

A Survey of LLM Prompt Datasets: Taxonomy, Linguistic Patterns, and Practical Uses

Yuanming Zhang^{1*} Yan Lin^{2*} Arijit Khan^{3,2†} Huaiyu Wan¹

¹School of Computer Science and Technology, Beijing Jiaotong University, Beijing, China

²Aalborg University, Aalborg, Denmark

³Bowling Green State University, Ohio, USA

{ymzhang23, hywan}@bjtu.edu.cn, lyan@cs.aau.dk,
arijitk@bgsu.edu

Abstract

We compile 129 public LLM prompt datasets with more than 1.22 TB and more than 673M instances and organise them into a unified taxonomy. We use seven datasets for detailed analysis and identify lexical, syntactic, and semantic patterns that distinguish prompts from general text. We evaluate these features in prompt filtering, source domain routing, and elicited response quality analysis. A 63 dimensional linguistic feature set that can be extracted on a CPU achieves over 91% of the F1 of GPU sentence embeddings and reduces latency per request by nearly half. In cross dataset routing, sentence embeddings reach 0.74 Macro-F1 and a structural feature subset reaches 0.61, compared with the chance level of 0.20. Structural routing features remain negatively associated with elicited response quality after prompt length is controlled on UltraFeedback. The apparent positive association of type to token ratio becomes much smaller under the same control. These findings support prompt processing pipelines that combine efficient structural features with robust semantic routing. Our catalogue, annotations, and code are available at <https://github.com/ymzhang-cs/prompt-dataset-analysis>.

1 Introduction

Recent advances in large language models (LLMs) have led to many custom prompts across GitHub repositories such as f/awesome-chatgpt-prompts (Akin, 2022), Reddit forums such as ChatGPTPromptGenius (community, 2023), commercial platforms such as PromptBase (PromptBase, 2024), and research releases (Conover et al., 2023; Chen et al., 2024a). Few studies examine this growing collection as a whole. The variation in prompt vocabulary, syntax, and semantics across sources

and its effects on prompt filtering, source domain routing, and elicited response quality analysis remain unclear.

To address this gap, we compile and analyse 129 diverse prompt datasets with more than 1.22 TB and more than 673M instances. We derive a hierarchical taxonomy and evaluate the extracted linguistic features in prompt filtering, source domain routing, and elicited response quality analysis. The 63 dimensional linguistic feature set can be extracted on a CPU and recovers more than 91% of GPU embedding F1. Its latency per request is 1.9 times lower at 3.0 ms compared with 5.7 ms. We state separately which findings use the full catalogue, the seven datasets, and the data for each application.

2 Related Work

Datasets for LLMs. Liu et al. (2025a) survey a broad range of LLM datasets and focus on training and model tuning corpora. Zhang et al. (2026) and OpenCodeInstruct (Ahmad et al., 2025a) focus on instruction tuning collections. Luo et al. (2025) organise instruction dataset quality evaluation into human, statistical, model based, and LLM based methods. Work on individual resources includes AuToGen, which generates tool learning examples from domain specific structured data (Zeng et al., 2025), and MSEED, which constructs preference data for safety alignment and multidimensional evaluation (Wang and Sun, 2025). Adjacent reviews survey capabilities such as mathematical problem solving (He et al., 2025), and topic analyses map themes in ChatGPT research (Zhang et al., 2025). These studies evaluate training data, construct particular resources, or map adjacent research; they do not provide a linguistic analysis at multiple levels for prompt datasets at this scale.

Prompt engineering tools. Schulhoff et al. (2024) survey prompt engineering techniques. PromptAid (Mishra et al., 2025) and PromptLand-

*Equal contribution.

†Corresponding author.

scape (Wang et al., 2024a) provide visual tools. Promptware (Chen et al., 2026) applies software engineering principles, and PromptAgent (Wang et al., 2024b) creates prompts for expert tasks. We characterise complete dataset collections, while these studies develop individual prompts or tools.

3 Prompt Datasets Discovery and Refinement

Data discovery guideline. We target datasets that meet three conditions. First, they contain natural language instructions that describe a task and guide an LLM towards a desired output. Second, they cover domains that range from everyday tasks such as travel planning to professional and specialised tasks such as academic writing, healthcare, and finance. Third, they include several prompt forms, such as single instructions and conversations with multiple turns.

Data discovery process. We collect public datasets from four source types. Searches on Hugging Face (Lhoest et al., 2021), Kaggle, Google Dataset Search (Noy et al., 2019), and Papers with Code yield 60 datasets. We query “prompt dataset” and “instruction following dataset.” Publications on prompt engineering and dialogue systems at NeurIPS, ICLR, and ICML from 2023 through 2025 contribute 73 datasets. GitHub repositories add 21 datasets. These repositories include Awe-some Instruction Datasets (Nie, 2023), Prompt Engineering Guide (Saravia, 2022), and LLM-DataHub (Zhao, 2023). Prompt sharing websites contribute 14 datasets. Examples include Prompt Genius (Genius, 2024) and BoredHumans (BoredHumans, 2024).

Data filtering. We first remove duplicate catalogue entries. For example, CVQA (Romero et al., 2024) appears on Hugging Face and in NeurIPS 2024. We then apply four criteria. Curated datasets must contain at least 1K prompts, and datasets shared by users must contain at least 50. We exclude samples with inconsistent formatting or unclear structure, including illustrative prompts scattered through Prompt Engineering Guide. We retain prompts tied to clearly defined tasks and omit collections such as PersonaHub (Ge et al., 2024), which mostly contains persona descriptions without task instructions. Each dataset must be publicly available under a license that allows research. This process yields 129 prompt datasets for the taxonomy in §4.

4 Dataset Taxonomy

We categorise the prompt datasets along the Source, Content, and Target axes. These axes form the hierarchical taxonomy in Figure 1. Appendix G gives details for each category.

Source captures publisher, release channel, and generation process. Publishers include *end users*, *LLM researchers*, and *domain scientists*. Examples are Prompt Genius, OpenMathReasoning, and ChatGPT Data Science Prompts (Tang, 2023). Release channels include GitHub, Hugging Face, Kaggle, prompt sharing platforms, and social media. Prompt sharing platforms may provide open or commercial access. Generation processes include human writing, model generation, and dataset derivation. Examples include databricks-dolly-15k, Self Instruct (Wang et al., 2023), simulations with several agents (Li et al., 2023), and Flan 2022 (SirNeural, 2022).

Content captures linguistic and structural attributes. Language may be monolingual or multilingual. Display form covers instructions and conversations with one, several, or branching turns. Format may be free form, structured, or mixed, and it can affect response quality (Liu et al., 2025b). Placeholders support template substitution (Shin et al., 2020). Prompt engineering methods include few shot prompting (Brown et al., 2020), role play (Zhang et al., 2018), and chain of thought prompting (Wei et al., 2022). Other attributes include safety labels, metadata, ground truth, update patterns, and modality. UltraFeedback provides one example of additional labels (Cui et al., 2024).

Target defines purpose and application. Cognitive intents include information retrieval, logical reasoning, creative generation, and tasks that involve emotion. DeepSeek-Prover-V1 supports logical reasoning (Xin et al., 2024), No Robots supports creative generation (Rajani et al., 2023), and empathetic-dialogues supports emotional tasks (Rashkin et al., 2019). Downstream tasks focus on data, models, or prompts. Data tasks include synthesis and augmentation. Model tasks include SFT, benchmarking, and RLHF. Instruction tuning collections such as Alpaca (Taori et al., 2023) and OASST1 (Köpf et al., 2023) form a major subset (Zhang et al., 2026).

Annotation protocol and reliability. We assign catalogue labels at the dataset level. Two authors independently annotated all 129 datasets for publisher type, generation process, display form, do-

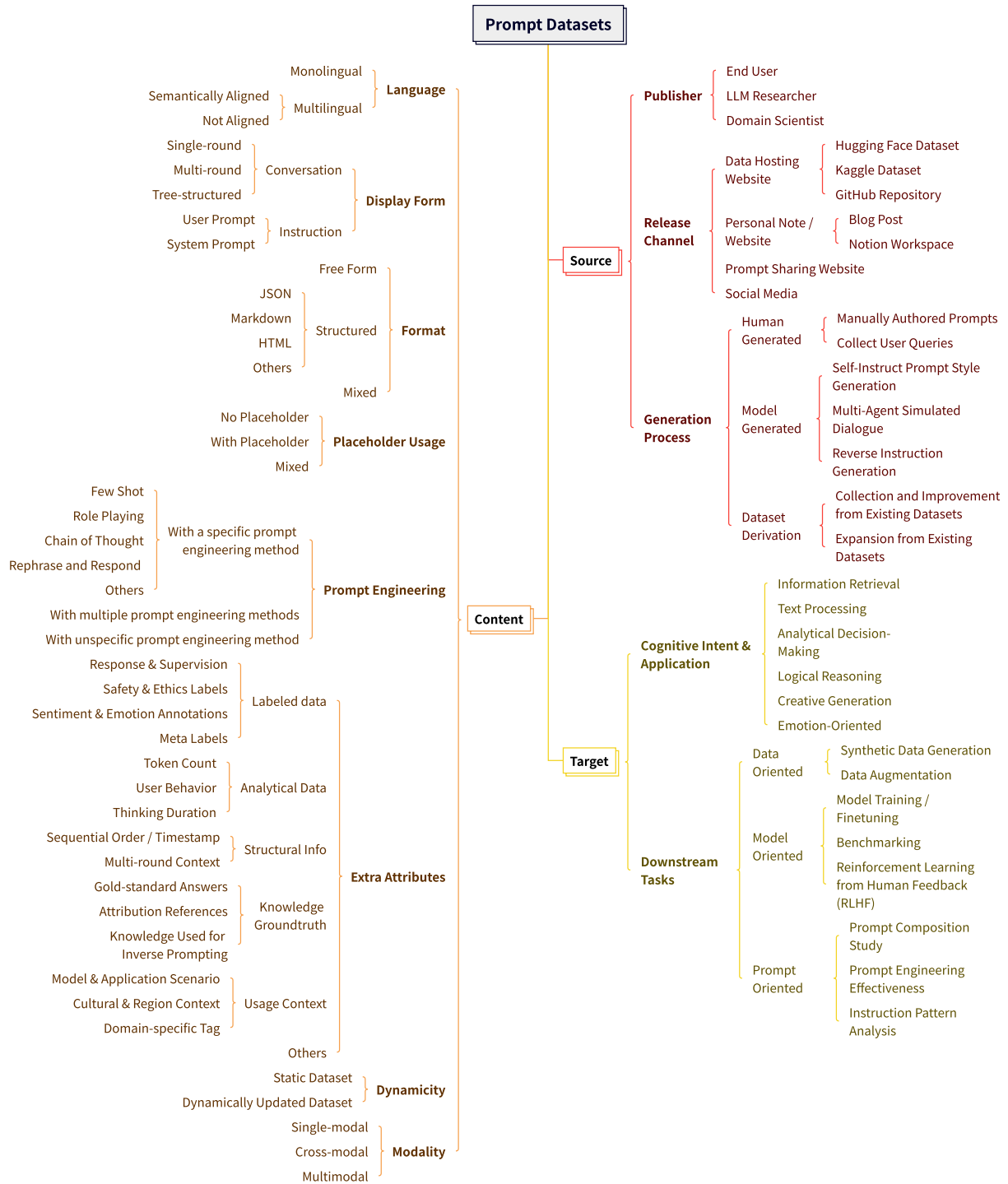


Figure 1: The hierarchical taxonomy of prompt datasets.

main, cognitive intent, and downstream task. They used dataset cards, official documentation, publisher descriptions of collection methods, and sampled prompts. Each scored facet receives one dominant label. Other documented facets permit *Mixed*, *multiple*, and *Unknown* values when needed. Cohen’s κ for the six facets is 0.829, 0.813, 0.862, 0.857, 0.606, and 0.838 in the listed order. Cognitive intent is the lowest and lies just above the

conventional moderate–substantial boundary. Appendix G.1 gives decision rules, agreement details, and catalogue-wide distributions. The agreement study does not cover the remaining design facets.

5 Prompt Data Analysis

We next analyse prompts from seven sources at the lexical, syntactic, and semantic levels. These sources form a sample for detailed analysis. The

sample does not provide a statistical representation of all 129 datasets. We use statistical and machine learning methods to identify composition patterns and variation across sources.

5.1 Dataset Selection

We use three principles to select prompt datasets for analysis across sources. Each dataset uses English and provides accessible raw prompts with wording that supports linguistic analysis. We exclude benchmark prompts such as PHYBench (Qiu et al., 2025). The selected datasets vary in publisher type, generation process, display form, domain, and target use. Table 1 summarises the seven datasets. Appendix G.4 describes each dataset.

Evidence scope and representativeness check.

Claims about scale, coverage, and taxonomy use all 129 datasets. The detailed linguistic results use the seven selected datasets. We test one main group pattern on 11 more prompt sources—nine catalogue datasets and two extension sources—and three general text baselines with the same feature pipeline. We sample up to 400 texts from each source with seed 42 and require at least 10 characters. Mean unknown vocabulary rates measured by the X tag are 0.008 for the seven datasets, 0.020 for the additional prompt datasets, and 0.003 for general text. The distributions for individual datasets overlap. This check therefore supports a group mean pattern and does not show that the seven datasets represent the full catalogue. Appendix G.3 gives the protocol and source lists.

5.2 Token Level Analysis

We use n gram models to analyse local text patterns (Jurafsky and Martin, 2000; Cavnar and Trenkle, 1994; Manning and Schütze, 1999). After lemmatising all tokens, we extract sequences of three, four, and five words. Their frequency distributions show common instruction templates, keyword combinations, and syntactic patterns. Appendix H.1 gives the full results.

Analysis of results. Frequent word sequences show differences in domain and prompt engineering. OASST1 contains role cues such as “you to act as,” and medical-o1 contains reasoning phrases such as “the most likely diagnosis.” Sequences of three words capture general commands. Sequences of four and five words capture patterns for specific tasks, such as ShareGPT’s “please write in English language.” Google Books contains sequences of five words such as “at the end of the” and “in

whole or in part.” Prompt corpora contain more questions and commands, which separates them from narrative and descriptive text.

5.3 Syntactic Analysis

We examine prompt structure through dependency parsing (Nivre, 2003), part of speech (POS) tagging (Brill, 1992), and term frequency inverse document frequency (TF-IDF) scoring (Salton and Buckley, 1988). These features describe the datasets and form vector representations for tasks such as prompt classification.

We compare the prompts with the English Universal Dependencies corpora EWT (Silveira et al., 2014) and ParTUT (Sanguinetti and Bosco, 2014). EWT contains blogs, social media, reviews, email, and web text. ParTUT contains formal text from legal documents, news, and Wikipedia.

5.3.1 Dependency Parsing

We use the spaCy `en_core_web_sm` parser (Honnibal and Montani, 2017) to extract syntactic dependencies and measure key grammatical relations in each dataset. We use the official dependency annotations for EWT and ParTUT. The results show consistent differences in linguistic style across prompt sources. We also track verb and object pairs to measure variation across prompt tasks. Figure 2 shows these pairs.

Analysis of Results. Table 2 shows eight common dependency types across seven prompt datasets and the EWT and ParTUT reference corpora. Three patterns stand out.

First, **medical-o1** has a high rate of adjectival modifiers at 0.11 and a low rate of direct objects at 0.03. This pattern reflects descriptions of medical conditions and diagnoses that use linking verbs. Second, **1.1k-business** often uses imperatives that state goals. Its direct object rate is 0.09, and its determiner rate is 0.05. This pattern fits its focus on project planning. Third, verb and noun pairs separate domains. Medical prompts use pairs such as “have history” and “experience pain,” while ShareGPT uses general pairs such as “write answer” and “use code.”

5.3.2 Part of Speech Tagging

We annotate each dataset with POS tags and measure the distribution of nouns, verbs, adjectives, and adverbs. Table 3 compares content words and function words. **1.1k-business** and **medical-o1** have noun proportions near 0.26, which exceed the

Table 1: Key characteristics of seven datasets. Size gives the number of prompts retained after standard preprocessing and quality checks. See Appendix G.4 for the mapping from native source units to retained prompt counts.

Dataset	Size	Publisher Type	Generation Method	Display Form	Domain
1.1k-business	1235	End User	Unknown	User Prompt	Business
BoredHumans	956	End User	Dataset Derivation	User Prompt	General
dolly-15k	14779	LLM Researcher	Human Generated	Single turn Conversation	General
medical-o1	19679	Domain Scientist	Model Generated	Single turn Conversation	Medical
OASST1	22079	LLM Researcher	Human Generated	Branching Conversation	General
Self-Instruct	81673	LLM Researcher	Model Generated	Single turn Conversation	General
ShareGPT	181577	End User	Human Generated	Multiple turn Conversation	General

Table 2: The eight most common dependency types and their dataset proportions. **punct** marks punctuation, **prep** marks prepositions, **det** marks determiners such as “the” and “a,” **pobj** marks prepositional objects, **dobj** marks direct objects, **nsubj** marks nominal subjects, and **ROOT** marks the main verb or predicate. The spaCy model `en_core_web_sm` uses some labels outside the Universal Dependencies standard. A dash marks these labels in comparisons across corpora. Table 11 gives the full data.

Dependency Type	EWI	ParTUT	1.1k-business	BoredHumans	dolly-15k	medical-o1	OASST1	Self-Instruct	ShareGPT
punct	<u>0.12</u>	<u>0.12</u>	0.1227	0.1985	0.1445	0.1216	0.1273	0.1863	0.1540
prep	-	-	0.0759	<u>0.0672</u>	0.0866	0.1013	0.0816	0.0676	0.0764
det	0.08	0.09	<u>0.0518</u>	0.0692	0.0961	0.0906	0.0841	0.0838	0.0693
pobj	-	-	0.0718	<u>0.0620</u>	0.0817	0.0979	0.0760	0.0645	0.0711
nsubj	0.08	0.06	0.0596	0.0545	0.0650	<u>0.0469</u>	0.0739	0.0596	0.0562
ROOT	0.07	0.04	0.0528	0.0462	0.0768	0.0444	0.0604	0.0792	<u>0.0437</u>
amod	0.05	0.06	0.0573	0.0527	0.0469	0.1072	0.0523	<u>0.0384</u>	0.0480
dobj	-	-	0.0904	0.0665	0.0447	<u>0.0315</u>	0.0594	0.0570	0.0519

proportion in the formal ParTUT corpus. These datasets therefore focus heavily on domain entities. `medical-o1` also has a high adjective ratio of 0.11, which fits the descriptive nature of clinical reasoning tasks.

5.3.3 TF-IDF Analysis

We analyse lexical patterns across prompt datasets with TF-IDF. We combine the prompts from each dataset into one document, which produces seven documents. The vectorizer uses at most 5,000 words and removes English stopwords. We measure lexical similarity across datasets with pairwise cosine similarity in Figure 2c. We also extract the three tokens with the highest weights in each dataset. Appendix Table 13 gives these tokens.

Analysis of results. Each corpus has a distinct lexical focus. `1.1k-business` emphasises business terms such as “content” and “email.” `BoredHumans` emphasises role prompts through words such as “act.” `Self-Instruct` emphasises structural tokens such as “output” with a weight of 0.766, which reflects its instruction and response format. Across datasets, `OASST1` and `ShareGPT` have the highest cosine similarity because both contain conversations between people and LLMs. `Self-Instruct` is lexically distant from `1.1k-business` and `medical-`

`o1`, which reflects differences in domain and style.

5.4 Semantic Analysis

We encode each prompt as a dense vector with 384 dimensions by using Sentence-BERT’s `all-MiniLM-L6-v2` model (Reimers and Gurevych, 2019). We sample 500 prompts from each dataset and project them to two dimensions with PCA. Appendix Figure 8 shows a scatter plot for each dataset. `Self-Instruct` has a broad and even semantic distribution, which fits its use of many instruction types. `medical-o1` and `1.1k-business` form concentrated clusters for their domains. `dolly-15k`, `OASST1`, and `ShareGPT` overlap because they share styles from interactions between people and LLMs.

K-Means with $K=7$ on the 384 dimensional embeddings yields $NMI=0.39$, $ARI=0.31$, $Purity=0.56$, and $silhouette=0.03$. These values show real boundaries with substantial overlap. The mean cosine distance to the centroid is 0.50 for `1.1k-business` and 0.55 for `medical-o1`, which confirms their concentrated clusters. `OASST1` and `ShareGPT` have distances of 0.79, and `dolly-15k` has a distance of 0.75. These three datasets have the greatest dispersion.

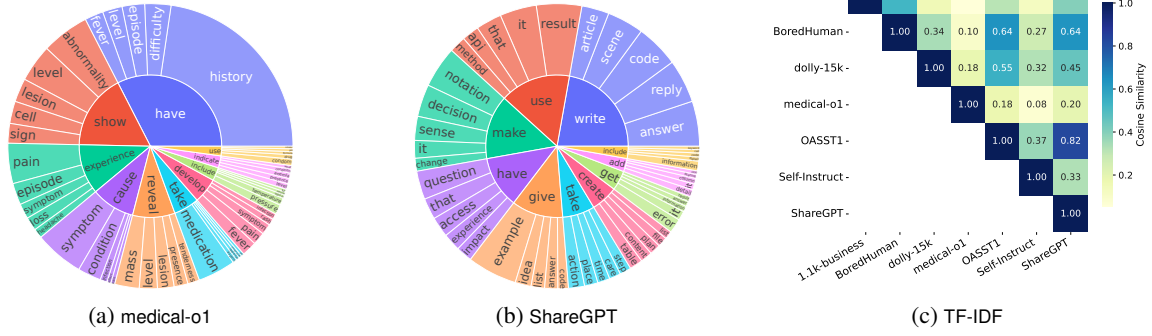


Figure 2: Panels (a) and (b) show the ten most common verbs and their five most common direct noun objects in two prompt datasets. Figure 7 gives the other five datasets. Panel (c) shows cosine similarity between TF-IDF vectors for each dataset.

Table 3: The seven most common parts of speech and their dataset proportions. Table 12 gives the full data.

POS	EWT	ParTUT	1.1k-business	BoredHumans	dolly-15k	medical-o1	OASST1	Self-Instruct	ShareGPT
NOUN	0.17	0.21	0.2637	0.2103	0.1899	0.2590	0.1946	0.2027	0.1944
PUNCT	0.12	0.12	<u>0.1094</u>	0.1942	0.1435	0.1158	0.1231	0.1839	0.1450
VERB	0.11	0.10	0.1302	0.1094	0.0871	<u>0.0775</u>	0.1069	0.0999	0.0979
ADP	0.09	0.12	0.0758	<u>0.0678</u>	0.0858	0.0998	0.0851	0.0701	0.0789
DET	0.08	0.11	<u>0.0506</u>	0.0693	0.0949	0.0893	0.0839	0.0844	0.0696
PRON	0.09	0.04	0.0912	0.0708	0.0695	<u>0.0369</u>	0.0870	0.0701	0.0583
ADJ	0.07	0.08	0.0588	0.0543	0.0538	0.1104	0.0632	<u>0.0498</u>	0.0563

6 Application

We use the extracted linguistic features in three applications. Prompt filtering separates LLM prompts from general text. Source domain routing assigns prompts to five source domains. Our feature-based elicited response quality analysis estimates the judged quality of responses elicited by a prompt without LLM inference at prediction time. These applications show that the patterns in §5 support classification and association tasks. To connect them to the taxonomy, Appendix G.2 groups source-level linguistic measurements by cognitive intent and downstream task. Appendix H.5 reports zero- and few-shot LLM baselines and inference cost for all three applications.

6.1 Prompt Filtering

Task. Given a text snippet, the classifier determines whether it is an LLM prompt or general text. General text includes blog posts, Reddit comments, and Wikipedia passages. This task tests whether the structural features in our analysis capture the characteristics of prompts.

Setup. We sample 3,000 prompts from seven analysed datasets and 3,000 general texts. The general texts contain 1,000 examples from each of the blog, Reddit, and Wikipedia corpora. We extract

six representations for each text. They are an 18 dimensional POS distribution, a 45 dimensional dependency distribution, their 63 dimensional Syntactic concatenation, a 5,000 dimensional TF-IDF vector, a 384 dimensional sentence embedding, and a Combined vector with 5,447 dimensions. We fit TF-IDF in each training fold. We train MLP classifiers with five fold stratified cross validation. We also evaluate an Isolation Forest fitted on prompts.

Results. Table 4 reports the results. The Combined representation reaches F1=0.902, and sentence embeddings reach 0.899. The 63 dimensional Syntactic representation reaches F1=0.822, which is over 91% of embedding F1. It uses $\approx 6x$ fewer dimensions and requires no embedding inference. In our runtime benchmark, CPU linguistic features take 3.0 ms per request, and GPU embeddings take 5.7 ms. Section 6.5 discusses this comparison. The Isolation Forest (Liu et al., 2008) performs near chance, which shows the value of supervised learning. Appendix Figure 9 gives the training curves and F1 for each feature set. Supervised F1 ranges from 0.789 to 0.902 across the six representations.

6.2 Source Domain Routing

Task. We use the same feature representations to route prompts into Business, Medical, Coding, Finance, and Creative Writing. This task

Table 4: Classification of prompts and general text across six feature representations. The table reports the mean and standard deviation from five fold stratified cross validation for the supervised MLP and results for an Isolation Forest baseline. **Bold** marks the best supervised value for each metric.

Feature	Supervised (MLP)				Unsupervised (Isolation Forest)			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
POS (18 dim)	.789±.007	.787±.007	.792±.021	.789±.010	.507	.505	.700	.587
Dependency (45 dim)	.793±.017	.790±.012	.798±.029	.794±.019	.532	.524	.700	.599
Syntactic (63 dim)	.822±.012	.822±.012	.822±.013	.822±.012	.515	.511	.700	.591
TF-IDF (5k dim)	.848±.009	.843±.019	.855±.029	.849±.010	.561	.547	.700	.614
Embedding (384 dim)	.900±.005	.907±.002	.892±.014	.899±.006	.599	.582	.700	.636
Combined (5.4k dim)	.903±.008	.912±.008	.893±.018	.902±.009	.569	.554	.700	.619

tests whether linguistic features separate source domains.

Datasets. We use five training sources. 1.1k-business covers Business, and medical-o1-reasoning-SFT covers Medical. OpenCodeReasoning is an NVIDIA corpus distilled for competitive programming and covers Coding (Ahmad et al., 2025b). finance-alpaca contains financial instructions and covers Finance. WritingPrompts contains Reddit story prompts and covers Creative Writing (Fan et al., 2018). Each domain contributes 1,200 prompts, for a total of 6,000.

Evaluation protocol. We conduct four evaluation phases. Phase A uses five fold stratified cross validation with MLP classifiers. Phase B uses unsupervised K-Means clustering to test whether domain structure appears without labels and supervised Nearest Centroid classification as a simple geometric baseline. We report NMI, ARI, Purity, and accuracy. Phase C uses an 80/20 stratified split. Phase D trains on the five original sources and evaluates transfer to one target collection for each domain. The target collections are the business subset of 2.5k-templates, UltraMedical (Zhang et al., 2024), OpenCodeInstruct (Ahmad et al., 2025a), FiQA (Maia et al., 2018), and the generation subset of No Robots (Rajani et al., 2023).

Phase A results. Figure 3a presents five fold cross validation results. Appendix Figure 10 gives the diagnostics. The Combined representation reaches 97.3% Macro-F1, and sentence embeddings reach 97.1%. The Syntactic representation reaches 90.2%, and TF-IDF 93.3%. Coding is recognised most consistently. Finance is generally the hardest domain because its lexical and structural features overlap with general business text.

Phase B and C results. Figures 3b and 3c report Phase B and Phase C metrics for three representative feature sets. Appendix Tables 14 and 15 and Figure 13 give results for all six feature sets.

Table 5: Phase D source domain transfer. Pooled cross validation uses both collections. Cross dataset evaluation trains on the original collection and tests on the target collection. Chance Macro-F1 is 0.20.

Representation	Pooled CV	Cross dataset
Formatting syntactic (8 dim)	.677	.329
Full syntactic (55 dim)	.913	.525
Structural syntactic (47 dim)	.893	.605
Embedding (384 dim)	.964	.740

Sentence embeddings yield K-Means NMI = 0.80 and Purity = 0.92. The Combined representation yields a lower K-Means NMI of 0.69 because of its high dimensionality. On the 20% test split, sentence embeddings reach 97.4% Macro-F1. Most remaining errors occur in Finance, as shown in Appendix Figure 11. Maximum softmax confidence separates 1,200 test prompts from the five domains and 1,000 general prompts with ROC-AUC = 0.843. Appendix Figure 12 gives this result.

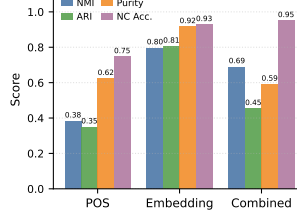
Phase D results. All representations exceed chance on the target collections. The decrease from pooled cross validation shows that the representations capture signals specific to each collection. Removing eight formatting features improves syntactic transfer from 0.525 to 0.605. Sentence embeddings transfer best and reach 0.740. Creative recall ranges from 0.10 to 0.26 because the two creative collections contain different writing tasks. Syntactic features provide an efficient option within known distributions. Sentence embeddings provide stronger transfer across collections. Appendix H.4.2 gives the results for each domain.

6.3 Feature Importance via Gradient Analysis

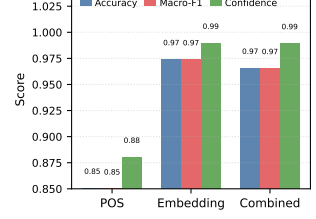
We apply mean absolute gradient saliency (Simonyan et al., 2014; Baehrens et al., 2010) to the trained POS, Dependency, and Syntactic MLPs to identify the features that drive classification. For



(a) Phase A reports five fold Macro-F1 and F1 for each domain across six feature representations. The first column gives Macro-F1 in bold.



(b) Phase B reports K-Means NMI, ARI, and Purity together with Nearest Centroid accuracy.



(c) Phase C reports accuracy, Macro-F1, and average maximum softmax confidence on the 20% test split.

Figure 3: Source domain routing under three protocols within each source. Panel (a) shows that Coding is recognised most consistently and Finance is generally the hardest domain. Panels (b)-(c) show that sentence embeddings lead the clustering metrics and reach 97.4% Macro-F1 on the test split. The Appendix gives results for all six feature sets.

each input feature x_i and output class c , the importance score is the expected absolute partial derivative of the class logit with respect to that feature. We average the score over all classes and the training distribution.

$$\text{Imp}(i) = \frac{1}{|C|} \sum_{c \in C} \mathbb{E}_x \left[\left| \frac{\partial \text{logit}_c(x)}{\partial x_i} \right| \right]. \quad (1)$$

Larger values of $\text{Imp}(i)$ mean that small changes to feature i produce larger changes in the model logits. Appendix Table 17 gives the five most important features.

Appendix Figures 14 and 15 show gradient importance for both tasks.

Structural signals dominate both tasks.

SPACE represents whitespace tokens, and `dep` represents unresolved dependency arcs. SPACE leads the POS-only prompt-filtering representation, while `dep` leads the separate Dependency-only representation. In the 63 dimensional Syntactic prompt-filtering representation, the five largest values are `nsubj` (.648), SPACE (.640), `attr` (.622), `PRON` (.594), and `ADP` (.561); `dep` is not among its five leading features. For source domain routing, SPACE and `dep` remain leading signals in the Syntactic and Dependency-only representations, respectively. SPACE captures numbered steps, code blocks, and bullet lists that are less common in linear prose. `dep` captures technical noun phrases and imperative fragments. Prompt filtering also uses an imperative and conversational register that is less common in general text.

Domain lexical features and practical rules.

Source domain routing relies heavily on X, which represents unknown vocabulary tokens such as code identifiers, medical abbreviations, and financial tickers. Its importance is 0.87 in the POS-only representation. The NUM and `nummod` features

identify quantitative content in Finance and Medical prompts. The compound feature captures domain terms such as *blood pressure* and *cash flow*. The ROOT feature captures action verbs such as *write* and *implement* in Coding, *diagnose* and *explain* in Medical, and *calculate* in Finance. These results suggest three rules that do not require a trained model. High rates of `PRON`, `ADV`, `nsubj`, and `attr` help identify prompt text. The X tag ratio helps identify a domain without labels. The combined density of NUM and `nummod` separates quantitative and qualitative domains.

6.4 Elicited Response Quality

The prior experiments show that linguistic features identify the type and source domain of a prompt. We next test whether the same features correlate with the judged quality of responses elicited by a prompt. This target depends on the response models, task difficulty, and judge.

Setup. We use UltraFeedback (Cui et al., 2024). Four LLMs answer each prompt, and GPT-4 rates every response from 1 to 10 with the `overall_score` field. We average the four response scores for each prompt. We then draw a stratified sample of 10,000 unique prompts without adding annotations. The final scores range from 1.25 to 9.38 and have a mean of 6.41. We extract 55 linguistic features that contain 17 POS rates, 27 dependency rates, and 11 basic statistics. We report Spearman ρ with Benjamini and Hochberg FDR at $\alpha=0.05$ (Benjamini and Hochberg, 1995). We also report Cohen’s d for Q4 compared with Q1 (Cohen, 1988) and logistic regression AUC for Q1 compared with Q4 under five fold cross validation. We test judge dependence on 2,000 HelpSteer2 prompts with human helpfulness ratings (Wang et al., 2024c). For both datasets, we

Table 6: Selected correlations with elicited response quality before length control on 10,000 UltraFeedback prompts. The columns report Spearman ρ and Cohen’s d for Q4 compared with Q1.

Feature	ρ	d
<i>Positive correlations</i>		
Lexical diversity	+.257	+.721
ADP (preposition)	+.168	+.397
<i>Negative correlations</i>		
dep (syntactic ambiguity)	−.320	−.678
appos (appositive)	−.304	−.598
SPACE (formatting tokens)	−.308	−.607
Sentence count	−.294	−.445
Word count	−.270	−.379
X (OOV/technical tokens)	−.248	−.388
NUM (numerals)	−.233	−.384
PROPN (proper nouns)	−.221	−.448

compute partial Spearman correlations that control for prompt word count.

Results before length control. Table 6 reports selected associations. After Benjamini and Hochberg FDR at $\alpha=0.05$, 49 of 55 features remain significant. Within UltraFeedback, prediction of Q1 compared with Q4 reaches AUC 0.799 with all features and 0.773 with Dependency alone. HelpSteer2 human ratings show the same directions before length control. Lexical diversity has $\rho=+.150$. The correlations for dep, SPACE, NUM, and X are $-.128$, $-.127$, $-.172$, and $-.077$.

Prompt length confounds lexical diversity.

The reported diversity feature is type to token ratio, which depends on length (McCarthy and Jarvis, 2010). After word count is controlled, its association changes from $+.257$ to $+.024$ in UltraFeedback and from $+.150$ to $-.017$ in HelpSteer2. In UltraFeedback, the routing features dep, SPACE, X, and NUM remain negative. Their partial correlations are $-.212$, $-.197$, $-.153$, and $-.136$. In HelpSteer2, only NUM remains negative after length control. We interpret these findings as observational associations that depend on the dataset and judge. They provide no evidence that lexical diversity causes better elicited responses. Appendix H.4.4 gives the full robustness results.

6.5 Discussion

The 63 dimensional linguistic features reach more than 91% of GPU embedding F1. In our benchmark, their latency per request is 3.0 ms compared

with 5.7 ms for embeddings. They also require no corpus vocabulary and support gradient importance analysis. The separate 55 dimensional linguistic representation used in Phase D transfers less reliably than sentence embeddings across target collections. Gradient importance identifies representation-specific structural signals: `nsubj` and `SPACE` lead the Syntactic prompt-filtering model, while `dep` leads the Dependency-only model and is important for routing. Several routing features remain negatively associated with UltraFeedback scores after length control. The positive association for type to token ratio becomes negligible after the same control. These results support a two stage pipeline that uses efficient structural features for known traffic and semantic features or more processing when the distribution changes.

Comparison with LLM inference. On matched but non-paired samples, zero- and few-shot `gpt-4o-mini` is slightly stronger for filtering (F1 .919–.943 vs. .902) and cross-dataset routing (Macro-F1 .921–.924 vs. .740), but our linguistic classifier is stronger for response-quality quartile prediction (AUC .799 vs. .457–.479; LLM $\rho \approx 0$). LLM inference also takes 1.54–1.60 s per prompt, 270–510 times the 3.0–5.7 ms local pipelines, and incurs per-query API cost. Thus our approach offers competitive filtering, better within-dataset quality discrimination, interpretability, and substantially lower latency; LLM inference is most useful for semantic routing (Appendix H.5, Table 19).

7 Conclusion

We compiled 129 LLM prompt datasets into a structured taxonomy and measured annotation reliability at the dataset level. We analysed seven diverse datasets in detail. The 63 dimensional linguistic features provide efficient and interpretable signals within known distributions. Sentence embeddings transfer more reliably across target collections. In UltraFeedback, several structural routing features remain negatively associated with elicited response quality after length control. The association between type to token ratio and quality becomes much smaller after the same control. These findings support pipelines that combine efficient structural processing, semantic routing, and checks for changes in the data distribution.

Limitations

Our corpus contains mainly English prompts from Western platforms. The reported linguistic patterns may therefore reflect technically skilled users who write in English. Cross dataset routing uses two collections for each domain and transfers weakly between the two Creative tasks. The elicited response quality study is observational. UltraFeedback inherits biases from its GPT-4 judge (Liu et al., 2024), HelpSteer2 uses one protocol for human ratings, and prompt length confounds type to token ratio. Taxonomy labels apply at the dataset level. Dominant labels and descriptions of mixed attributes preserve major forms of variation but omit some differences within a dataset. Appendix A discusses classifier scope and licensing constraints.

Acknowledgments

Arijit Khan acknowledges support from the Novo Nordisk Foundation Grant NNF22OC0072415.

References

- Wasi Uddin Ahmad, Aleksander Ficek, Mehrzad Samadi, Jocelyn Huang, Vahid Noroozi, Somshubra Majumdar, and Boris Ginsburg. 2025a. [Opencode-instruct: A large-scale instruction tuning dataset for code llms](#). *arXiv preprint arXiv:2504.04030*.
- Wasi Uddin Ahmad, Sean Narenthiran, Somshubra Majumdar, Aleksander Ficek, Siddhartha Jain, Jocelyn Huang, Vahid Noroozi, and Boris Ginsburg. 2025b. [Opencodereasoning: Advancing data distillation for competitive coding](#). *arXiv preprint arXiv:2504.01943*.
- Fatih Kadir Akın. 2022. [awesome-chatgpt-prompts](#). <https://github.com/f/awesome-chatgpt-prompts>.
- David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. 2010. [How to explain individual classification decisions](#). *J. Mach. Learn. Res.*, 11(61):1803–1831.
- Yoav Benjamini and Yosef Hochberg. 1995. [Controlling the false discovery rate: A practical and powerful approach to multiple testing](#). *J. R. Stat. Soc. B*, 57(1):289–300.
- BoredHumans. 2024. [Boredhumans](#). <https://boredhumans.com/prompts.php>.
- Eric Brill. 1992. [A simple rule-based part of speech tagger](#). In *ANLP*, pages 152–155.
- Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- William B Cavnar and John M Trenkle. 1994. N-gram-based text categorization. In *SDAIR*, pages 161–175.
- Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024a. [Huatuogpt-o1, towards medical complex reasoning with llms](#). *arXiv preprint arXiv:2412.18925*.
- Lingjiao Chen, Matei Zaharia, and James Zou. 2024b. [Frugalgpt: How to use large language models while reducing cost and improving performance](#). *Transactions on Machine Learning Research*.
- Zhenpeng Chen, Chong Wang, Weisong Sun, Xuanzhe Liu, Jie M. Zhang, and Yang Liu. 2026. [Promptware engineering: Software engineering for prompt-enabled systems](#). *ACM Transactions on Software Engineering and Methodology*, 35(9):1–22.
- Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. Lawrence Erlbaum Associates.
- ChatGPT Prompt Genius community. 2023. [Chatgpt prompt genius](#). <https://www.reddit.com/r/ChatGPTPromptGenius/>.
- 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](#). <https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm>.
- Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. [ULTRAFEEDBACK: Boosting language models with scaled AI feedback](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 9722–9744. PMLR.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. [Hierarchical neural story generation](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898. Association for Computational Linguistics.
- 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*.
- Prompt Genius. 2024. [Prompt genius](#). <https://www.promptgenius.site/>.

- Feijuan He, Han Lai, Jiaqi Liu, Bo Wang, Haoran Chen, Haohan Liu, and Chenxi Zhang. 2025. [Solving mathematical problems using large language models: A survey](#). *Data Intelligence*, 7(4):907–946.
- Matthew Honnibal and Ines Montani. 2017. *spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing*. <https://spacy.io>.
- Daniel Jurafsky and James H Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR.
- Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2023. [Openassistant conversations – democratizing large language model alignment](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 47669–47681. Curran Associates, Inc.
- Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, and 13 others. 2021. [Datasets: A community library for natural language processing](#). In *EMNLP*, pages 175–184.
- Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. [Camel: Communicative agents for “mind” exploration of large language model society](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 51991–52008. Curran Associates, Inc.
- Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. [Isolation forest](#). In *ICDM*, pages 413–422.
- Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. [Lost in the middle: How language models use long contexts](#). *Trans. Assoc. Comput. Linguist.*, 12:157–173.
- Yang Liu, Jiahuan Cao, Chongyu Liu, Kai Ding, and Lianwen Jin. 2025a. [Datasets for large language models: A comprehensive survey](#). *Artificial Intelligence Review*, 58:403.
- Yuanye Liu, Jiahang Xu, Li Lyna Zhang, Qi Chen, Xuan Feng, Yang Chen, Zhongxin Guo, Yuqing Yang, and Peng Cheng. 2025b. [Beyond prompt content: Enhancing llm performance via content-format integrated prompt optimization](#). *arXiv preprint arXiv:2502.04295*.
- Yitian Luo, Yu Liu, Lu Zhang, Feng Gao, and Jinguang Gu. 2025. [A survey on quality evaluation of instruction fine-tuning datasets for large language models](#). *Data Intelligence*, 7(3):527–566.
- Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. [WWW’18 open challenge: Financial opinion mining and question answering](#). In *Companion Proceedings of the Web Conference 2018*, pages 1941–1942. Association for Computing Machinery.
- Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.
- Philip M McCarthy and Scott Jarvis. 2010. [MtlD, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment](#). *Behav. Res. Methods*, 42(2):381–392.
- Aditi Mishra, Bretho Danzy, Utkarsh Soni, Anjana Arunkumar, Jinbin Huang, Bum Chul Kwon, and Chris Bryan. 2025. [Promptaid: Visual prompt exploration, perturbation, testing and iteration for large language models](#). *IEEE Trans. Vis. Comput. Graph.*, 31(10):6946–6962.
- JianZheng Nie. 2023. [awesome-instruction-datasets](https://github.com/jianzhnie/awesome-instruction-datasets). <https://github.com/jianzhnie/awesome-instruction-datasets>.
- Joakim Nivre. 2003. [An efficient algorithm for projective dependency parsing](#). In *IWPT*, pages 149–160.
- Natasha Noy, Matthew Burgess, and Dan Brickley. 2019. [Google dataset search: Building a search engine for datasets in an open web ecosystem](#). In *WWW*, pages 1365–1375.
- Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E Gonzalez, M Waleed Kadous, and Ion Stoica. 2025. [Routellm: Learning to route llms with preference data](#). In *International Conference on Learning Representations*.
- PromptBase. 2024. [Promptbase](https://promptbase.com/). <https://promptbase.com/>.
- Shi Qiu, Shaoyang Guo, Zhuo-Yang Song, Yunbo Sun, Zeyu Cai, Jiashen Wei, Tianyu Luo, Yixuan Yin, Zhang Haoxu, Yi Hu, Chenyang Wang, Chencheng Tang, Haoling Chang, Qi Liu, Ziheng Zhou, Tianyu Zhang, Jingtian Zhang, Zhangyi Liu, Minghao Li, and 34 others. 2025. [Phybench: Holistic evaluation of physical perception and reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc.
- Nazneen Rajani, Lewis Tunstall, Edward Beeching, Nathan Lambert, Alexander M Rush, and Thomas Wolf. 2023. [No robots](https://huggingface.co/datasets/HuggingFaceH4/no_robots). https://huggingface.co/datasets/HuggingFaceH4/no_robots.

- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. [Towards empathetic open-domain conversation models: A new benchmark and dataset](#). In *ACL*, pages 5370–5381.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-bert: Sentence embeddings using siamese bert-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992. Association for Computational Linguistics.
- David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, and 57 others. 2024. [Cvqa: Culturally-diverse multilingual visual question answering benchmark](#). In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 37, pages 11479–11505.
- Gerard Salton and Christopher Buckley. 1988. [Term-weighting approaches in automatic text retrieval](#). *Inf. Process. Manag.*, 24(5):513–523.
- Manuela Sanguinetti and Cristina Bosco. 2014. [Converting the parallel treebank partut in universal stanford dependencies](#). In *CLiC-it*.
- Elvis Saravia. 2022. Prompt engineering guide. <https://github.com/dair-ai/Prompt-Engineering-Guide>.
- Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yin-heng Li, Aayush Gupta, Hyojung Han, Sevien Schulhoff, Pranav Sandeep Dulepet, Saurav Vidyadhara, Dayeon Ki, Sweta Agrawal, Chau Pham, Gerson Kroiz, Feileen Li, Hudson Tao, Ashay Srivastava, and 12 others. 2024. [The prompt report: A systematic survey of prompt engineering techniques](#). *arXiv preprint arXiv:2406.06608*.
- Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. [Autoprompt: Eliciting knowledge from language models with automatically generated prompts](#). In *EMNLP*, pages 4222–4235.
- Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D Manning. 2014. [A gold standard dependency corpus for english](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 2897–2904. European Language Resources Association (ELRA).
- Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. [Deep inside convolutional networks: Visualising image classification models and saliency maps](#). In *ICLR Workshop*.
- SirNeural. 2022. Flan 2022. https://huggingface.co/datasets/SirNeural/flan_v2.
- Travis Tang. 2023. Chatgpt-data-science-prompts. <https://github.com/travistangvh/ChatGPT-Data-Science-Prompts>.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.
- Chaoyi Wang and Yuan Sun. 2025. [MSEED: Human preference dataset for multidimensional safety enhancement and evaluation of large language models](#). *Data Intelligence*, 7(4):1243–1269.
- Junpeng Wang, Chin-Chia Michael Yeh, Yujie Fan, Xin Dai, Yan Zheng, Liang Wang, and Wei Zhang. 2024a. [Promptlandscape: Guiding prompts exploration and analysis with visualization](#). In *PacificVis*, pages 319–324.
- Xinyuan Wang, Chenxi Li, Zhen Wang, Fan Bai, Haotian Luo, Jiayou Zhang, Nebojsa Jojic, Eric P Xing, and Zhiting Hu. 2024b. [Promptagent: Strategic planning with language models enables expert-level prompt optimization](#). In *The Twelfth International Conference on Learning Representations*.
- 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 *ACL*, pages 13484–13508.
- Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. 2024c. [Helpsteer2: Open-source dataset for training top-performing reward models](#). In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.
- Huajian Xin, Daya Guo, Zhihong Shao, Zhizhou Ren, Qihao Zhu, Bo Liu, Chong Ruan, Wenda Li, and Xiaodan Liang. 2024. [Deepseek-prover: Advancing theorem proving in llms through large-scale synthetic data](#). In *NeurIPS Workshop on MATH-AI*.
- Daojian Zeng, Lin Zhou, Zhiheng Zhang, and Lincheng Jiang. 2025. [AuToGen: Automated tool learning data generation with domain-specific structured data](#). *Data Intelligence*, 7(4):1108–1128.

- Biao Zhang, Yunwei Chen, and Daobin Gao. 2025. [Development characteristics and topic analysis of international research on ChatGPT](#). *Data Intelligence*, 7(4):1192–1217.
- Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, Xingtai Lv, Jinfang Hu, Zhiyuan Liu, and Bowen Zhou. 2024. [Ultrammedical: Building specialized generalists in biomedicine](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 26045–26081. Curran Associates, Inc.
- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. [Personalizing dialogue agents: I have a dog, do you have pets too?](#) In *ACL*, pages 2204–2213.
- Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2026. [Instruction tuning for large language models: A survey](#). *ACM Computing Surveys*, 58(7):1–36.
- Junhao Zhao. 2023. [Llmdatabub. https://github.com/Zjh-819/LLMDataHub](https://github.com/Zjh-819/LLMDataHub).

A Limitations and Discussion

Dataset coverage and language bias. Our survey covers 129 sources with more than 1.22 TB and more than 673M instances. Most prompts are in English and come from Western platforms such as GitHub, Reddit, Hugging Face, and PromptBase. The catalogue has less coverage of prompts in other languages, private enterprise prompts, and traffic from commercial assistants such as ChatGPT and Claude. The reported linguistic patterns may therefore reflect technically skilled users who write in English.

Taxonomy scope and residual heterogeneity. Two independent annotators assign the taxonomy in §4 at the dataset level. They use source documents and sampled prompts. Five of the six scored facets have almost perfect agreement. Cognitive intent has the lowest agreement at $\kappa = .606$, just above the conventional moderate–substantial boundary. Broad or mixed collections still have uncertain category boundaries. For documented but unscored attributes, mixed and multiple descriptions preserve supported categories but omit some differences between prompts. Future work can extend the taxonomy to prompt-level annotation within the catalogue instances.

Elicited response quality depends on data and judges. Our primary analysis in §6.4 uses UltraFeedback with GPT-4 as the reference judge. GPT-4 has style and length biases (Liu et al., 2024). HelpSteer2 adds human helpfulness ratings from one response set and one rating protocol. Type to token ratio depends on length. Its positive association becomes negligible after prompt word count is controlled. The reported AUC and associations describe these data and judging settings. They do not define a universal or intrinsic measure of elicited response quality.

Association and causal interpretation. Gradient importance in §6.3, Spearman correlations, and Cohen’s d report associations between linguistic features and classifier decisions or elicited response quality. These results provide no evidence that removing a feature such as whitespace tokens would improve an answer. Length control addresses one measured confound. Prompt topic, task difficulty, user expertise, response model capability, and judge preferences may also affect the results. Causal interpretation requires controlled rewriting experiments.

Classifier scope. We evaluate prompt filtering as a binary task that separates prompts and general text. We evaluate source domain routing on balanced samples from Business, Medical, Coding, Finance, and Creative Writing. The transfer study uses two collections for each domain. The two Creative collections represent different writing tasks. Macro-F1 may decrease on the full taxonomy or highly imbalanced traffic. The linguistic pipeline may also become less competitive as the number of classes increases.

Licensing and redistribution constraints. Several sources have unclear or restrictive licenses. Our release contains catalogue metadata, taxonomy annotations, processed statistics, and code. It excludes raw prompts when the source terms prohibit redistribution. Researchers who reproduce token level statistics for these sources must download the data from the original hosts.

Prompt effect evaluation. The selected datasets cover diverse tasks, so this study does not evaluate causal prompt effects by generating and comparing new task responses. The LLM baselines in Appendix H.5 instead predict task labels or existing response scores from prompt text. Future work could run the same prompts through several LLMs and compare their task responses. Prompt data continues to grow, and repeated analyses could track new patterns that affect prompt design.

B LLM Usage

LLMs enter this work in the following roles.

LLMs as analysis objects. Some analysed datasets contain prompts generated by LLMs. Self-Instruct (Wang et al., 2023) expands human seeds with GPT-3. medical-o1-reasoning-SFT (Chen et al., 2024a) contains medical questions reformulated by GPT-4o. We treat these datasets as observed data and generate no new training prompts or task responses for the linguistic and elicited-response-quality analyses.

Existing labels in the quality analysis. The elicited response quality study in §6.4 uses scores released with UltraFeedback (Cui et al., 2024) and human helpfulness ratings released with Help-Steer2 (Wang et al., 2024c). We use both label sets as released and add no LLM annotations. The statistical analysis uses Spearman ρ , partial correlations, Benjamini and Hochberg FDR, Cohen’s d ,

and logistic regression AUC.

Existing encoders in the feature pipeline. The semantic analysis in §5.4 and the Embedding representation in §6.1 and §6.2 use the frozen Sentence-BERT all-MiniLM-L6-v2 encoder (Reimers and Gurevych, 2019). The encoder maps each prompt to a vector with 384 dimensions. We use it as a fixed feature extractor and perform no model tuning, prompting, or generation. We use spaCy in the same way to extract POS and dependency tags.

LLMs as evaluation baselines. The zero- and few-shot baselines in Appendix H.5 query `openai/gpt-4o-mini` to predict prompt-filtering labels, source-domain labels, or an existing elicited-response score from prompt text. These calls generate task-level labels or scores for evaluation; they do not generate answers to the source prompts or add examples to the analysed datasets.

During paper preparation. The authors used LLM assistants for English editing, code completion, and checks of the organisation and consistency of analysis files. The authors verified all experimental designs, numerical results, citations, interpretations, and final text. The authors remain responsible for the paper.

C Ethics Statement

We compile and analyse existing prompt datasets and collect no new sensitive data. We document the datasets, respect their sources and licenses, and release code and derived artefacts for review and reproduction. The work analyses existing data and releases no new generative technology. We document the methods to support external review.

Personally identifying information and offensive content. Several analysed datasets, including ShareGPT90K and OASST1, come from real user interactions. They may contain personally identifying information or offensive content even though users shared them on public platforms. We do not redistribute raw prompts from users. The released artefacts contain catalogue metadata, annotations, code, aggregate statistics, and derived results. This work does not provide a full content audit of the source corpora. Users who download the raw data should screen it for identifying information and harmful content based on their use case.

D Reproducibility Statement

The repository linked in the abstract contains the code, data processing instructions, and experiment commands. It also specifies dependency versions. The datasets are public or have documented sources. We set random seeds when the method permits. These resources support reproduction of the analyses and results.

E Compute Resources

All local experiments can run on one workstation and require no cluster or distributed training. The zero- and few-shot LLM baselines additionally require access to the external OpenRouter API; their end-to-end latency, token counts, and estimated cost are reported separately.

Compute workers. Linguistic feature extraction, TF-IDF and scikit-learn baselines, and statistical analyses run on a standard CPU with several cores. Sentence-BERT embedding and the PyTorch MLP experiments run on one consumer GPU. The encoder can also run in CPU batches at the reported scale of at most 10,000 prompts for each feature extraction job.

Memory use. The largest feature objects are the sentence embedding matrix with $n \times 384$ values and the TF-IDF representation with 5,000 dimensions. The sample size is at most 10^4 . Both objects fit in workstation memory at this scale.

Runtime. The application experiments can run on the workstation described above. Runtime depends on data loading, feature caching, and hardware. Embedding extraction and repeated MLP folds use most of the computation.

Execution details. The repository at <https://github.com/ymzhang-cs/prompt-dataset-analysis> provides commands, random seeds, and hyperparameters for each experiment.

F Broader Impacts

We survey and analyse public LLM prompt datasets and train compact classifiers with spaCy, TF-IDF, and Sentence-BERT features. We release no generative foundation model, image generator, or new scraped corpus. The following paragraphs describe potential benefits and risks.

Positive impacts. The Syntactic representation used in prompt filtering and Phases A through C of source domain routing has 63 dimensions and reaches over 91% of embedding F1 for prompt filtering. Its features can be extracted on a CPU and inspected through gradient importance. This pipeline may reduce GPU cost, latency, and energy use in systems that process requests as they arrive. It may also make prompt research more accessible to groups with limited compute. The released taxonomy, statistics for each dataset, and code provide a reference for future studies and audits of changes in prompt distributions.

Potential negative impacts. Formatting density, unresolved dependency arcs, and technical tokens correlate with writing style. Classifiers trained on these features may penalise prompts from people who learned English as another language, new users, or users of assistive technologies. Type to token ratio also depends on prompt length and should not be treated as an intrinsic quality measure. A deployment should include fairness audits across relevant user groups. The reported structural signals could also help attackers craft prompt injection or jailbreak inputs that match expected statistics. Practitioners should combine these features with other signals and avoid using them as a sole detector.

Mitigations. Downstream users should audit performance across relevant user groups before deployment. They should combine structural classifiers with safety models that inspect content. They should also monitor calibration as prompt distributions change.

G Datasets

We collected 129 prompt datasets from four source channels. They are dataset platforms, publications at NeurIPS, ICLR, and ICML from 2023 through 2025, GitHub repositories, and prompt sharing websites. Dataset platforms include Hugging Face, Kaggle, Google Dataset Search, and Papers with Code. The complete catalogue and taxonomy annotations are available at <https://github.com/ymzhang-cs/prompt-dataset-analysis>. Each catalogue entry provides its Publisher, Size, License, Link, and Description. The license labels reflect rights declared by the publishers when we collected the data. Original license

terms still apply to collections included within another dataset.

Source units and analysed prompts. Catalogue cards report the native source unit, whereas Table 1 reports prompts retained after preprocessing. BoredHumans has 964 source instances and 956 retained prompts. Dolly has 15,011 source records and 14,779 retained prompts. The medical-o1 release reports 90,120 records across its configurations; selection of the analysed English prompt view, deduplication, and quality checks yield 19,679 retained prompts. For Self-Instruct, the analysed self-generated configuration has 82,612 source records and 81,673 retained prompts; the separate 52,445-example P3 configuration is not analysed. ShareGPT has 90,665 source threads, which expand to 226,625 user turns and yield 181,577 retained prompts. OASST1 has 161,443 messages; the 84,437/4,401 train/validation rows include both roles, while our user-role view contains 55,507 messages, 22,850 English candidates, and 22,079 retained prompts. The released catalogue records both native and retained units for these six cases.

G.1 Taxonomy Annotation Protocol

Unit and evidence. Each annotation applies to one dataset. Two authors independently labelled all 129 catalogue entries. They used the dataset card, official documents, the collection method stated by the publisher, and a fixed sample of prompts. The six scored facets cover all three taxonomy dimensions. Source includes publisher type and generation process. Content includes display form and domain. Target includes cognitive intent and downstream task. We document language and modality without κ because their labels are nearly deterministic. We exclude format, placeholder use, prompt engineering method, and extra attributes from the agreement statistics because they require more inspection at the source or prompt level. The released supplement records two additional dataset-level operational labels, RLHF use and dynamicity, but we do not retroactively include them in the reported six-facet agreement study.

Ambiguity rules. Annotators record a dominant value only when the evidence supports it. Generation process permits *Unknown*. Format and placeholder use permit *Mixed*. Prompt engineering method permits multiple or unspecified values. When source inspection provides enough evidence,

Table 7: Independent taxonomy agreement at the dataset level across all 129 datasets.

Dimension	Facet	Cohen’s κ
Source	Publisher type	.829
Source	Generation process	.813
Content	Display form	.862
Content	Domain	.857
Target	Cognitive intent	.606
Target	Downstream task	.838

the catalogue descriptions preserve supported parts of mixed and multiple attributes; the release does not claim exhaustive structured labels for these unscored attributes. For example, BoredHumans combines role play and chain of thought methods. It contains prompts with and without placeholders. 2.5k-templates also combines templates with slots and prompts without slots. We do not convert remaining variation within a dataset into unsupported labels for individual prompts.

Five facets reach almost perfect agreement. Cognitive intent is the lowest at $\kappa = .606$, only .006 above the conventional moderate–substantial boundary. Both annotation files contain values in all 774 cells and contain no invalid labels. The exact match across all six facets is 62 of 129, or 0.481. Cognitive-intent disagreements occur mainly in broad or multi-purpose prompt libraries whose prompts genuinely span overlapping intents, rather than indicating a mechanically identifiable annotator error. Other disagreements occur at the boundary between prompt instructions and records that contain one instruction and one response. The released repository includes confusion matrices for each facet and both annotation files.

Zero-count categories. The zero counts for Emotion and Data-oriented mean that no catalogue dataset receives either category as its dominant dataset-level label. They do not imply that mixed-purpose collections contain no emotion-related prompts or that no collection supports data synthesis or augmentation as a secondary use.

For the auxiliary RLHF label, *yes* includes datasets intended for reinforcement learning and datasets with preference, DPO, rating, reward-model, or verifier signals suitable for preference optimisation. For dynamicity, *dynamic* denotes a living collection or platform that accepts or advertises continuing updates; a versioned snapshot is *static*. The released CSV exposes every decision, and the statistics script reconstructs Table 8.

Table 8: Catalogue-wide dataset counts for the six scored taxonomy facets and two review-requested auxiliary facets. Percentages use all 129 datasets as denominator and may not sum to exactly 100 because of rounding.

Facet	Count (percentage)
Publisher type	LLM researcher 81 (62.8%); end user 29 (22.5%); domain scientist 19 (14.7%)
Generation process	human 52 (40.3%); model 41 (31.8%); derivative 35 (27.1%); unknown 1 (0.8%)
Display form	single-round 84 (65.1%); instruction 28 (21.7%); multi-round 15 (11.6%); tree 2 (1.6%)
Domain	general 82 (63.6%); other 30 (23.3%); medical 7 (5.4%); coding 5 (3.9%); business 3 (2.3%); creative 1 (0.8%); finance 1 (0.8%)
Cognitive intent	text processing 64 (49.6%); reasoning 35 (27.1%); retrieval 14 (10.9%); analytical 10 (7.8%); creative 6 (4.7%); emotion 0 (0.0%)
Downstream task	model-oriented 103 (79.8%); prompt-oriented 26 (20.2%); data-oriented 0 (0.0%)
RLHF use (auxiliary)	no 113 (87.6%); yes 16 (12.4%)
Dynamicity (auxiliary)	static 118 (91.5%); dynamic 11 (8.5%)

Table 9: Descriptive source-level linguistic means grouped by taxonomy labels. Small groups are not population estimates.

Facet	Label	<i>n</i> src.	SPACE	X	TTR	dep
Cognitive intent	Analytical	2	.000	.035	.862	.015
	Reasoning	2	.048	.005	.622	.051
	Text processing	12	.019	.016	.828	.023
Downstream task	Model-oriented	7	.031	.003	.802	.035
	Prompt-oriented	9	.012	.027	.811	.018

G.2 Taxonomy-organised Linguistic Check

We connect the taxonomy to the bounded linguistic extension by joining the source-level means with catalogue labels. The analysis contains 16 catalogue sources: the seven analysed datasets and nine additional catalogue sources. We exclude the two extension sources that are not catalogue entries and the three general-text controls. Each source receives equal weight, so a large corpus does not dominate a category mean.

The higher formatting and unresolved-arc rates in the two reasoning sources are driven by medical and coding reasoning formats. Model-oriented sources show the same direction relative to prompt-oriented libraries. These values demonstrate how the taxonomy can organise the linguistic measurements, but the small and unbalanced source groups do not establish a catalogue-wide difference.

G.3 Bounded Representativeness Check

We compare the seven analysed datasets with 11 more prompt sources and three general text baselines. Nine of the additional sources are in the current 129-dataset catalogue; finance-alpaca and WritingPrompts are explicitly marked extension sources rather than catalogue entries. The additional sources are 2.5k-templates, awesome-chatgpt, chatgpt-cheatsheet, chatgpt-datascience, finance-alpaca, lamini, OpenCodeReasoning, prompt-genius, prompt-hackers, the-prompt-index, and WritingPrompts. The general text baselines are

Table 10: Group means in the bounded representativeness check. Type to token ratio is included as a description and depends on length.

Group	OOV/x	TTR
Seven analysed prompt datasets	.008	.877
Eleven additional prompt datasets	.020	.776
General text baselines	.003	.705

blog, Reddit, and Wikipedia. We sample up to 400 texts from each source with seed 42 and retain texts with at least 10 characters. We apply the same spaCy `en_core_web_sm` feature pipeline as in the main analysis.

Both prompt groups have higher mean OOV/x rates than the general text baselines. Type to token ratio follows the same ordering of group means and depends on length. It does not provide an independent quality signal. The distributions for individual datasets overlap. Group means for formatting and dependency features do not support a stable distinction across the catalogue. We use this experiment as a bounded check of central tendency on the measured dimensions.

G.4 Cards for the Seven Analysed Datasets

The following cards give the Publisher, Size, License, Link, and Description fields for the seven datasets used in detailed analysis. Table 1 summarises these datasets.

1. 1100+ ChatGPT Prompts for Business

- **Publisher:** Chris Porter
- **Size:** 1235 instances
- **License:** -
- **Link:** <https://chatgpt-business-prompts.notion.site/1100-ChatGPT-Prompts-for-Business>
- **Description:** “1100+ ChatGPT Prompts for Business” is a Notion dataset with 1,235 prompts for business tasks. Topics

include buyer personas, content strategy, digital marketing, email campaigns, market research, product development, and finance. The prompts provide guidance for professional and business users.

2. BoredHumans

- **Publisher:** Impulse Communications, Inc.
- **Size:** 964 source instances; 956 retained prompts
- **License:** -
- **Link:** <https://boredhumans.com/prompts.php>
- **Description:** BoredHumans combines prompts from sources that include Awesome ChatGPT Prompts, Data Science Prompts, and Tree of Thought Prompting. It covers several domains and prompt styles.

3. databricks-dolly-15K

- **Publisher:** Databricks
- **Size:** 15,011 source records; 14,779 retained prompts
- **License:** CC-BY-SA-3.0
- **Link:** <https://huggingface.co/datasets/databricks/databricks-dolly-15k>
- **Description:** Databricks Dolly 15K contains over 15,000 instruction and response pairs written by Databricks employees. Its eight categories include brainstorming, classification, closed and open question answering, generation, information extraction, and summarisation. Some records use context from Wikipedia. The English dataset supports model tuning, synthetic data generation, and data augmentation under the stated license.

4. medical-o1-reasoning-SFT

- **Publisher:** The Chinese University of Hong Kong, Shenzhen et al.
- **Size:** 90,120 source records across configurations; 19,679 retained English prompts
- **License:** Apache-2.0
- **Link:** <https://huggingface.co/datasets/FreedomIntelligence/medical-o1-reasoning-SFT>

- **Description:** medical-o1-reasoning-SFT supports supervised model tuning for medical reasoning in HuatuoGPT-o1. It contains English and Chinese instruction and response pairs generated by GPT-4o for clinical problems and checked by a medical verifier. Separate configurations provide data in one language or a mix of languages. The dataset supports question answering and text generation under the stated license.

5. OASST1

- **Publisher:** OpenAssistant
- **Size:** 161,443 source messages; 22,079 retained English prompt messages
- **License:** Apache-2.0
- **Link:** <https://huggingface.co/datasets/OpenAssistant/oasst1>
- **Description:** OpenAssistant Conversations contains 161,443 messages in 66,497 conversation trees. The 84,437 training and 4,401 validation rows include both user and assistant roles and therefore are not prompt counts. Our prompt view contains 55,507 user-role messages, 22,850 English candidates, and 22,079 prompts after final deduplication and length checks. The dataset includes quality ratings and supports supervised model tuning and reward-model development.

6. Self-Instruct

- **Publisher:** University of Washington et al.
- **Size:** 82,612 source records; 81,673 retained prompts
- **License:** Apache-2.0
- **Link:** https://huggingface.co/datasets/yizhongw/self_instruct
- **Description:** The analysed configuration contains 82,612 self-generated prompt-completion records and yields 81,673 prompts after extraction, deduplication, and quality checks. The separate 52,445-example P3 configuration caused the earlier card mismatch and is not analysed. Other configurations include Super Natural Instructions and expert-written evaluation tasks.

7. ShareGPT90K

- **Publisher:** RyokoAI
- **Size:** 90,665 source threads; 181,577 retained prompt turns
- **License:** CC0
- **Link:** <https://huggingface.co/datasets/liyucheng/ShareGPT90K>
- **Description:** ShareGPT90K contains 90,665 conversational threads. Expanding the threads yields 226,625 user prompt turns; language, deduplication, and quality checks retain 181,577 prompts. Thus this card reports both source threads and the processed prompt turns used in Table 1.

H Additional Experimental Results

H.1 Token Level Analysis

Figure 5 compares sequences of three, four, and five words for all datasets. Figure 4 gives the ShareGPT results. Figure 6 compares the five most common sequences across datasets.

The common sequences mainly contain commands and topic terms, which matches the main paper. The figures also show repetition and overlapping sequences.

1. Some datasets contain unusual common sequences. Examples include “*identify which instrument be string*” in dolly-15 and “*The quick brown fox jumps over the lazy dog*” in Self-Instruct. These sequences show repeated content in template inputs or construction instructions. Such repetition may affect the balance of the content.
2. Fixed sentences produce several overlapping sequences. For example, “*The quick brown fox jumps over the lazy dog*” produces sequences such as “*quick brown fox jump over*” and “*jump over the lazy dog.*”

H.2 Syntactic Analysis

Table 11 reports the proportion of every dependency type found in each dataset. Table 12 reports the proportion of every part of speech tag. Figure 7 shows the ten most common verbs and their five most common direct noun objects. The main paper gives the corresponding figures for medical-o1 and ShareGPT.

Table 11: Proportions of all detected dependency types. A dash marks a type that was not detected in a dataset.

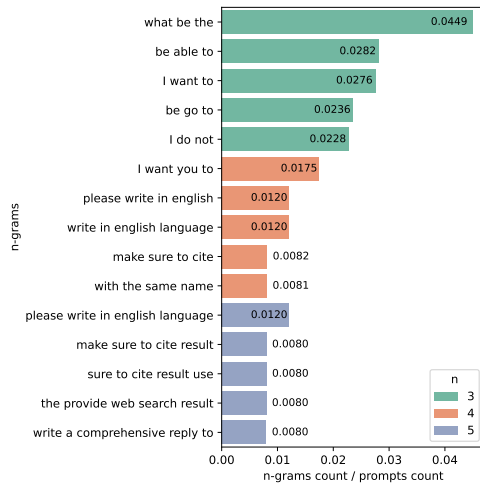
Dependency Type	1.1k-business	BoredHumans	dolly-15k	medical-o1	OASST1	Self-Instruct	ShareGPT
punct	0.1227	0.1985	0.1445	<u>0.1216</u>	0.1273	0.1863	0.154
prep	0.0759	<u>0.0672</u>	0.0866	0.1013	0.0816	0.0676	0.0764
det	<u>0.0518</u>	0.0692	0.0961	0.0906	0.0841	0.0838	0.0693
pobj	0.0718	<u>0.062</u>	0.0817	0.0979	0.076	0.0645	0.0711
nsubj	0.0596	0.0545	0.065	<u>0.0469</u>	0.0739	0.0596	0.0562
ROOT	0.0528	0.0462	0.0768	0.0444	0.0604	0.0792	<u>0.0437</u>
amod	0.0573	0.0527	0.0469	0.1072	0.0523	<u>0.0384</u>	0.048
doj	0.0904	0.0665	0.0447	<u>0.0315</u>	0.0594	0.057	0.0519
compound	0.0742	0.0471	0.0719	0.0716	0.0436	<u>0.022</u>	0.0576
conj	0.0457	0.0494	0.0569	0.0391	<u>0.0343</u>	0.0359	0.0371
aux	0.0642	0.0425	0.0257	<u>0.0143</u>	0.0495	0.0302	0.0355
dep	0.0095	0.0306	<u>0.007</u>	0.0183	0.0218	0.0611	0.0577
cc	0.04	0.0297	<u>0.0203</u>	0.0291	0.0289	<u>0.0203</u>	0.0287
advmod	0.0269	0.0273	0.0263	0.023	0.0383	<u>0.0224</u>	0.0299
poss	0.0401	0.0183	<u>0.0084</u>	0.0132	0.0118	0.0124	0.0116
appos	<u>0.003</u>	0.0223	0.017	0.0099	0.0129	0.0207	0.0238
attr	<u>0.0044</u>	0.005	0.0306	0.014	0.0163	0.0113	0.0083
nummod	<u>0.003</u>	0.0073	0.0096	0.0276	0.0093	0.0136	0.0155
mmod	0.0252	0.0126	<u>0.0042</u>	0.0095	0.0068	0.0043	0.0126
ccomp	0.0058	0.0129	0.0089	<u>0.0044</u>	0.013	0.0142	0.013
relcl	0.0146	0.0088	0.0097	<u>0.0072</u>	0.0109	0.0097	0.0094
xcomp	0.0194	0.0104	<u>0.0042</u>	<u>0.0042</u>	0.0106	0.0079	0.0099
advcl	0.0104	0.0109	<u>0.0056</u>	0.0066	0.0115	0.0086	0.0122
npadvmod	<u>0.0026</u>	0.0068	0.0052	0.0141	0.0051	0.0052	0.0066
acom	<u>0.0034</u>	0.0041	0.0059	0.007	0.0086	0.0091	0.0068
mark	<u>0.002</u>	0.0057	0.0038	0.0033	0.0087	0.0114	0.0095
acl	<u>0.0028</u>	0.0062	0.0055	0.0077	0.0056	0.0079	0.0063
auxpass	<u>0.0016</u>	0.0017	0.0062	0.0101	0.0055	0.0052	0.0059
pcomp	0.008	0.0048	0.0036	0.0053	0.0057	0.0035	0.0051
nsubjpass	<u>0.0014</u>	0.0015	0.005	0.009	0.0046	0.0048	0.0051
neg	<u>0.0015</u>	0.0031	0.0016	0.002	0.0043	0.0033	0.0045
case	0.0035	0.0013	0.0032	0.0024	0.002	<u>0.0011</u>	0.0023
dative	0.0003	0.0029	0.0041	<u>0.0001</u>	0.0035	0.0023	0.0016
prt	0.0014	0.0018	0.0014	<u>0.0006</u>	0.0025	0.0047	0.0025
intj	0.0004	0.0038	0.0012	<u>0.0002</u>	0.0033	0.0016	0.0025
agent	<u>0.0002</u>	0.0003	0.001	0.002	0.001	0.0015	0.0012
expl	0.0	0.0002	0.0004	0.0007	0.0016	0.0014	0.0011
quantmod	0.0	0.0002	0.0005	0.0007	0.0009	0.0014	0.0016
meta	0.0001	0.0016	0.0001	0.0	0.0003	0.0018	0.0012
oprd	<u>0.0002</u>	0.0009	0.0009	0.0007	0.0009	0.0004	0.0008
predet	-	0.0003	0.0005	<u>0.0001</u>	0.0006	0.0009	0.0005
csbj	0.0005	<u>0.0001</u>	0.0004	0.0004	0.0005	<u>0.0001</u>	0.0007
parataxis	-	0.0009	0.0001	0.0	0.0003	0.0002	0.0006
preconj	0.0	0.0001	0.0008	0.0002	0.0002	0.0002	0.0003
csbjpass	0.0	-	0.0001	0.0	0.0	0.0	0.0

Table 12: Proportions of all detected part of speech tags. A dash marks a tag that was not detected in a dataset.

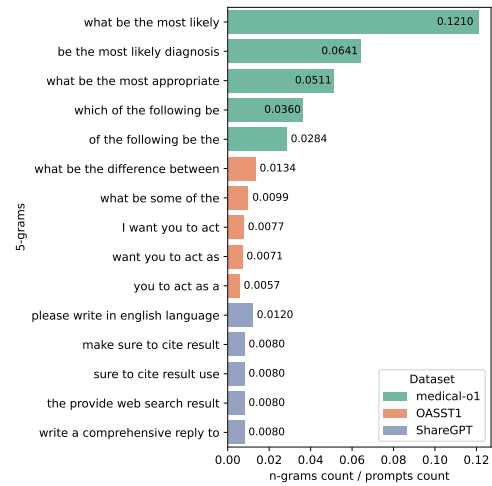
POS	1.1k-business	BoredHumans	dolly-15k	medical-o1	OASST1	Self-Instruct	ShareGPT
NOUN	0.2637	0.2103	<u>0.1899</u>	0.259	0.1946	0.2027	0.1944
PUNCT	<u>0.1094</u>	0.1942	0.1435	0.1158	0.1231	0.1839	0.145
VERB	0.1302	0.1094	0.0871	<u>0.0775</u>	0.1069	0.0999	0.0979
ADP	0.0758	<u>0.0678</u>	0.0858	0.0998	0.0851	0.0701	0.0789
DET	<u>0.0506</u>	0.0693	0.0949	0.0893	0.0839	0.0844	0.0696
PRON	0.0912	0.0708	0.0695	<u>0.0369</u>	0.0869	0.0701	0.0583
ADJ	0.0588	0.0543	0.0538	0.1104	0.0632	<u>0.0498</u>	0.0563
PROPN	<u>0.0219</u>	0.0372	0.1272	0.0515	0.0471	0.0294	0.0703
AUX	0.0458	<u>0.0379</u>	0.0608	0.0382	0.0644	0.0453	0.0423
CCONJ	0.0399	0.0294	0.0209	0.0291	0.0288	<u>0.0204</u>	0.0286
SPACE	-	0.0267	<u>0.0053</u>	0.0175	0.019	0.0504	0.0517
PART	0.0358	0.0223	0.013	<u>0.01</u>	0.0222	0.0172	0.0213
NUM	<u>0.0041</u>	0.0146	0.014	0.0309	0.0153	0.0282	0.0273
ADV	<u>0.0097</u>	0.0259	0.0107	0.0209	0.0238	0.0153	0.0247
SCONJ	0.0199	0.0105	0.0198	<u>0.007</u>	0.0242	0.0197	0.0152
X	0.035	0.0128	0.0015	<u>0.0005</u>	0.0044	0.0075	0.0082
SYM	0.008	0.0037	<u>0.0008</u>	0.0049	0.0035	0.0031	0.0073
INTJ	<u>0.0002</u>	0.003	0.0012	0.001	0.0037	0.0027	0.0028

The additional data support two findings. The medical-o1 dataset has a high proportion of numerical modifiers at 0.0276 and passive auxiliaries at 0.0101. It also has a high numeral rate of 0.0309. These features reflect medical language that emphasises terms, precision, processes, and outcomes. The verb and noun pairs in 1.1k-business include business phrases such as “*create plan*” and “*create strategy.*” The pairs in BoredHumans, OASST1, and Self-Instruct cover more general uses.

In dolly-15k, the most common verb and noun pairs have a skewed distribution. The most common nouns occur mainly with one or two verbs. Pairs such as “*tell i,*” “*give list,*” and “*classify each*” also provide little information about the task. The writing habits of individual annotators may contribute to this pattern.



(a) Five most common n grams in ShareGPT for n equal to 3, 4, and 5.



(b) Five most common sequences of five words in medical-o1, OASST1, and ShareGPT.

Figure 4: Sequences of three, four, and five words within a dataset and across datasets. Each ratio is the count of a sequence divided by the number of prompts in the dataset.

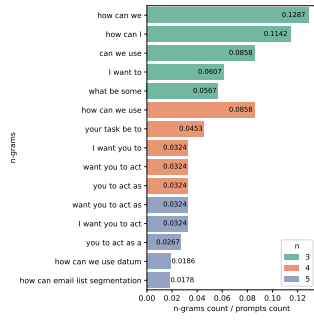
Table 13: Three tokens with the highest TF-IDF weights in each dataset.

Dataset	Three largest token weights
1.1k-business	content (0.308), email (0.284), marketing (0.245)
BoredHumans	act (0.269), want (0.261), write (0.217)
dolly-15k	list (0.338), given (0.246), following (0.241)
medical-o1	old (0.427), year (0.367), patient (0.256)
OASST1	write (0.307), like (0.216), does (0.207)
Self-Instruct	output (0.766), input (0.292), task (0.243)
ShareGPT	write (0.190), use (0.180), data (0.157)

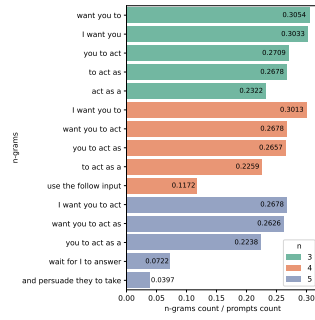
H.3 Semantic Analysis

Figure 8 shows the sampled embedding points after PCA for all seven analysed datasets. The overview panel places the datasets written by people in a shared PCA space.

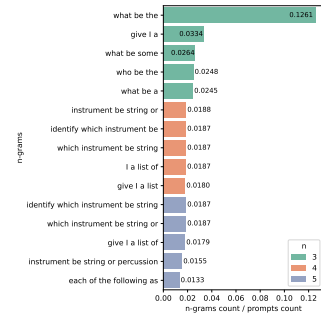
Datasets with focused topics form more concentrated groups. Examples include 1.1k-business and medical-o1. Datasets with broader topics form more dispersed groups. Examples include ShareGPT and Self-Instruct.



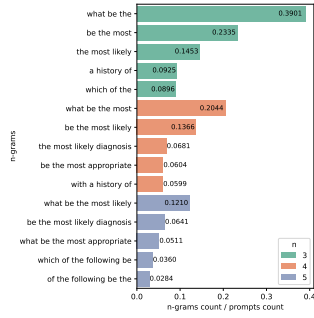
(a) Five most common n grams in 1.1k-business for n equal to 3, 4, and 5.



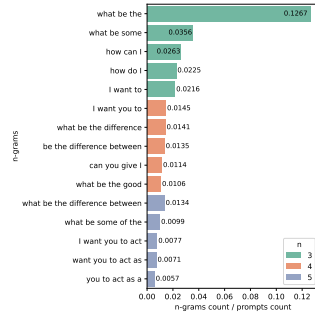
(b) Five most common n grams in BoredHumans for n equal to 3, 4, and 5.



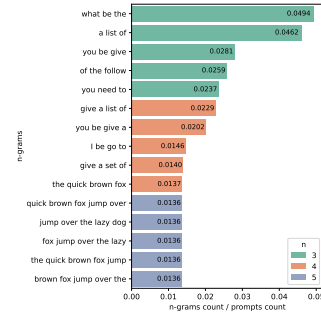
(c) Five most common n grams in dolly-15k for n equal to 3, 4, and 5.



(d) Five most common n grams in medical-o1 for n equal to 3, 4, and 5.

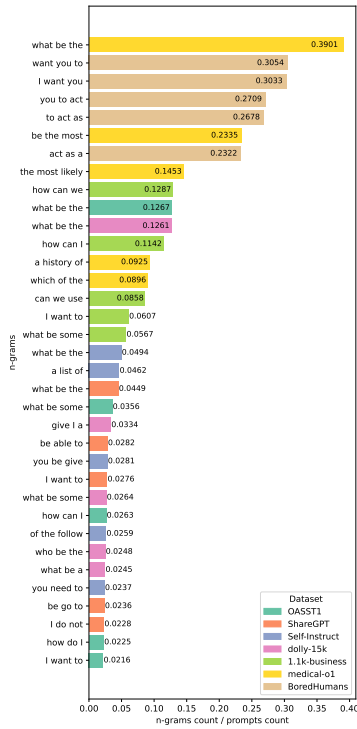


(e) Five most common n grams in OASST1 for n equal to 3, 4, and 5.

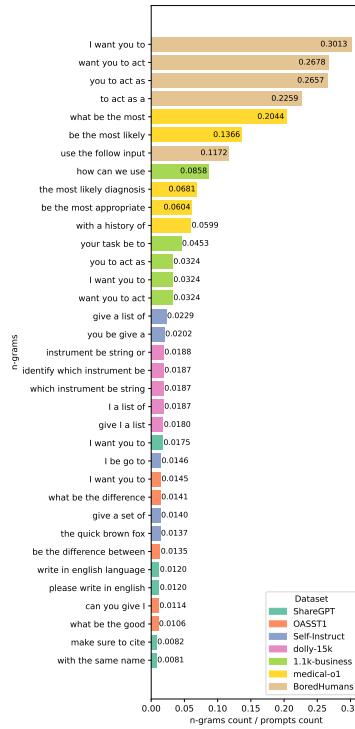


(f) Five most common n grams in Self-Instruct for n equal to 3, 4, and 5.

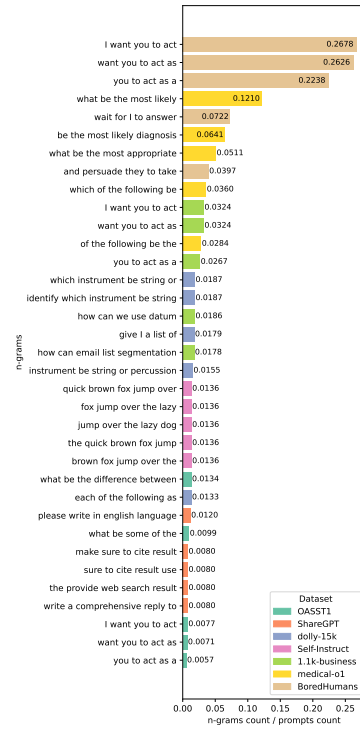
Figure 5: Sequences of three, four, and five words within each dataset.



(a) Sequences of three words.



(b) Sequences of four words.



(c) Sequences of five words.

Figure 6: The five most common n grams across datasets.



(a) 1.1k-business



(b) BoredHumans



(c) dolly-15k



(d) OASST1



(e) Self-Instruct

Figure 7: The ten most common verbs and their five most common direct noun objects in prompt datasets.

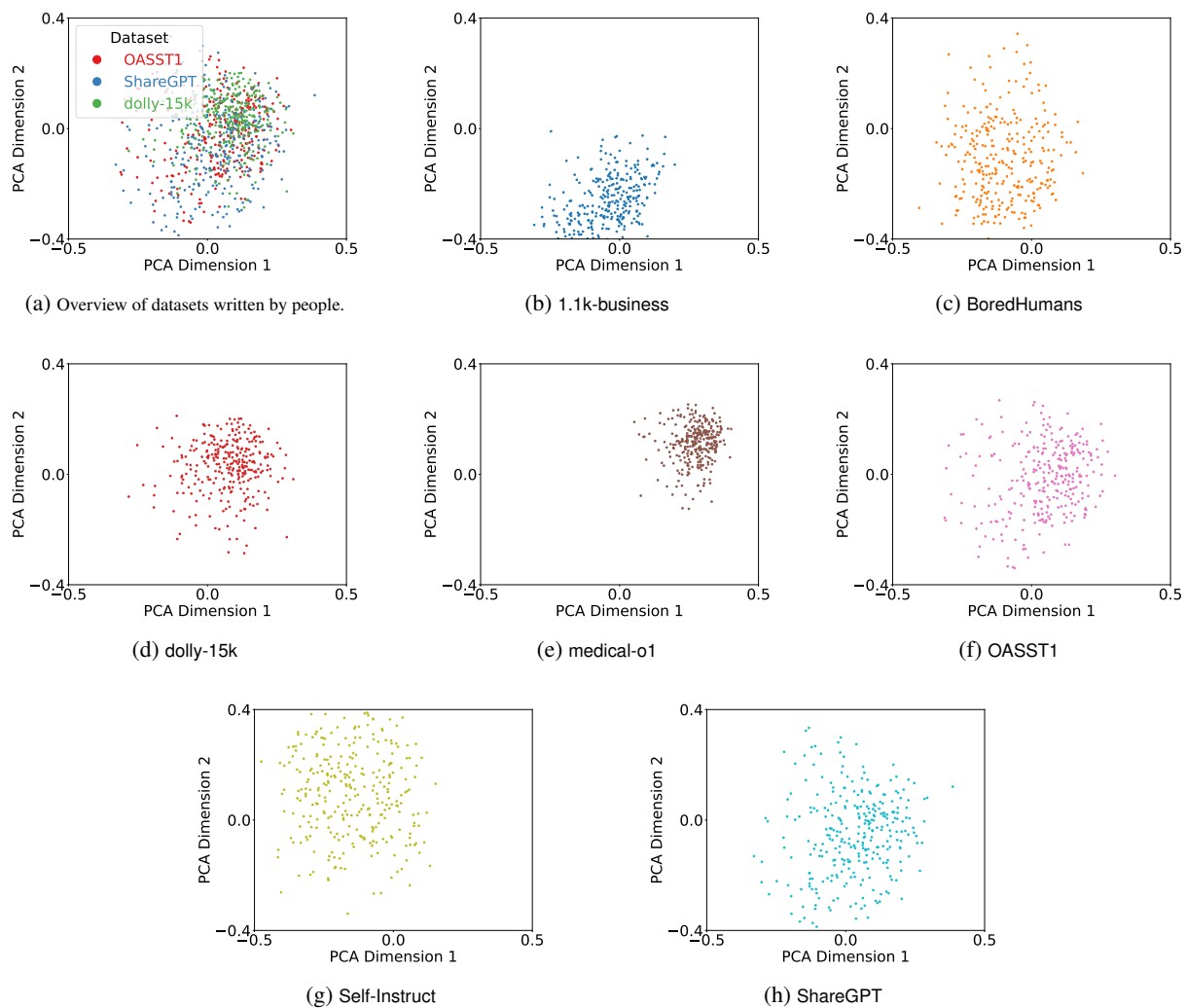


Figure 8: PCA projection of prompt embeddings from 500 sampled prompts in each dataset. Panel (a) places all datasets written by people in a shared space. Panels (b) through (h) show each analysed dataset.

H.4 Application

This section gives supplementary figures and tables for the application experiments in §6. Sections H.4.1 and H.4.2 show MLP training and classification errors. Section H.4.3 shows the input features that drive classifier decisions. Section H.4.4 reports associations between the same linguistic features and elicited response quality under GPT-4 and human ratings.

H.4.1 Prompt Filtering

Figure 9 gives the prompt filtering training results that complement Table 4. Panel (a) shows training and validation loss and F1 for POS, Dependency, Syntactic, TF-IDF, Embedding, and Combined. These curves allow comparison of convergence, final loss, and the generalisation gap. The validation curves follow the training curves closely under a limit of 100 epochs and early stopping. This pattern shows that the 63 dimensional Syntactic representation has limited overfitting on the 6,000 samples. Panel (b) reports F1 under five fold cross validation. Sentence embeddings reach about 0.90. The Syntactic representation reaches 0.822, which is over 91% of embedding F1, and requires no embedding inference or corpus vocabulary. The Combined representation with 5,447 dimensions provides a small improvement over embeddings. For the Isolation Forest, contamination is fixed at 0.30 and recall is evaluated on the fitted prompt examples; exactly 30% are therefore designated outliers, mechanically producing recall 0.700 for every representation. This is a fixed operating point rather than an independently estimated recall.

H.4.2 Source Domain Routing

Tables 14 and 15 extend the Phase B and Phase C results in Figures 3b and 3c. The tables report POS, Dependency, Syntactic, TF-IDF, Embedding, and Combined. K-Means on the Sentence-BERT representation with 384 dimensions recovers the five domains without labels. It reaches NMI = 0.80 and Purity = 0.92. POS, Dependency, and Syntactic reach NMI at most 0.42, and TF-IDF reaches 0.45. The Combined representation with 5,447 dimensions reaches NMI = 0.69 under K-Means and remains strong under supervised training.

Figure 10 shows the supervised Phase A results. The top row shows training, and the bottom row shows confusion matrices. Panel (a) shows stable convergence under early stopping. The Macro-F1 ranking in panel (b) matches Figure 3a. Panels (c)

Table 14: Complete Phase B evaluation of source domain routing. The table reports K-Means NMI, ARI, and Purity and Nearest Centroid accuracy for all six feature representations. Sentence embeddings form the clearest five clusters. The high dimensional Combined representation gives weaker distance based clusters and remains strong under supervised classification.

Feature	K-Means NMI	K-Means ARI	Purity	NC Accuracy
POS	.383	.349	.623	.746
Dependency	.380	.345	.603	.771
Syntactic	.421	.406	.675	.794
TF-IDF	.450	.200	.396	.915
Embedding	.796	.805	.916	.927
Combined	.689	.453	.591	.950

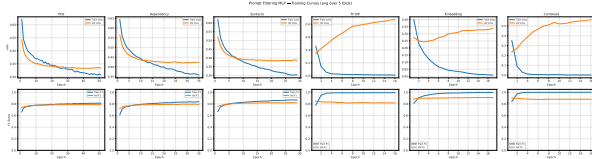
Table 15: Complete Phase C evaluation of source domain routing on an 80/20 stratified split. POS with 18 dimensions reaches 85.0% Macro-F1. Sentence embeddings reach 97.4%, and Combined reaches 96.6%. Confidence is the average maximum softmax probability for the predicted class on the test prompts.

Feature	Accuracy	Macro-F1	Avg Confidence
POS	.851	.850	.881
Embedding	.974	.974	.990
Combined	.966	.966	.990

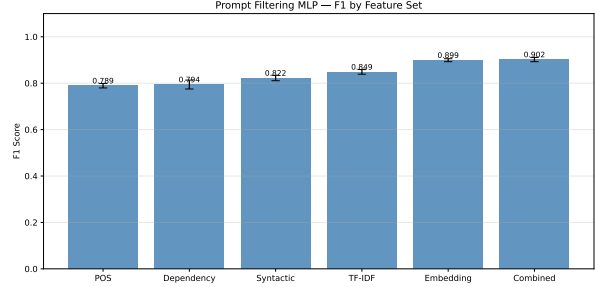
through (e) compare POS, Embedding, and Combined. POS produces a strong diagonal pattern, especially for Medical and Coding. Embedding and Combined further reduce confusion between Business and Finance and between Business and Creative. These pairs account for nearly all remaining errors.

Figure 11 repeats the confusion analysis on the 20% test split in Phase C. All three feature sets preserve the diagonal pattern. This result confirms that Phase A performance does not result from leakage across cross validation folds. POS has the most confusion between Business and Creative because both use command verbs such as “write” and “draft.” Embedding and Combined produce almost perfect confusion matrices.

Cross dataset transfer in Phase D. Each domain and collection contributes up to 800 prompts with seed 42. Pooled cross validation uses five stratified folds over collections A and B. Transfer trains on all available examples from A and evaluates on target collection B. The full linguistic representation has 55 dimensions. It contains 17 POS rates, 27 dependency rates, and 11 basic statistics. The structural representation has 47 dimensions after removing eight features that depend on formatting. These representations differ from the Syntactic representation with 63 dimensions used in Phases A through C.



(a) Training and validation curves for each feature set.



(b) F1 comparison across feature sets.

Figure 9: Prompt filtering results. Panel (a) shows MLP training. Panel (b) shows final F1 for each feature set. Combined reaches the highest F1, and sentence embeddings give a similar result. The Syntactic representation has 63 dimensions and reaches over 91% of embedding F1 without embedding inference.

Table 16: Phase D recall for each target collection. Table 5 gives Macro-F1.

Representation	Bus.	Med.	Code	Fin.	Creative
Formatting (8 dim)	.000	.561	.860	.249	.101
Full (55 dim)	.121	.973	.549	.770	.264
Structural (47 dim)	.192	.938	.995	.751	.243
Embedding (384 dim)	.938	.968	.821	.945	.201

The large decrease from pooled cross validation confirms that Phase A performance partly reflects conventions specific to each collection. Removing formatting features improves overall syntactic transfer. Sentence embeddings transfer best. Creative recall is low because **WritingPrompts** contains story premises and the **No Robots** generation subset contains broader generation instructions. Syntactic features also give weak Business transfer, and sentence embeddings give strong Business transfer.

Figure 12 extends Phase C with a routing gate based on confidence. The left panel compares maximum softmax confidence for 1,200 test prompts from the five domains and 1,000 general prompts from **ShareGPT** and **OASST1**. The right panel changes the decision threshold θ and reports precision, recall, and rejection rate. Confidence gives ROC-AUC 0.843. This value shows useful separation with some remaining errors and no additional distribution detector (Ong et al., 2025; Chen et al., 2024b).

Figure 13 shows the Phase B metrics from Table 14 for all six feature sets. Sentence embeddings lead every clustering metric. TF-IDF and Combined give lower scores even though they include the full vocabulary with 5,000 words. POS, Dependency, and Syntactic form a lower group. This result supports the claim in §6.2 that semantic embedding geometry captures domain identity more

strongly than lexical overlap.

H.4.3 Feature Importance

Table 17 summarises mean absolute gradient importance. The score $\text{Imp}(i)$ is the expected absolute partial derivative $\mathbb{E}_x[|\partial \logit_c(x)/\partial x_i|]$ averaged across classes. A large value means that small changes to feature i produce large changes in classifier logits. Each task has a global ranking, a heatmap for each class, and a radar chart for the leading features.

Figure 14 reports importance for prompt filtering. Panel (a) ranks the 18 POS tags by global gradient magnitude. **SPACE** captures whitespace in numbered steps, code blocks, and bullet lists. **PRON** captures second person pronouns in commands, and **ADV** captures directives such as “step by step.” These three tags have the largest values in the POS-only representation. Panel (b) shows importance for each class in the Syntactic representation with 63 dimensions. Its global top five are *nsubj* (.648), *SPACE* (.640), *attr* (.622), *PRON* (.594), and *ADP* (.561). The general *dep* arc instead leads the separate Dependency-only representation at .616. Panel (c) shows the leading Syntactic features in a radar chart.

Figure 15 reports three views for source domain routing and compares the two tasks in a fourth panel. Panel (a) ranks POS tags across all domains. The **X** tag captures unknown vocabulary tokens such as the code identifier `np.array`, the medical abbreviation *CBC*, and the financial ticker *AAPL*. It is the second largest signal after **SPACE**. Panel (b) shows a distinct POS subset for each domain. **NUM** and *nummod* lead in Finance, **X** leads in Coding, and **ADJ** leads in Medical. Differences across cells show that one common prompt style cannot explain all five domains. Panel (c) shows

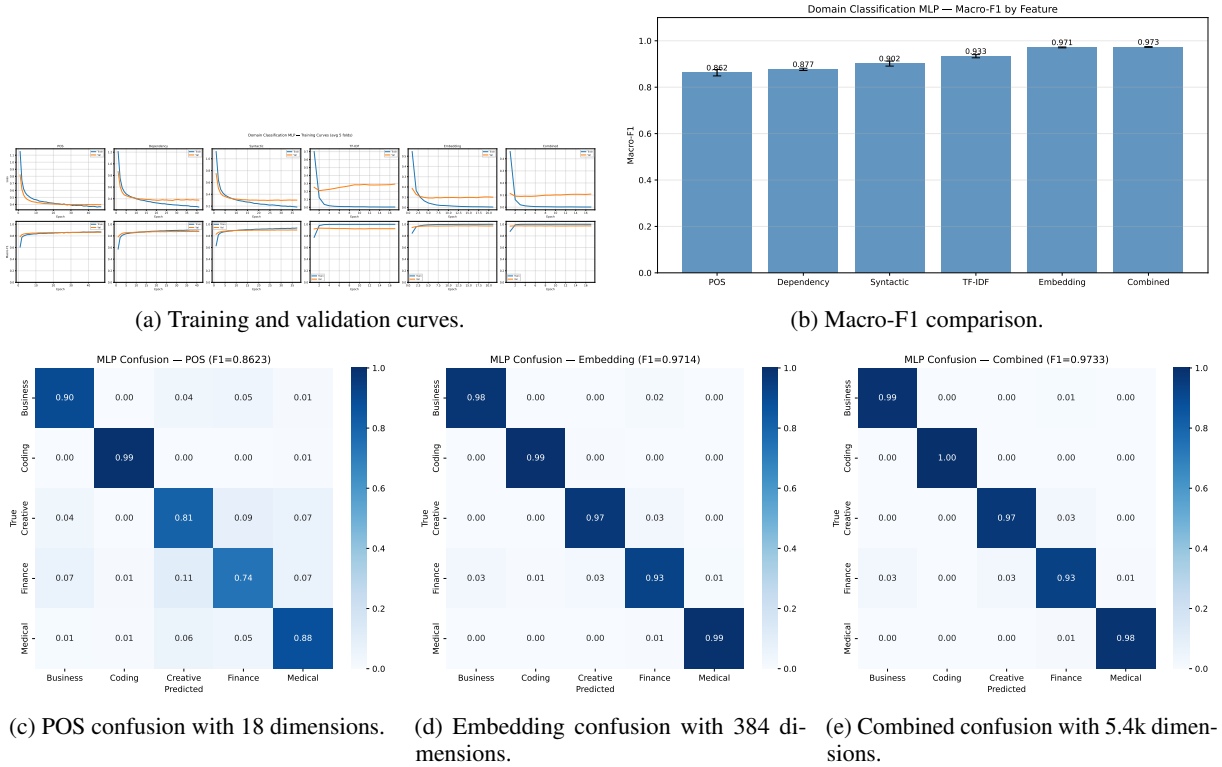


Figure 10: Phase A source domain routing under five fold cross validation. Panels (a) and (b) show MLP training and Macro-F1 for each feature set. Panels (c) through (e) show confusion matrices for three feature sets. POS produces a strong diagonal pattern. Embedding and Combined further reduce confusion between Business and Finance and between Business and Creative.

Table 17: Five largest gradient importance values for Prompt Filtering and Source Domain Routing.

Prompt Filtering		Source Domain Routing	
Feature set	Five largest values	Feature set	Five largest values
POS	SPACE(.766), PRON(.468), ADV(.461), NOUN(.394), PROPEN(.346)	POS	SPACE(2.07), X(.871), PROPEN(.829), NUM(.812), NOUN(.661)
Dependency	dep(.616), meta(.544), attr(.529), aux(.492), nsubj(.474)	Dependency	dep(1.85), ROOT(1.26), compound(.747), nummod(.634), aux(.507)
Syntactic	nsubj(.648), SPACE(.640), attr(.622), PRON(.594), ADP(.561)	Syntactic	SPACE(1.69), ROOT(1.11), X(.961), dep(.639), NOUN(.602)

dep, ROOT, compound, nummod, and aux in a radar chart. Together they capture technical noun phrases and imperative root verbs across domains. Panel (d) compares POS importance between the tasks. Gradient magnitudes are two to three times larger for source domain routing than for prompt filtering. This difference supports using the Syntactic representation first for routing and then for prompt detection when both tasks are needed.

H.4.4 Elicited Response Quality

Figure 16 presents the original UltraFeedback diagnostics before length control. The top row reports associations and differences between score groups. The bottom row reports prediction of Q1 compared with Q4 within UltraFeedback. These panels do not control for prompt length.

Top row. Panel (a) shows Spearman ρ with mean UltraFeedback scores before length control. Type to token ratio has a positive correlation of +0.257.

dep, appos, SPACE, sentence count, and word count have the largest negative correlations. Panel (b) shows Cohen’s d for Q4 compared with Q1. Panel (c) shows the distribution of each feature by quartile. The inverse relation between type to token ratio and prompt length explains most of its visible separation. The partial correlations below show this result.

Bottom row. Panel (d) shows scatter plots before length control. Panel (e) reports correlations for each GPT-4 rating dimension. Panel (f) reports logistic regression AUC for Q1 compared with Q4. POS, Dependency, and basic statistics reach AUC 0.766, 0.773, and 0.727. All features together reach 0.799. These results measure prediction within UltraFeedback. They provide no causal estimate or intrinsic prompt score.

Human ratings and length control. We sample 2,000 unique HelpSteer2 prompts with seed 42 and

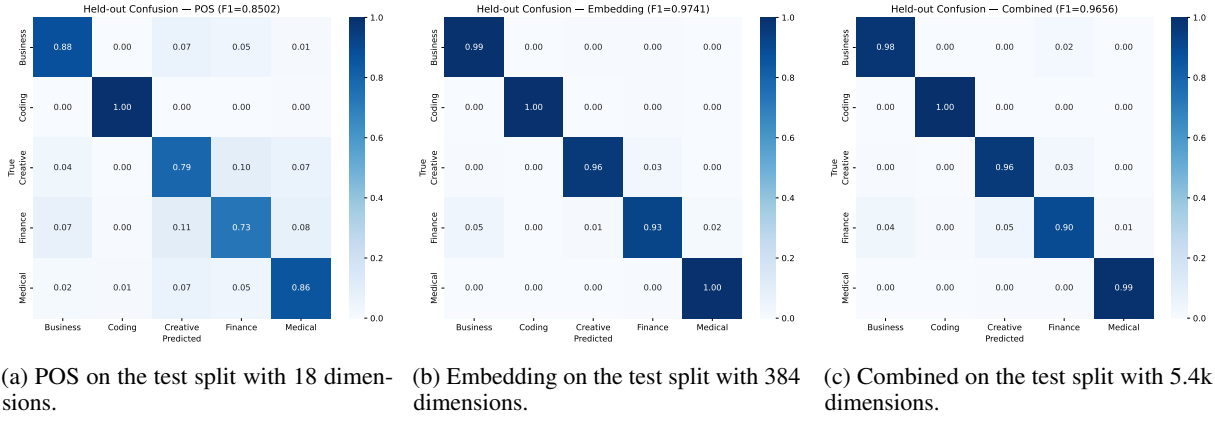
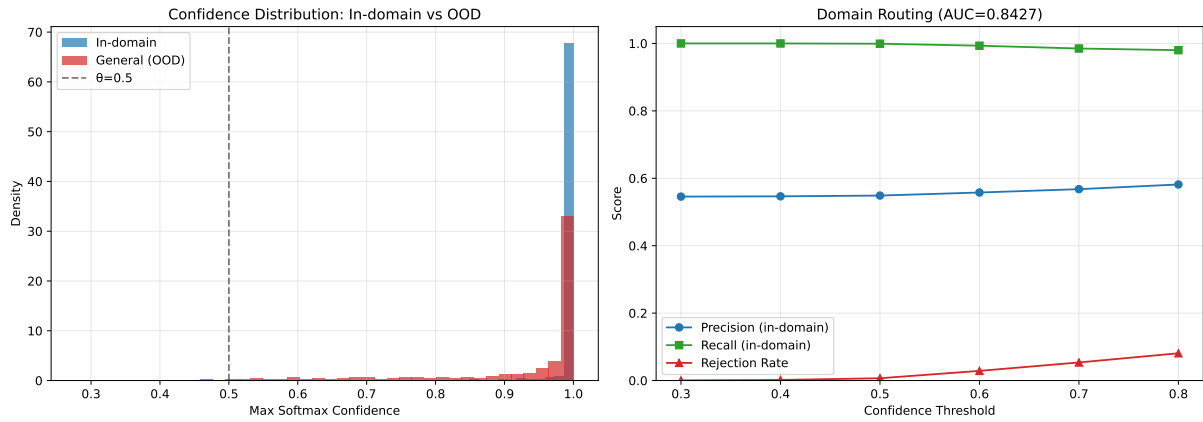


Figure 11: Phase C source domain routing on the 20% test split. POS in panel (a) maintains an almost diagonal confusion matrix. Embedding in panel (b) and Combined in panel (c) further reduce confusion, especially between Business and Creative.



average human helpfulness scores for each prompt. For both datasets, we rank transform each variable. We remove the linear association of prompt word count from the feature and rating ranks. We then compute Spearman correlations between the residuals.

HelpSteer2 reproduces the directions before length control, so the pattern also appears in human ratings. Length control changes the interpretation. The association for type to token ratio becomes nearly zero in both datasets. The structural features remain negative in UltraFeedback. In HelpSteer2, only NUM remains negative, and X becomes positive after control. We therefore limit the claim to observational associations that depend on the dataset. Prompt length, task difficulty, response model capability, and judge preferences may all affect the measured outcome.

H.5 Zero- and Few-shot LLM Baselines

We query `openai/gpt-4o-mini` through the OpenRouter chat-completions endpoint with temperature 0, seed 42, and JSON-object output. The retry path uses a strict JSON schema when deterministic malformed output persists. The protocol fixes 500 evaluation examples per task. Prompt filtering is balanced between 250 prompts, source-stratified across the seven analysed datasets, and 250 general texts from blog, Reddit, and Wikipedia. Source domain routing contains 100 examples per domain from the five Phase D target collections, so evaluation remains cross-dataset relative to the original training sources. Elicited response quality uses a score-stratified 500-prompt subset of UltraFeedback. Few-shot prompts use four demonstrations for filtering, five for routing, and ten for score prediction; demonstrations are disjoint from evaluation examples and, for routing, come from the original rather than target collections. These

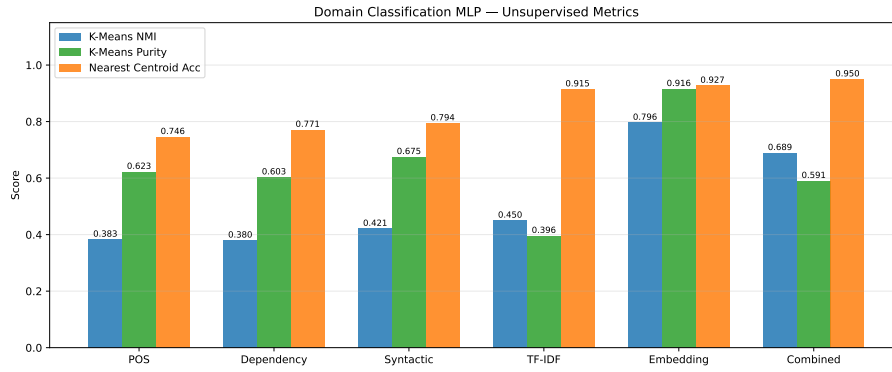


Figure 13: Phase B source domain clustering and centroid classification. Unsupervised K-Means with sentence embeddings forms clear clusters for the five domains and reaches high NMI, ARI, and Purity. The supervised Nearest Centroid accuracy is reported as a simple geometric baseline. The high dimensional Combined representation gives weaker distance based clusters.

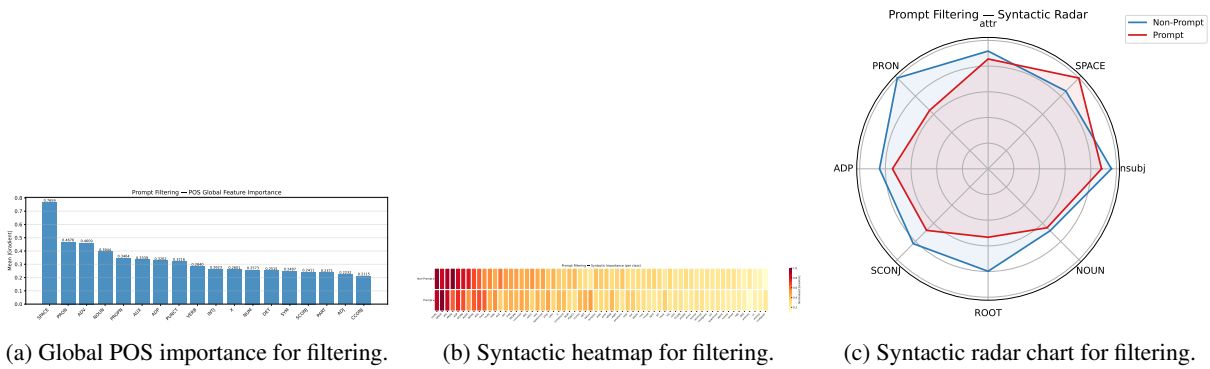


Figure 14: Gradient importance for **prompt filtering**. Panel (a) shows that `SPACE`, `PRON`, and `ADV` lead the POS ranking. Panel (b) shows the per-class gradients in the 63 dimensional Syntactic representation; its global leaders are `nsubj`, `SPACE`, and `attr`. Panel (c) summarises the leading Syntactic features.

subsets match the label spaces and source types of the corresponding applications but not their exact evaluation instances. The effectiveness values are therefore not paired comparisons with the main supervised results.

Effectiveness. Table 19 shows that the LLM performs strongly on prompt filtering and source domain routing. Few-shot prompting raises filtering F1 from .919 to .943 but changes routing Macro-F1 only from .921 to .924. For elicited response scores, it reduces MAE from 1.723 to 1.608 but does not recover score ordering: Spearman ρ remains indistinguishable from zero and Q1–Q4 AUC remains below .5. On the matched but non-paired samples, the LLM therefore exceeds the local models for routing, is only modestly stronger for filtering, and is substantially weaker for within-dataset quality-quartile discrimination.

Efficiency. All 3,000 baseline API outputs parse successfully after retry. The six 500-example runs consume 1,531,480 tokens and cost an estimated \$0.249, or \$0.000083 per evaluated example. Median API latency is 1.54–1.60 s per prompt, about

270 times the 5.7 ms embedding pipeline and 510 times the 3.0 ms Syntactic pipeline. These ratios include network and provider time, whereas the local measurements do not, and the local figures omit hardware amortisation. The local methods therefore provide competitive filtering, stronger quality discrimination, interpretable features, and substantially lower latency without a per-query API call. The release includes the evaluation script, aggregate metrics, token counts, and cost calculations. Raw prompt text and per-example prediction records remain in git-ignored local caches and are not redistributed.

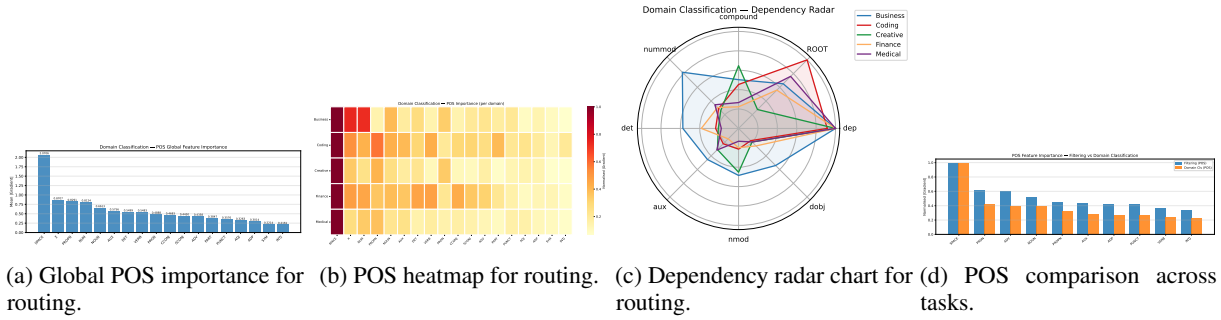
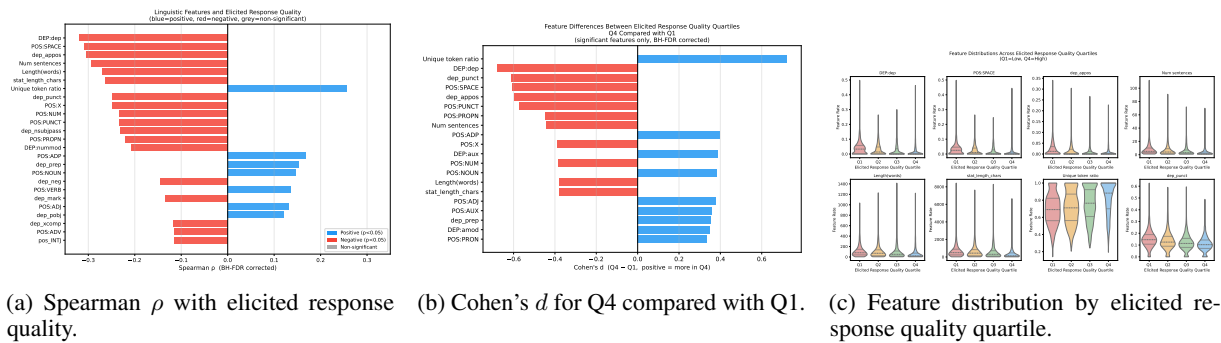


Figure 15: Gradient importance for **source domain routing**. Panel (a) shows that the X tag is the second largest POS signal. This tag captures unknown vocabulary tokens such as code identifiers, medical abbreviations, and financial tickers. Panel (b) shows that each domain uses a distinct POS subset. Panel (c) shows that `dep`, `ROOT`, and `compound` are the main dependency features for domains. Panel (d) shows that POS gradient magnitudes are two to three times larger for source domain routing than for prompt filtering.

Table 18: Spearman correlations with elicited response quality before and after word count control. “n.s.” marks $p \geq .05$.

Feature	UltraFeedback ($n = 10,000$)		HelpSteer2 ($n = 2,000$)	
	Before control	Partial	Before control	Partial
Type to token ratio	+.257	+.024	+.150	-.017 n.s.
dep	-.320	-.212	-.128	-.015 n.s.
SPACE	-.308	-.197	-.127	-.010 n.s.
X	-.248	-.153	-.077	+.052
NUM	-.233	-.136	-.172	-.116
Word count	-.270	—	-.166	—



(a) Spearman ρ with elicited response quality. (b) Cohen’s d for Q4 compared with Q1. (c) Feature distribution by elicited response quality quartile.

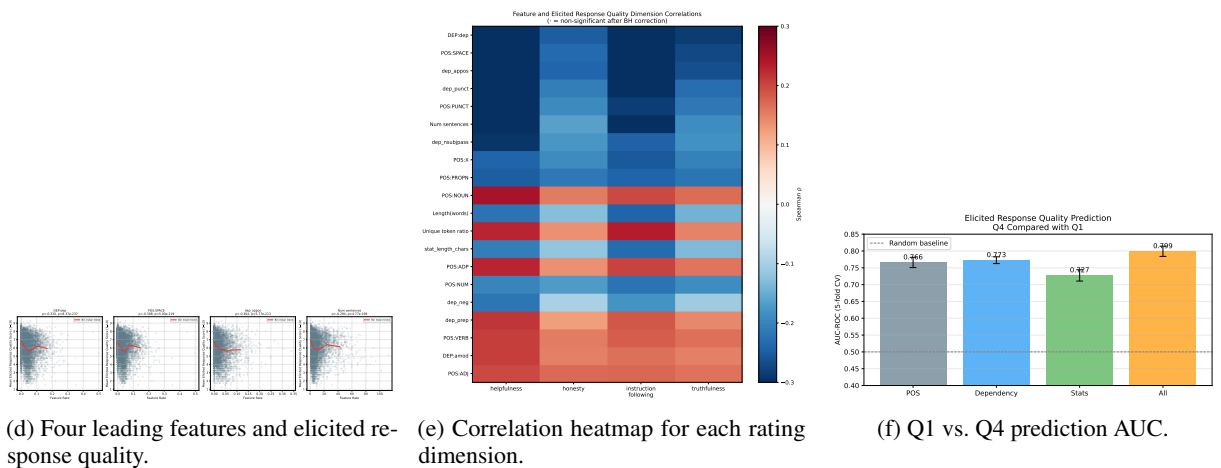


Figure 16: **UltraFeedback diagnostics before length control**. Panels (a) through (c) show Spearman ρ , Cohen’s d for Q4 compared with Q1, and quartile distributions. Panels (d) through (f) show scatter plots, correlations across GPT-4 rating dimensions, and prediction AUC for Q1 compared with Q4 within the dataset. These panels do not control for prompt length.

Table 19: Zero- and few-shot LLM baselines on fixed 500-example task samples. Latency is median end-to-end API time per example. Cost is estimated from recorded input/output tokens at \$0.15/\$0.60 per million tokens. For score prediction, Spearman ρ , Q1–Q4 AUC, and MAE are computed from the same API outputs. Latency and cost are reported once for this run; dashes denote that no additional inference is required.

Task	Effectiveness metric	Value	Zero-shot		Value	Few-shot	
			Latency (s)	Cost (\$)		Latency (s)	Cost (\$)
Prompt filtering	F1	.919	1.600	.0229	.943	1.537	.0749
Source domain routing	Macro-F1	.921	1.550	.0190	.924	1.571	.0612
Elicited response score	Spearman ρ	−.001	1.579	.0188	−.024	1.583	.0524
	Q1–Q4 AUC	.479	—	—	.457	—	—
	MAE	1.723	—	—	1.608	—	—