (almost) no human labor. *arXiv preprint arXiv:2212.09689*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Mengkang Hu, Pu Zhao, Can Xu, Qingfeng Sun, Jianguang Lou, Qingwei Lin, Ping Luo, and Saravan Rajmohan. 2024. Agentgen: Enhancing planning abilities for large language model based agent via environment and task generation. *arXiv preprint arXiv:2408.00764*.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SciTail: A textual entailment dataset from science question answering. In *AAAI*.

KisanVaani. 2024. agriculture-qa-english-only. https://huggingface.co/datasets/KisanVaani/agriculture-qa-english-only. Accessed: 2025-04-11.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Abdullatif Köksal, Timo Schick, Anna Korhonen, and Hinrich Schütze. 2023. Longform: Effective instruction tuning with reverse instructions. *arXiv preprint arXiv:2304.08460*.

Po-Nien Kung and Nanyun Peng. 2023. Do models really learn to follow instructions? an empirical study of instruction tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1317–1328, Toronto, Canada. Association for Computational Linguistics.

Po-Nien Kung, Fan Yin, Di Wu, Kai-Wei Chang, and Nanyun Peng. 2023. Active instruction tuning: Improving cross-task generalization by training on prompt sensitive tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1813–1829, Singapore. Association for Computational Linguistics.

Po-Nien Kung, Sheng-Siang Yin, Yi-Cheng Chen, Tse-Hsuan Yang, and Yun-Nung Chen. 2021. Efficient multi-task auxiliary learning: Selecting auxiliary data by feature similarity. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 416–428, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2023. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. *arXiv preprint arXiv:2308.12032*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, et al. 2023. Domain specialization as the key to make large language models disruptive: A comprehensive survey. *arXiv preprint arXiv:2305.18703*.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pages 22631–22648. PMLR.

Mingyu Derek Ma, Xiaoxuan Wang, Po-Nien Kung, P Jeffrey Brantingham, Nanyun Peng, and Wei Wang. 2024. Star: boosting low-resource information extraction by structure-to-text data generation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18751–18759.

Mahesh2841. 2024. Agriculture. https://huggingface.co/datasets/Mahesh2841/Agriculture. Accessed: 2025-04-11.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *ACL*.

Rosamond Mitchell and Cynthia Martin. 1997. Rote learning, creativity and'understanding'in classroom foreign language teaching. *Language Teaching Research*, 1(1):1–27.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Ayana Niwa and Hayate Iso. 2024. Ambignlg: Addressing task ambiguity in instruction for nlg. *arXiv preprint arXiv:2402.17717*.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. https://openai.com/blog/chatgpt. Accessed: 2025-04-10.

OpenAI. 2023a. Gpt-4 technical report. https://openai.com/research/gpt-4. Accessed: 2025-04-10.

OpenAI. 2023b. Sharegpt. https://sharegpt.com/. Accessed from https://sharegpt.com/.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. 2024. Infobench: Evaluating instruction following ability in large language models. *arXiv preprint arXiv:2401.03601*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020. Getting closer to AI complete question answering: A set of prerequisite real tasks. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8722–8731. AAAI Press.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision. *Advances in Neural Information Processing Systems*, 36:2511–2565.

Hieu Tran, Zhichao Yang, Zonghai Yao, and Hong Yu. 2024. Bioinstruct: instruction tuning of large language models for biomedical natural language processing. *Journal of the American Medical Informatics Association*, 31(9):1821–1832.

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

Daniel Vila-Suero and Francisco Aranda. 2023. Argilla: Open-source framework for data-centric nlp. https://github.com/argilla-io/argilla. Released on 2023-01-12.

Jiexin Wang, Adam Jatowt, and Masatoshi Yoshikawa. 2022a. Archivalqa: A large-scale benchmark dataset for open-domain question answering over historical news collections. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3025–3035.

Neng Wang, Hongyang Yang, and Christina Dan Wang. 2023. Fingpt: Instruction tuning benchmark for open-source large language models in financial datasets. *arXiv preprint arXiv:2310.04793*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022b. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022c. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *arXiv preprint arXiv:2204.07705*.

Yunjuan Wang, Hussein Hazimeh, Natalia Ponomareva, Alexey Kurakin, Ibrahim Hammoud, and Raman Arora. 2024. Dart: A principled approach to adversarially robust unsupervised domain adaptation. *arXiv preprint arXiv:2402.11120*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Zhengxuan Wu, Yuhao Zhang, Peng Qi, Yumo Xu, Rujun Han, Yian Zhang, Jifan Chen, Bonan Min, and Zhiheng Huang. 2024. Dancing in chains: Reconciling instruction following and faithfulness in language models. *arXiv preprint arXiv:2407.21417*.

Congying Xia, Chen Xing, Jiangshu Du, Xinyi Yang, Yihao Feng, Ran Xu, Wenpeng Yin, and Caiming Xiong. 2024. Fofo: A benchmark to evaluate llms' format-following capability. *arXiv preprint arXiv:2402.18667*.

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*.

Asaf Yehudai, Boaz Carmeli, Yosi Mass, Ofir Arviv, Nathaniel Mills, Assaf Toledo, Eyal Shnarch, and Leshem Choshen. 2024. Genie: Achieving human parity in content-grounded datasets generation. *arXiv preprint arXiv:2401.14367*.

Da Yin, Xiao Liu, Fan Yin, Ming Zhong, Hritik Bansal, Jiawei Han, and Kai-Wei Chang. 2023a. Dynosaur: A dynamic growth paradigm for instruction-tuning data curation. *arXiv preprint arXiv:2305.14327*.

Fan Yin, Jesse Vig, Philippe Laban, Shafiq Joty, Caiming Xiong, and Chien-Sheng Jason Wu. 2023b. Did you read the instructions? rethinking the effectiveness of task definitions in instruction learning. *arXiv preprint arXiv:2306.01150*.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*.

Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. 2023. Causal parrots: Large language models may talk causality but are not causal. *arXiv preprint arXiv:2308.13067*.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, et al. 2023. Huatuogpt, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075*.

Honghua Zhang, Po-Nien Kung, Masahiro Yoshida, Guy Van den Broeck, and Nanyun Peng. 2024a. Adaptable logical control for large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 115563–115587. Curran Associates, Inc.

Xuanming Zhang, Anthony Diaz, Zixun Chen, Qingyang Wu, Kun Qian, Erik Voss, and Zhou Yu. 2024b. Decor: Improving coherence in l2 english writing with a novel benchmark for incoherence detection, reasoning, and rewriting. *arXiv preprint arXiv:2406.19650*.

Jianqun Zhou, Yuanlei Zheng, Wei Chen, Qianqian Zheng, Hui Su, Wei Zhang, Rui Meng, and Xiaoyu Shen. 2024. Beyond content relevance: Evaluating instruction following in retrieval models. *arXiv preprint arXiv:2410.23841*.

Appendix

A Ablations on Subdomain Sampling

We compared subdomain sampling (Section 2.1) against 3 strategies: (i) Word2Vec, (ii) multi-level tree expansion without cosine filtering, and (iii) multi-level tree expansion with varying cosine thresholds $\tau \in \{0.2, 0.4, 0.6, 0.8\}$. As shown in Table 3, Word2Vec degrades beyond the top 200–300 terms, yielding irrelevant tokens or gibberish symbols (e.g., `####`, `@ad`). In contrast, unfiltered multi-level tree expansion introduced off-topic synonyms beyond 2–3 levels. To constrain the expansion to words relevant to our domain keyword, we incorporated cosine filtering and evaluated thresholds of 0.2, 0.4, 0.6, and 0.8. Lower thresholds (e.g., 0.2) enable faster processing but admit a higher proportion of spurious subdomains, while higher thresholds (e.g., 0.6, 0.8) prune aggressively, stalling expansion, adding few new terms beyond level 2. Threshold of **0.4** delivers the best trade-off between the relevance of subdomain words and computational efficiency. Notably, for niche domains such as Mesopotamia, multi-level tree expansion proves superior — Word2Vec returns no viable results—whereas for leukemia, the top 2000 subdomain words from Word2Vec yield only 936 Bing articles.

B Additional information on Datasets - Statistics

We report all dataset attributes – total records, token and verb counts, average instruction/response lengths, and other miscellaneous information(instruction and response generator models) in Table 4. Since the dataset sizes vary widely, raw counts of verbs, unique unigrams, and bigrams can be misleading. To enable fair comparison, we normalize these by computing the average frequency per instruction or response. DomAINS achieves substantially higher per-instance averages for tokens, unigrams, and bigrams, for both instruction and response, demonstrating greater lexical diversity even though it was generated with a smaller LLM compared to the SOTA baselines. Additionally, to assess the extent of domain alignment with the target domain, we measure the domain frequency ratios (average count of domain-keyword unigrams and bigrams per instruction sample). Table 5 presents these averages across all 9 domains for all datasets. SOTA-generated collections exhibit near-zero domain term occurrences, indicating poor alignment. In contrast, DomAINS shows 100–400x higher domain-term frequencies, reflecting strong domain coverage. We observed low scores for Mesopotamia, Music and Astronomy. Upon further analysis, we observed that these domains comprise morphologically related terms (e.g., “Mesopotamian,” “Sumerian”, “musician,” “musical,” “astronomically,” etc.) rather than the exact keyword. Despite this, Figure 17 still confirms strong domain alignment for all the datasets.

C DomAINS framework Task Template

Task meta-data template can be seen in Figure 7. We manually wrote the task meta for 31 tasks. For each task, we define its description, input expectations, positive and negative examples, with reasons highlighting the correctness or flaws for the respective case. For response, we only specify a good reasoning strategy suitable for that task to provide a coherent flow of logical reasoning. Task meta-data can be found in our Github Repository[9].

D DomAINS framework Prompt Templates

Instruction and Response generation prompts are presented in Figure 8 and 9, respectively. Followed by Figure 10 and 11 provides the Instruction and Response evaluation prompts, respectively. Finally, GPT-eval prompt for Win-Rate computations can be seen in Figure 12.

E Bing Search Filters

Initially, we constructed Bing search queries with only *domain* and *subdomain* keywords – *"Blogs on <domain> and <subdomain>"*. However, it often returned undesirable file types (e.g., PDFs, slides, multimedia) and commercial pages (e-commerce sites), which provided little instructional content. To focus the search on only text-rich articles, we then added search filters. Our final query – *"Blogs on <domain> AND <subdomain>" -filetype:pdf -filetype:ppt -filetype:doc -site:amazon.com -site:ebay.com -site:youtube.com -site:vimeo.com"*

[9] Task Meta-data templates - will be released on publication

F DomAINS dataset sample examples

We provide examples from our generated datasets in Figure 13, 14 15 and 16. Entire datasets will be made available on HuggingFace after publication.

G Training hyperparameters and configurations

```
MAX_SEQ_LENGTH = 3000
epochs = 5
lr = 2e-5
PER_DEVICE_TRAIN_BATCH_SIZE = 2
PER_DEVICE_EVAL_BATCH_SIZE = 1
GRAD_ACC_STEPS = 16
optim = "paged_adamw_32bit"
fp16 = True
lr_scheduler_type = "constant_with_warmup"
warmup_steps = 20

## PEFT -> LORA
rank = 256
alpha = 2 * rank
lora_dropout = 0.05
```

H Evaluation dataset

We compile a diverse evaluation suite by combining a variety of publicly available, human-curated benchmarks to rigorously assess our domain expert's performance. For NLI tasks, we include (Nie et al., 2020; Khot et al., 2018) datasets; for behavioral task categories such as brainstorming, information extraction, open-domain QA, and summarization, we add (Conover et al., 2023). For MCQs we add (Du et al., 2025) that introduces newly authored MCQs along with other existing MCQ collections. For reading comprehension across varied genres—news articles, user stories, fiction, blogs, and movie scripts, we incorporate (Rogers et al., 2020; Kočiský et al., 2018; Rajpurkar et al., 2016). Finally, to ensure coverage of underrepresented domains, we add specialized benchmarks: (KisanVaani, 2024; Mahesh2841, 2024) for Agriculture, (Wang et al., 2022a) for History and Mesopotamia, and (Tran et al., 2024) for Leukemia and Virology. Thus, our curated evaluation dataset enables a fair and thorough assessment of the domain-expert models. It must be noted, after our 2-stage filtration, we randomly select at-most 1000 samples for each domain due to budget constraints. Only for Mesopotamia and Leukemia we were able to retrieve 337 and 97 samples, respectively.

I DomAINS datasets additional plots and charts

Figure 18 presents root-verbs and task distribution for all the domains. It is clearly evident from the root-verbs-nouns plots that DomAINS successfully captures the appropriate domain-relevant verb-nouns pairs. DomAINS also attains uniform task coverage compared to skewed distributions in the SOTA frameworks.

Table 3: Ablation study on Subdomain sampling. Table presents top-25 words from Word2Vec, Multi-level tree sampling w & w/o cosine filtering.

Domain	Word2Vec	Multi-level Tree w/o cosine filtering	Multi-level Tree w cosine filtering (0.4)
history	‘foodborne outbreak’, ‘severest downturn’, ‘#,###-####’, ‘pizza’, ‘XXth century’, ‘endeavor’, ‘Mary Anning’, ‘nonfunctional’, ‘imperial overlords’, ‘Henry Wiencek’, ‘scholarly tomes’, ‘historyâ € ™’, ‘bitterest rivalries’, ‘Jamestown colony’, ‘pre Incan’, ‘ancient Babylonians’, ‘Pottawatomie Massacre’, ‘novelist Kingsley Amis’, ‘Chuck Langerman’, ‘pussy’, ‘millenary’, ‘inextricably woven’, ‘momentous’, ‘winningest coach’, ‘landmark Composers Inventors’, ‘lance’, ‘turbulent’	‘Egyptian History’, ‘Renaissance’, ‘Theater’, ‘European History’, ‘Chinese History’, ‘American History’, ‘Egyptian Roman’, ‘Roman Theater’, ‘Theater History’, ‘Roman Epigraphy’, ‘African History’, ‘Cinema History’, ‘Egyptian’, ‘Roman’, ‘African American’, ‘Egyptian Sarcophagi’, ‘Egyptian Synaxarion’, ‘Egyptian Martyrology’, ‘Egyptian Theater’, ‘Kushite Egyptian’, ‘American West’, ‘Roman Republic’, ‘Roman Britain’, ‘unmounted’, ‘Christianity’, ‘Henry Timrod’, ‘sorority’	‘Roman History’, ‘Chinese History’, ‘American History’, ‘Egyptian History’, ‘African History’, ‘European History’, ‘Chinese Theater’, ‘Egyptian Navy’, ‘Roman Navy’, ‘Roman Theater History’, ‘Roman Theater’, ‘Chinese Andragogy’, ‘Chinese Idol’, ‘Chinese Woodcarving’, ‘Chinese Cryology’, ‘Chinese Calligraphers’, ‘American Flag’, ‘Hollywood’, ‘African American’, ‘British’, ‘Cinema’, ‘Western’, ‘Mexican American’, ‘European Theater’, ‘Egyptian Sarcophagi’, ‘Egyptian Stonecarving’
astronomy	’Fran Bagenal’, ’astronautics’, ’reclassify’, ’Hawaii Mauna Kea’, ’galaxy clusters’, ’Capak’, ’geoinformatics’, ’Earthlike planets’, ’KIPAC’, ’Planetary Sciences’, ’J####+####’, ’Limbu’, ’demote Pluto’, ’Dutchman’, ’Lagoon Nebula’, ’Condensed Matter Physics’, ’undercook’, ’Terrestrial Planet Finder’, ’cellular biology’, ’Voorwerp’, ’Astrosat’, ’AB Aurigae’, ’planetary’, ’physicists’, ’sailor’, ’reionization’, ’ASTRONOMY’, ’classical antiquity’, ’Explorer WISE’, ’particle physics experiments’,	’Astronomy’, ’Planetarium’, ’Astronomy Research’, ’Comet Astronomy’, ’Gamma Astronomy’, ’Radio Astronomy’, ’Infrared Astronomy’, ’Ultraviolet Astronomy’, ’Neutrino Astronomy’, ’Neutron Astronomy’, ’Pulsar Astronomy’, ’X Ray Astronomy’, ’Astronaut Training’, ’Gravitational Waves Astronomy’, ’Astronaut Mission’, ’Gravitational Wave Astronomy’, ’Astronaut Selection’, ’Astronaut Helmet’, ’Astronaut Suit’, ’Dynamical System’, ’Astronaut Health’, ’Visual Multiple’, ’Rocket Control System’, ’Rocket Navigation System’, ’Space Physics’, ’Spacewalk Equipment’, ’Spacewalk Maintenance’	’Astronomy’, ’Robotics’, ’Photonics’, ’Astronomy Robotics’, ’Telepresence Robotics’, ’Astronomy Telepresence’, ’Mechatronics’, ’Planetarium’, ’Aerospace Robotics’, ’Teleoperation Robotics’, ’Aerospace’, ’Rocketry’, ’Robotics Haptics’, ’Robotics Telerobotics’, ’Astronomy Robotics System’, ’Space Robotics Telepresence’, ’Telepresence Technology Advancements’, ’Robotics Telepresence Platform’, ’Robotics for Astronomy’, ’Astronomy Research Robotics’, ’Astronomy Robotics Research’, ’Robotics and Mechatronics’, ’Robotics and Astronomy’, ’Astronomy and Robotics’
mesopotamia	*NA*	‘Iraq’, ‘Basra’, ‘Baghdad’, ‘Iraqi’, ‘Mesopotamia’, ‘Iraqi History’, ‘Baghdad History’, ‘Iraq War’, ‘Baghdad Era’, ‘Baghdad University’, ‘Iraqi University’, ‘Baghdad Bombing’, ‘Iraqi City’, ‘Basra City’, ‘Ancient Mesopotamia’, ‘Basra Culture’, ‘Iraqi Culture’, ‘Iraqi People’, ‘Iraqi Heritage’, ‘Iraqi Museum’, ‘Baghdad Museum’, ‘Iraq Museum’, ‘Baghdad Caliphate’, ‘Basra Port’, ‘Iraqi Border’, ‘Iraqi Politics’, ‘Persian History’, ‘Iraqi Agriculture’, ‘Iraqi Fauna’, ‘Iraqi Wildlife’, ‘Iraqi Architecture’, ‘Conquest of Mesopotamia’, ‘Basra Province’,	‘Baghdad’, ‘Iraqi’, ‘Mesopotamia’, ‘Iraqi Dinar’, ‘Basra’, ‘Iraq’, ‘Fallujah’, ‘Ramadi’, ‘Iraqi Dabke’, ‘Kuwait’, ‘Iraqi Dinar Exchange’, ‘Iran Iraq War’, ‘Baghdad History’, ‘Iraqi History’, ‘Islamic Baghdad’, ‘Baghdad Religion’, ‘Iraqi Religion’, ‘Iraqi Sufism’, ‘Sufi Mesopotamia’, ‘Baghdad Empire’, ‘Iraq War’, ‘Iraqi War’, ‘Baghdad Era’, ‘Kuwait War’, ‘Iraqi Insurgency’, ‘Baghdad Times’, ‘Baghdad Old’, ‘Baghdad Institute’, ‘Baghdad Bombing’, ‘Fallujah Bombings’, ‘Baghdad Bombings’, ‘Basra City’, ‘Baghdad City’, ‘Iraqi City’, ‘Fallujah City’, ‘Iraqi Army’

Table 4: Additional information on the Datasets – Stats and miscellaneous information.

(a) **Instruction Stats**

Dataset	#records	#tokens	#verbs	avg #token/ins	avg #verb/ins	uniq unigrams / ins	uniq bigrams / ins	Ins. Generator
SuperNI	61.8k	355M	45M	171	22	90	146	human
Self-Instruct	82.6k	4.62M	608k	31.91	3.99	25.45	29.99	text-davinci-003
Magpie-Pro	300k	198M	26M	16	2.5	15	15	Llama 3 70B Instruct
Evol-Instruct-70k	70k	26M	3.0M	99	12	57	83	gpt-3.5-turbo
Evol-Instruct-196k	196k	62M	7.6M	111	13.5	62	92	-
Longform-C	27.7k	12M	1.8M	172	24	74	135	text-davinci-003
OpenHermes	243k	64M	7.1M	55	6.3	36	48	~ GPT-4
UltraChat-200k	200k	220M	31M	223	32	117	193	GPT-3.5 and GPT-4
WildChat	1.5M	382M	39M	250	25	98	180	human
Mesopotamia	100k	63.4M	7.1M	236	27.4	119	198	Llama 3.1 8B Instruct
Fish	100k	60.0M	7.7M	217	29.9	113	187	Llama 3.1 8B Instruct
History	100k	63.9M	7.3M	239	28.7	123	203	Llama 3.1 8B Instruct
Art	100k	61.1M	7.6M	220	29.0	115	190	Llama 3.1 8B Instruct
Leukemia	100k	66.5M	8.1M	272	34.9	134	231	Llama 3.1 8B Instruct
Astronomy	100k	59.3M	7.3M	210	27.1	111	179	Llama 3.1 8B Instruct
Virology	100k	61.6M	7.4M	229	28.2	119	195	Llama 3.1 8B Instruct
Agriculture	100k	60.5M	7.8M	218	29.8	117	189	Llama 3.1 8B Instruct
Music	100k	59.7M	7.2M	208	26.4	111	179	Llama 3.1 8B Instruct

(b) **Response Stats**

Dataset	avg len/res	avg #tokens/res	uniq uni/res	uniq bi/res	Res. Generator
SuperNI	35	8	5.5	6.2	human
Self-Instruct	24.03	24.03	16.23	20.31	text-davinci-003
Magpie-Pro	3270	645	246	471	Llama 3 70B Instruct
Evol-Instruct-70k	1356	266	106	190	gpt-3.5-turbo
Evol-Instruct-196k	1705	321	125	230	-
Longform-C	1756	347	159	293	C4, WikiHow, Eron, BEA-2019, EL5 datasets
OpenHermes	1000	210	81	147	~ GPT-4
UltraChat-200k	4486	838	282	589	ChatGPT
WildChat	2681	471	172	330	GPT-3.5 and GPT-4
Mesopotamia	2129	398	134	248	Llama 3.1 8B Instruct
Fish	1991	382	134	245	Llama 3.1 8B Instruct
History	2128	399	136	250	Llama 3.1 8B Instruct
Art	2071	391	137	251	Llama 3.1 8B Instruct
Leukemia	2135	393	136	248	Llama 3.1 8B Instruct
Astronomy	2104	384	135	245	Llama 3.1 8B Instruct
Virology	2151	387	134	243	Llama 3.1 8B Instruct
Agriculture	2132	387	137	249	Llama 3.1 8B Instruct
Music	2031	389	136	250	Llama 3.1 8B Instruct

Table 5: Domain Frequency Ratio for all the 9 domains. We computed the avg. freq. of **domain** keyword and bi-grams(containing **domain** keyword) per instruction instance for all SOTA and DomAINS datasets.

Dataset	Music		Leukemia		Astronomy		History	
	freq	bigram	freq	bigram	freq	bigram	freq	bigram
SuperNI	0.023	0.040	0.000	0.001	0.000	0.001	0.028	0.054
Magpie-Pro	0.005	0.011	0.000	0.000	0.001	0.013	0.011	0.022
Evol-Instruct-70k	0.008	0.014	0.000	0.000	0.001	0.017	0.022	0.040
Evol-Instruct-196k	0.014	0.024	0.000	0.000	0.007	0.012	0.025	0.047
Longform-C	0.012	0.021	0.006	0.009	0.018	0.031	0.032	0.059
OpenHermes	0.004	0.008	0.000	0.000	0.003	0.007	0.083	0.158
UltraChat-200k	0.076	0.122	0.004	0.007	0.012	0.021	0.070	0.125
WildChat	0.014	0.020	0.003	0.005	0.002	0.004	0.017	0.028
DomAINS	0.048	0.072	2.898	3.878	0.375	0.593	0.687	1.102

Dataset	Mesopotamia		Agriculture		Art		Fish		Virology	
	freq	bigram	freq	bigram	freq	bigram	freq	bigram	freq	bigram
SuperNI	0.000	0.000	0.001	0.002	0.006	0.011	0.007	0.012	0.000	0.000
Magpie-Pro	0.000	0.000	0.001	0.002	0.006	0.011	0.000	0.001	0.000	0.000
Evol-Instruct-70k	0.000	0.000	0.003	0.006	0.008	0.015	0.005	0.008	0.000	0.000
Evol-Instruct-196k	0.000	0.000	0.003	0.006	0.012	0.022	0.004	0.007	0.000	0.000
Longform-C	0.000	0.000	0.002	0.004	0.007	0.011	0.008	0.014	0.000	0.000
OpenHermes	0.000	0.000	0.001	0.002	0.004	0.008	0.003	0.005	0.000	0.000
UltraChat-200k	0.000	0.000	0.008	0.013	0.054	0.085	0.015	0.025	0.000	0.000
WildChat	0.000	0.000	0.001	0.002	0.128	0.248	0.009	0.016	0.000	0.000
DomAINS	0.002	0.003	1.454	2.204	3.648	5.050	4.133	5.741	0.340	0.529

We split the table into 2 halves to fit it on the page. Both the sub-tables present the same experimental results.

TASK META-DATA TEMPLATE

```
{ #task_id
    "instruction" : { ## all the details for generating the instruction ##
        "type": " ## task type ## ",
        "desc": "## description on how to generate the instruction ##",
        "positive_example": {
            "content": "## bing article section ##",
            "response": "## generated instruction ##",
            "explanation": "## why the generated instruction is correct ##",
        },
        "negative_example": {
            "content": "## bing article section ##",
            "response": "## generated instruction ##",
            "explanation": "## why the generated instruction is incorrect ##",
        },
        "instruction_format": "## specify if there's any specific format ##",
    },

    "response" : { ## details for generating a response for the instruction ##
        "response_format": "## specify if there's any specific format ##",
        "example" : "## example response with step by step reasoning",
    },

    "cot" : ' ## step by step reasoning strategy ## ' ,
},
```

Figure 7: Task meta-data template

INSTRUCTION GENERATION PROMPT

<| begin_of_text |><| start_header_id |> *system* <| end_header_id |>

A chat between a curious user and an artificial intelligence assistant over a ***{domain}*** domain. A domain is the field or sector on which the user wishes to restrict the entire conversation.

The assistant generates a **good instruction** for the instruction following-model referring the given `meta-data` for the Bing article and task category limiting it to the given ***{domain}*** domain and ***{sub_domain}*** subdomain topics.

TASK CATEGORY META-DATA:
{instruction_category}

Assistant makes use of the given *bing article* of the following format:

- `keywords` - tells the domains used to describe the data
- `name` - title of the article
- `content` - actual information expounding the keywords
- `url` - from where the content is retrieved
- `preview` - thumbnail or short preview of the content of the article
- `positive_example` - Comprise content containing a sample passage, CORRECT response of what is expected, and rationale behind why it is correct
- `negative_example` - Comprise content containing a sample passage, INCORRECT response of what is expected, and rationale behind why it is incorrect. The assistant should prevent question like these.\n

BING ARTICLE META-DATA:
{bing_articles}

Assistant must go through the content THOROUGHLY and use any segment of the content to come up with the instruction.

OUTPUT FORMAT: The output should only be an instruction as described in the `TASK CATEGORY META-DATA`. The output should abide the following template:

INSTRUCTION: ``generated instruction response``

NOTE: We do not want anyone to know about the meta-data, articles or any other information. Assume you already knew about it. Do not leak any infiormation about the source. No need of any introduction or hallucination.

<| eot_id |><| start_header_id |> *assistant* <| end_header_id |> INSTRUCTION:

Figure 8: Instruction Generation Prompt

RESPONSE GENERATION PROMPT

<| begin_of_text |> <| start_header_id |> *system* <| end_header_id |>

A chat between a curious user and an artificial intelligence assistant over a ***{domain}*** domain. A domain is the field or sector on which the user wishes to restrict the entire conversation.

The assistant generates a **good response for the instruction** by THOROUGHLY referring to the given `meta-data` for the Bing article and limiting it to the given ***{domain}*** domain.

Format of the bing article meta-data is as follows:
- `keywords` - tells the domains used to describe the data\n
- `name` - title of the article \n
- `content` - actual information expounding the keywords\n
- `url` - from where the content is retrieved \n
- `preview` - thumbnail or short preview of the content of the article\n
- `positive_example` - Comprise content containing a sample passage, CORRECT response of what is expected, and rationale behind why it is correct\n
- `negative_example` - Comprise content containing a sample passage, INCORRECT response of what is expected, and rationale behind why it is incorrect\n

BING ARTICLE META-DATA:
{bing_articles}

The response must provide a valid solution, reasoning or missing information to the asked instruction by strictly abiding to the given content in the Bing article. Do not hallucinate.

OUTPUT FORMAT: The output should be a detailed, logical, and succinct answer to the user's instruction question. The output should abide the following template:

{response['response_format']}

RESPONSE EXAMPLE: {response['example']}

NOTE: We do not want anyone to know about the meta-data, articles or any other information. Assume the assistant already knew the information. Response should not mention phrase like "based on given meta-data, articles" or "according to the..." etc. Do not leak the source of the information. "

<| eot_id |><| start_header_id |> *assistant* <| end_header_id |>
INSTRUCTION: ***{instruction}***
RESPONSE:

Figure 9: Response Generation Prompt

INSTRUCTION RATING PROMPT

<| begin_of_text |><| start_header_id |> *system* <| end_header_id |>

You are a skilled human evaluator. I will give you an INSTRUCTIONs. Your job is to assess it based on the metrics provided. Finally, return it in the JSON format. Do not be lenient.

<| eot_id |>

<| start_header_id |> *user* <| end_header_id |>

METRICS:
- Coherence: Assess if the instruction is logically structured and consistent, ensuring that all parts contribute to a unified goal. Rate on a scale of 1 (severe disjointedness) to 5(cogent, well-organized).
- Ambiguity: Evaluate whether the instruction is clear or vague. Multiple interpretations could lead to confusion. Scale 1 (extremely ambiguous) to 5 (completely unambiguous).
- Difficulty: Assess the level of cognitive challenge required to complete the instruction. Scale 1 (shallow and direct) where little thought is required to 5 (challenging) demands significant analysis and thought.
- Completeness: Determine if the instruction provides all the necessary information to complete the task. Scale 1 (ill-defined) to 5(thorough).
- Quality: Finally, consider the overall readability, correctness, writing style, and how well the instruction is presented. Labels: ['terrible', 'poor', 'average', 'good', 'excellent'].

```
INSTRUCTION:
<< INS >>
```

OUTPUT FORMAT:
- Type: JSON
- ONLY valid keys for JSON: ['coherence', 'coherence_reason', 'completeness', 'completeness_reason', 'relevance', 'relevance_reason', 'adherence', 'adherence_reason', 'cot_strategy', 'cot_strategy_reason', 'quality']
- all keys in lowercase
- Metrics should be strictly within the range.
- labels for quality are all lowercase nd stick to the given labels.
- Example: {'coherence': 5, 'coherence_reason': 'instruction is logically organized and clearly states what is expected.', 'ambiguity': 4, 'ambiguity_reason': 'Most evident. Somewhat subjective but still understandable', 'difficulty': 3, 'difficulty_reason': 'Moderate challenge in justifying the answer based on the passage', 'completeness': 5, 'completeness_reason': 'All needed context is provided.', 'quality': 'excellent'}

<| eot_id |>

<| start_header_id |> *assistant* <| end_header_id |>

Figure 10: Instruction Rating Prompt

RESPONSE RATING PROMPT

<| begin_of_text |> <| start_header_id |> *system* <| end_header_id |>

You are a skilled human evaluator. I will give you an INSTRUCTION and its RESPONSE. Your job is to assess the RESPONSE based on the following metrics. Finally, provide a score for each metric and finally give its quality label. Do not be lenient.

<| eot_id |>

<| start_header_id |> *user* <| end_header_id |>

METRICS:
- coherence: Assess whether the response is logically structured, consistent, and easy to follow. Rate on a scale of 1 (severely disjointed) to 5 (cogent and well-organized).
- completeness: Evaluate if the response thoroughly addresses all necessary points. Rate on a scale of 1 (shallow) to 5 (comprehensive).
- relevance: Assess if the response directly answers the question in the instruction without unnecessary deviations or beating around the bush. Rate on a scale of 1 (irrelevant) to 5 (to the point).
- adherence: Measure how well the response follows the prompt's specific requirements, intent, and format constraints. Rate on a scale of 1 (noncompliant) to 5 (fully aligned/conforming).
- cot_strategy: Assess if the response demonstrates a clear and logical chain-of-thought (CoT) reasoning process. Rate on a scale of 1 (absent) to 5 (thorough and well-reasoned).
- quality: Finally, provide an overall assessment of the response, considering readability, correctness, reasoning effectiveness, and adherence to the evaluation criteria. Assign one of the following labels: ['terrible', 'poor', 'average', 'good', 'excellent'].

-->
```
**INSTRUCTION:**
***&lt;&lt; INS &gt;&gt;***
```

-->
```
**RESPONSE:**
**&lt;&lt;res&gt;&gt;**
```

OUTPUT FORMAT:
- Type: JSON
- ONLY valid keys for JSON: ['coherence', 'coherence_reason', 'completeness', 'completeness_reason', 'relevance', 'relevance_reason', 'adherence', 'adherence_reason', 'cot_strategy', 'cot_strategy_reason', 'quality']
- all keys in lowercase
- Metrics should be strictly within the range.
- labels for quality are all lowercase and stick to the given labels.
- Example: {'coherence': 3, 'coherence_reason': 'The response is somewhat structured but lacks smooth transitions between ideas.', 'completeness': 3, 'completeness_reason': 'Some key points are covered, but the response lacks depth and supporting details.', 'relevance': 4, 'relevance_reason': 'The response mostly addresses the instruction but includes some unnecessary information.', 'adherence': 2, 'adherence_reason': 'The response does not fully follow the specified format and partially misses the intent.', 'cot_strategy': 2, 'cot_strategy_reason': 'Limited logical reasoning is present; the response lacks a step-by-step approach.', 'quality': 'average', 'quality_reason': 'While readable and somewhat relevant, the response is lacking in depth, reasoning, and strict adherence to the prompt.'}

<| eot_id |>

<| start_header_id |> *assistant* <| end_header_id |>

Figure 11: Response Rating Prompt

GPT EVALUATION PROMPT

You are a skilled human evaluator. I will give you an INSTRUCTION and 2 RESPONSES. Your objective is to evaluate which response better satisfies the INSTRUCTION based on the metrics provided. Finally, provide a `PREFERENCE`.

METRICS:

- Completeness: Evaluate if the response thoroughly, coherent and addresses all necessary points. Rate on a scale of 1 (shallow) to 5 (comprehensive).
- Relevance: Assess if the response directly answers the question in the instruction without unnecessary deviations or beating around the bush. Rate on a scale of 1 (irrelevant) to 5 (to the point).
- Format-following: If the question asks for Step by Step reasoning, then we evaluate the comprehensive CoT Strategy. Assess if the response demonstrates a clear and logical chain-of-thought (CoT) reasoning process. Rate on a scale of 1 (absent) to 5 (thorough and and well reasoned).

PREFERENCE SCALE:

- **1:** Overall, Response 1 is the better output.
- **2:** Overall, Response 2 is the better output.
- **0:** Both predictions are equally good or equally bad.

``````

####################
####################

## INSTRUCTION : ***{0}***

####################
####################

## RESPONSE 1: ***{1}***

####################
####################

## RESPONSE 2 : ***{2}***

####################
####################

``````

Guidelines:
1. Read the provided INSTRUCTION and both RESPONSES thoroughly.
2. Use the provide METRICS to evaluate both the responses.
3. Final Judgment: `PREFERENCE`.
4. Return the output in the given OUTPUT FORMAT.

OUTPUT FORMAT: REASON : *<<generated reasoning based on metrics for the preference order>>* \n RESPONSE: PREFERENCE : *<< 0 / 1 / 2 >>*

Figure 12: GPT-4 Win Rate Evaluation prompt template

INSTRUCTION:

Generate relevant topic words for the given question.

PASSAGE: The Northern Grapes Project was funded in 2011 by the USDA NIFA Specialty Crops Research Initiative to address the challenges of cold hardy, Vitis riparia-based wine grape cultivars in the rapidly-expanding industry of small vineyard and winery enterprises in the Northeast and Midwest. The project aimed to optimize viticultural and enological practices, as well as marketing strategies, to increase the profitability and sustainability of emerging cold climate grape and wine industries. The research integrated production, distribution, and processing, and consumers and markets to elucidate and build on the relationships between these focus areas. Ultimately, the project helped producers overcome production and marketing constraints and increase the profitability and sustainability of the emerging industries.

Question: What are some good topic words related to the Northern Grapes Project's objectives and outcomes?

RESPONSE:

Step-by-Step Solution:

1. **Understand the Passage:** The passage describes the Northern Grapes Project, a research initiative focused on optimizing viticultural and enological practices, marketing strategies, and production processes for cold climate grape and wine industries in the Northeast and Midwest.

2. **Analyze the Question:** The question asks for good topic words related to the Northern Grapes Project's objectives and outcomes.

3. **Extract Relevant Information:**
- Key objectives: optimizing viticultural and enological practices, marketing strategies, and production processes.
- Focus areas: production, distribution, processing, consumers, and markets.
- Outcomes: increased profitability and sustainability of emerging industries.

4. **Identify Core Themes and Subjects:**
- Viticulture and enology practices
- Marketing strategies
- Production processes
- Sustainability and profitability

5. **Filter for Specificity and Relevance:**
- Focus on specific terms related to the project's objectives and outcomes.
- Avoid generic terms like "research" or "industry."

FINAL RESPONSE - TOPIC WORDS: Viticulture, Enology, Marketing, Sustainability, Profitability, Production, Distribution, Processing.

Figure 13: Dataset example from Domain="Agriculture," subdomain="Viticulture Practices," and task="Topic Words Generation"

INSTRUCTION:

The passage below contains intentional typos and grammatical errors. Identify and rectify the mistakes. PASSAGE: Ancint Chinese rituals are an importint part of Chinese culture and have been practiced for senturies. These rituals include ancestor worship, worship of the heavens, eating rituals, and birth rituals. They reflect the values and beliefs of the Chinese people and are often used to show respect, promote harmony, and seek guidance from the spiritual world.

RESPONSE:

Step-by-Step Solution:

1. **Read the Passage Thoroughly:**
- The passage discusses ancient Chinese rituals and their significance in Chinese culture.

2. **Identify Errors Systematically:**
- Sentence 1: "Ancint Chinese rituals are an importint part of Chinese culture and have been practiced for senturies." - Error 1: "Ancint" should be "Ancient", "importint" should be "important", and "senturies" should be "centuries."
- Sentence 2: "These rituals include ancestor worship, worship of the heavens, eating rituals, and birth rituals." - No major errors detected.
- Sentence 3: "They reflect the values and beliefs of the Chinese people and are often used to show respect, promote harmony, and seek guidance from the spiritual world." - No major errors detected.

3. **Correct Errors:**
- "Ancint" → "Ancient"
- "importint" → "important"
- "senturies" → "centuries"

4. **Review for Contextual Fit:**
- Verified that each corrected sentence aligns with the passage's context and retains its intended meaning.
- Ensured logical flow between sentences.

5. **Check for Consistency:**
- Checked terminology and consistent formatting.
- Verified appropriate tone and grammatical style throughout the passage.

6. **Polish the Passage:**
- Re-read the entire passage for fluency and professional tone.

FINAL RESPONSE: Ancient Chinese rituals are an important part of Chinese culture and have been practiced for centuries. These rituals include ancestor worship, worship of the heavens, eating rituals, and birth rituals. They reflect the values and beliefs of the Chinese people and are often used to show respect, promote harmony, and seek guidance from the spiritual world.

Figure 14: Dataset example from Domain="History," subdomain="ancient Chinese rituals," and task="Typo Rectification"

INSTRUCTION:

PASSAGE: The Roman Empire's rich and multifaceted visual culture is a manifestation of the sprawling geography of its provinces. In 2011, a group of twenty international scholars began a multi-year research seminar to study, discuss, and ponder the nature and development of art and archaeology in the Roman provinces. Their compelling research resulted in a book titled Beyond Boundaries: Connecting Visual Cultures in the Provinces of Ancient Rome. The book is the result of an international seminar on the arts of Rome's provinces, funded by the Getty Foundation, and included some twenty contributors from eleven different countries. The goal of the seminar was to consider the artistic remains of the ancient Roman provinces and how they can complicate and contradict our simple understanding of relations between the imperial center—that is, Rome—and the provincial periphery.

Question: What would be a good title that captures the main theme of the research seminar on the arts of Rome's provinces?

RESPONSE:

Step-by-Step Solution:

1. **Understand the Passage:** The passage describes a research seminar on the arts of Rome's provinces, highlighting its international scope, contributors, and goal of exploring the relationship between the imperial center and the provincial periphery.

2. **Central Theme:** The central theme of the research seminar is to understand the relationship between the Roman Empire's center and its provinces, specifically through the lens of art and archaeology.

3. **Identify the Question Requirements:** The question asks for a suitable title that captures the essence of the research seminar's main theme, which is the relationship between the Roman Empire's center and its provinces.

4. **Extract Key Components from the Passage:**
- The research seminar aimed to study the nature and development of art and archaeology in the Roman provinces.
- The seminar considered the artistic remains of the ancient Roman provinces and their impact on our understanding of the relationship between the imperial center and the provincial periphery.

5. **Topic Words:** Roman Empire, provinces, art, archaeology, center, periphery, relationship.

6. **TITLE:** "Beyond the Periphery: Exploring the Roman Empire's Visual Culture."

This title captures the main theme of the research seminar by highlighting the relationship between the Roman Empire's center and its provinces, and the exploration of the visual culture that reflects this relationship.

Figure 15: Dataset example from Domain="History," subdomain="Roman Empire," and task="Title Generation"

INSTRUCTION:

Given the passage, CONFIRM or REFUTE the CLAIM.

PASSAGE: 'The killifish genus Valencia is the only fish genus found only in Europe and the Mediterranean basin, and the recent description of a new member brings the total number of species to three. This trio forms the family Valenciidae and all members are critically-endangered with restricted natural ranges within which they have been subjected to various threats.'
CLAIM: 'All Valencia killifish species are critically endangered.'

RESPONSE:

Step by Step Solution:

1. **Understand the Claim:** Core Assertion: The claim states that all Valencia killifish species are critically endangered.
Key Components:
- Species: Valencia killifish
- Condition: Critically endangered

2. **Analyze the Passage:** Locate relevant evidence in the passage:
- The passage mentions that the killifish genus Valencia has three species.
- It states that all members of the family Valenciidae (which includes the Valencia genus) are critically-endangered.

3. **Verify the Claim:**
- Explicit Evidence: The passage confirms that all members of the family Valenciidae are critically-endangered.
- Implicit Evidence: Since the Valencia genus is part of the family Valenciidae, it can be inferred that all Valencia killifish species are also critically-endangered.
- Gaps or Contradictions: There are no significant contradictions or missing elements in the passage regarding the claim.

4. **Evaluate the Claim:**
Conclusion: The claim is CONFIRM.
Rationale: The passage explicitly states that all members of the family Valenciidae are critically-endangered, and since the Valencia genus is part of this family, it can be inferred that all Valencia killifish species are also critically-endangered.

5. **Write the Verification:**
Conclusion: CONFIRM.
Rationale: The passage confirms that all Valencia killifish species are critically-endangered, as they are part of the family Valenciidae, which includes all critically-endangered members.

6.**Review for Clarity and Accuracy.**

FINAL RESPONSE: CONFIRM

Figure 16: Dataset example from Domain="Fish," subdomain="fish remorse," and task="Fact Verification"

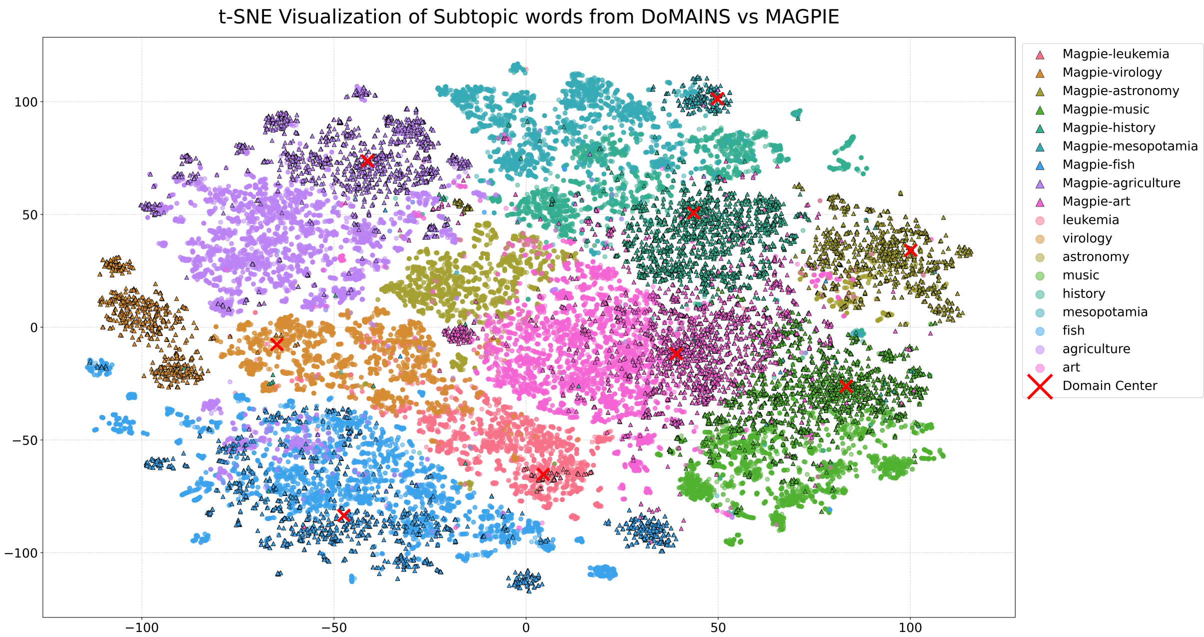
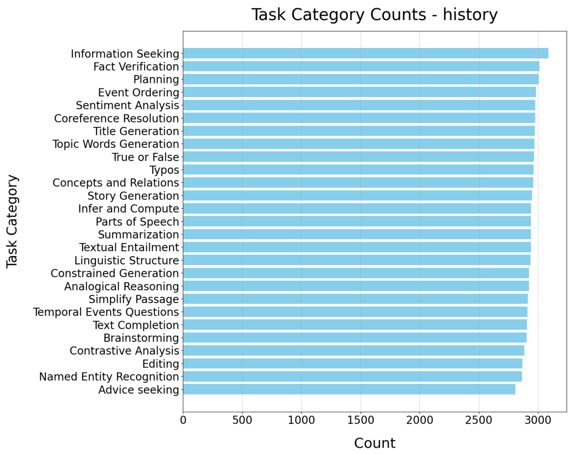


Figure 17: Extended Domain alignment plot of our generated dataset against Magpie over all 9 domains.

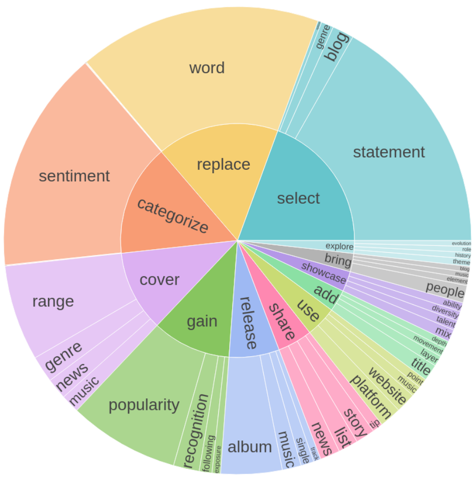
history

Root-Verb – history

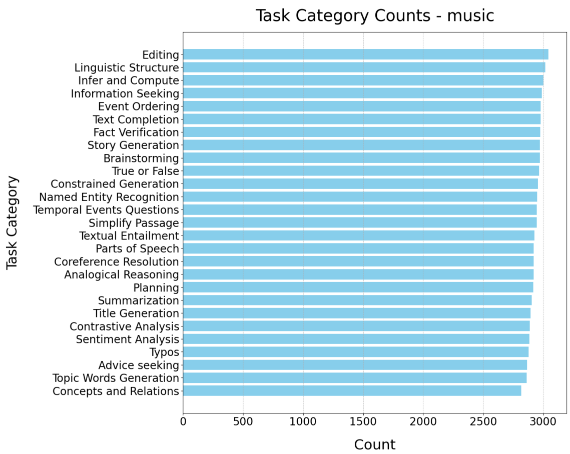


Tasks Distribution – history

music

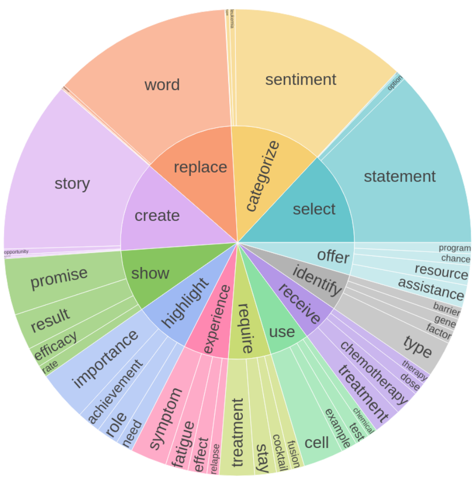


Root-Verb – music

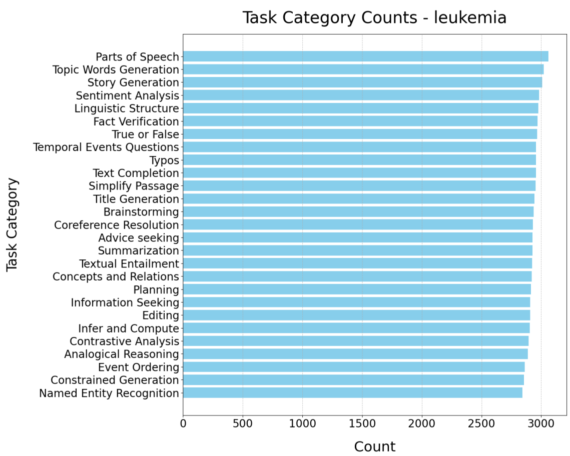


Tasks Distribution – music

leukemia

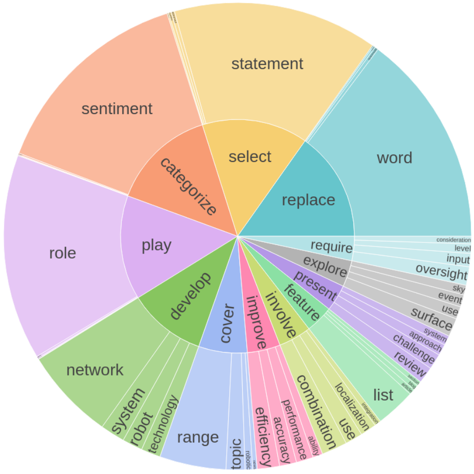


Root-Verb – leukemia

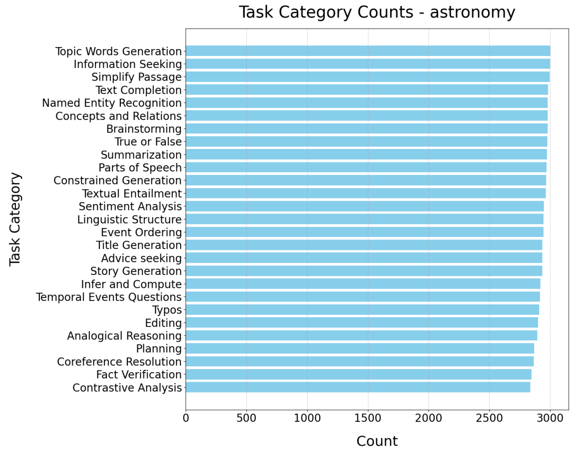


Tasks Distribution – leukemia

astronomy

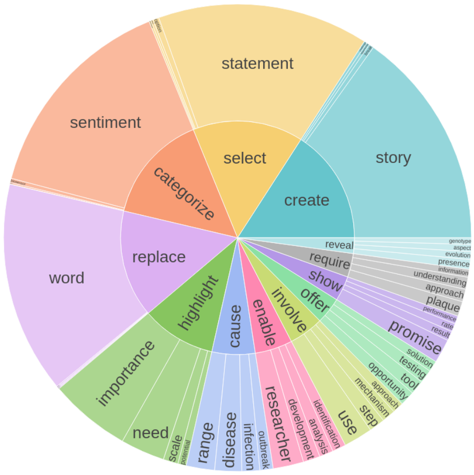


Root-Verb – astronomy

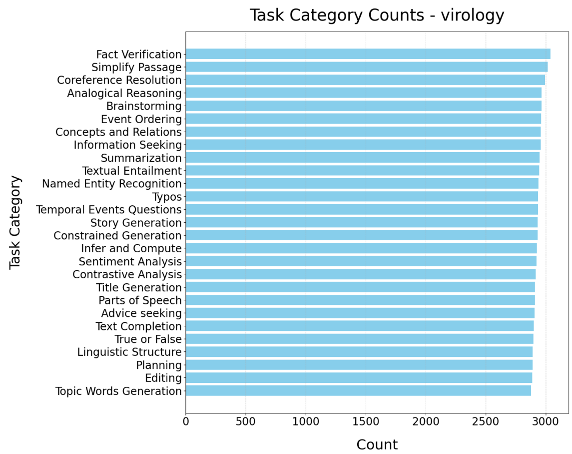


Tasks Distribution – astronomy

virology



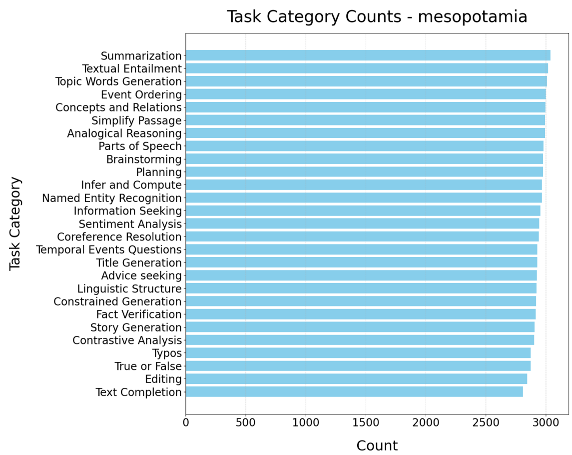
Root-Verb – virology



Tasks Distribution – virology

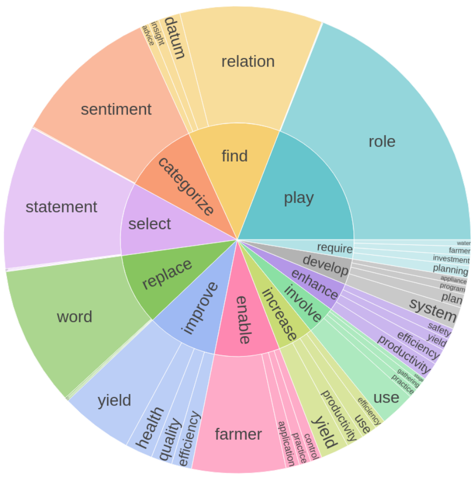
Mesopotamia

Root-Verb – Mesopotamia

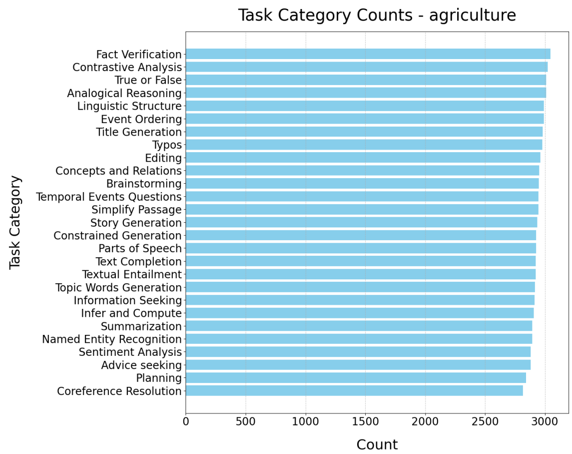


Tasks Distribution – Mesopotamia

agriculture

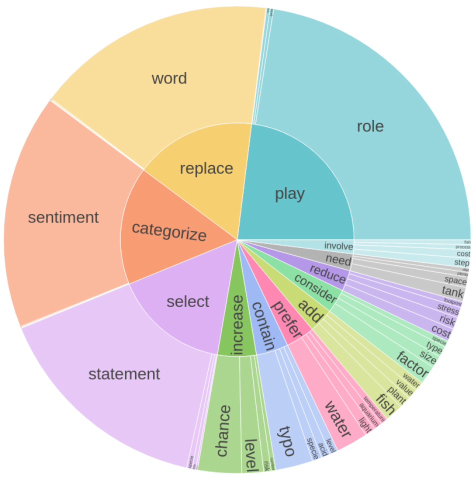


Root-Verb – agriculture

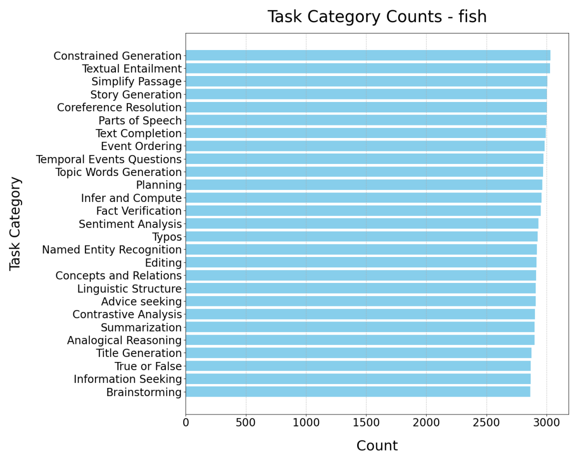


Tasks Distribution – agriculture

fish



Root-Verb – fish



Tasks Distribution – fish

Figure 18: Root-Verb and Task-Bar plots for each domain.