A Appendix about Benchmark

A.1 Data Availability

Dataset: The whole benchmark along with four split parts can be found in the supplementary materials. We will upload it to a repository in Github and provide the URL.

Code: The code for experiments can be found in the supplementary materials. Similarly, we will upload it to the same repository.

A.2 Details about Data Synthesis

The prompt for employing GPT-4 to generate samples based on task category names is shown in Figure 6. We randomly selected 30% existing task categories and generate 3 samples for each category. After filtering, we obtained a total of 633 synthetic samples.

> Generate an instruction represents the {task category} task, which contains two sentences. Note that the second generated sentences must contain the task word.

Figure 6: The prompt for generating the complex instructions.

Here are some generated examples:

Table 4: Examples of generated complex instructions.

Task category	Examples
Classify Animal	You are a biologist studying a new species discovered in the Amazon rainforest. Classify the animal based on its characteristics, habitat, and behavior.
Generate Rap	Imagine you are a famous rapper who's known for his/her unique style. Generate a rap verse that showcases your creativity and lyrical prowess.
Give Title	You have written an article about the impact of social media on mental health. Give a title to your article that will reflect the content of your article.
Make Poem	Imagine you are sitting by a serene lake during a beautiful sunset. Make a poem that captures this tranquil moment and the emotions it evokes.

A.3 Details about Quality Control

The prompt for employing GPT-4 to check whether instructions belong to its annotated category is shown in Figure 7.

> Check if the given instruction represents the {task category} task. Instruction: {instruction}. Please answer 'yes' or 'no'.

Figure 7: The prompt for generating the complex instructions.

For category merging, we will provide additional details about the merging procedure. Firstly, we select every two categories where both nouns and verbs are synonyms or same words. Then we calculate the cosine similarities of each pair of them by using word embeddings. For two categories where the values of both nouns and verbs pairs are above 0.5, we directly merge them as one category. For categories with values between 0.3 and 0.5, we use GPT-4 to determine whether they describe the same task. If they do, we merge them. For those below 0.3, we directly discard the merge. The prompt for this process is shown in Figure 8.

15

Are {task1} and {task2} represent the same task for instruction?. Please answer 'yes' or 'no'.

Figure 8: The prompt for generating the complex instructions.

A.4 Human Evaluation

While we have highlighted the quantity and diversity of the data in IEB, the quality remains uncertain. To assess this, we randomly select 100 task categories and choose one instance from each. An expert annotator, who is a co-author of this work, then evaluate whether each instruction belongs to its annotated category. The instruction for judgement is the same as Figure 7. The results indicate that 93% of the sample categories are accurate, showing that most of the annotated category labels are correct.

A.5 More Statistics

Besides the dataset partitioning, we provide more information about the statistics of proposed benchmark. We present the distribution of the number of instructions per category in Figure 9. Please note that for categories with more than 100 samples, we randomly retained only 100. Additionally, Figures 10 through 14 provide a more detailed view of the verb-noun distributions, where it is clear that there is no category overlap between EFT and IFT, but there is some overlap between the training and test sets within IFT.
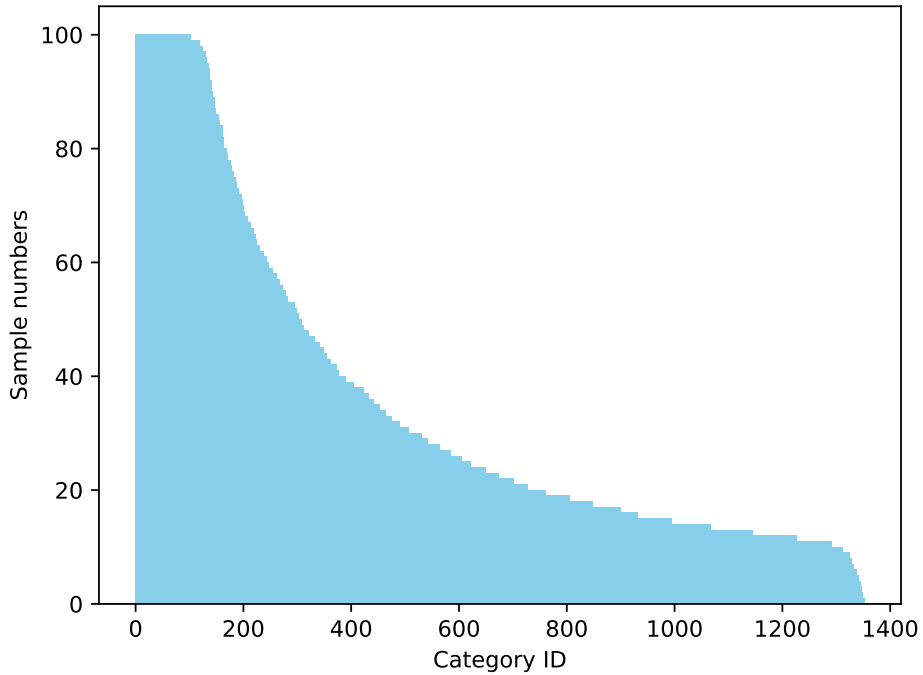
Figure 9: The prompt for generating the complex instructions.

16
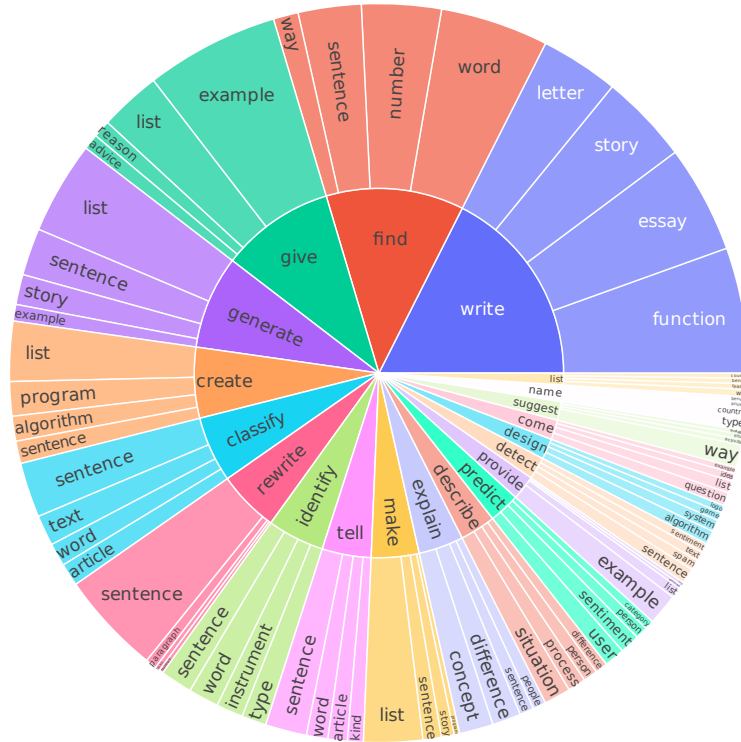
Figure 10: Verb-noun distributions of whole benchmark.
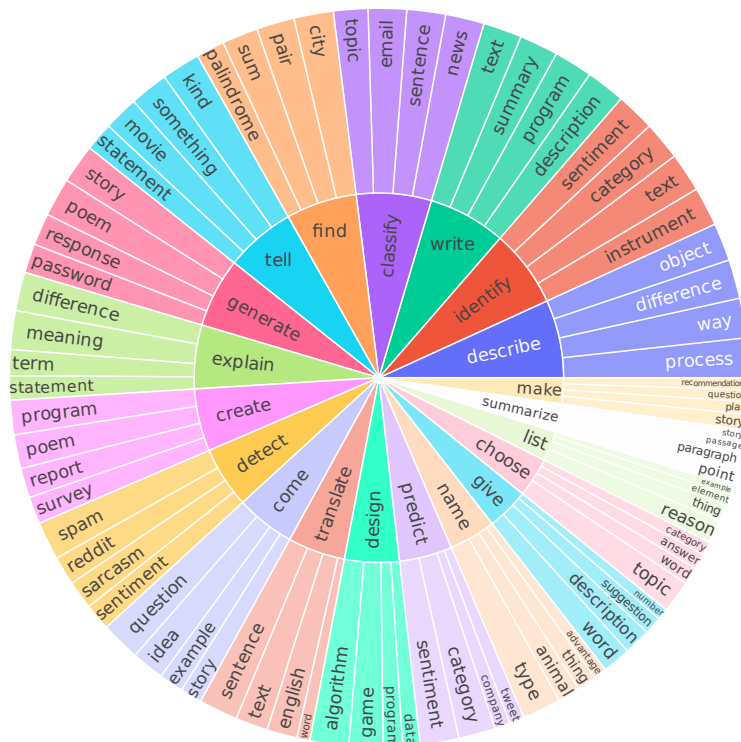
Figure 11: Verb-noun distributions of EFT-train.

17
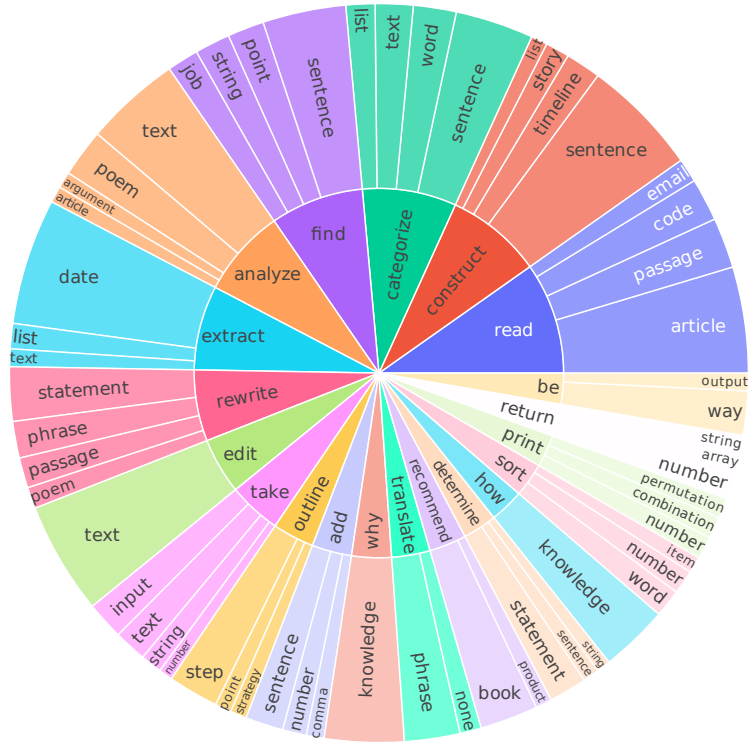
Figure 12: Verb-noun distributions of EFT-test.
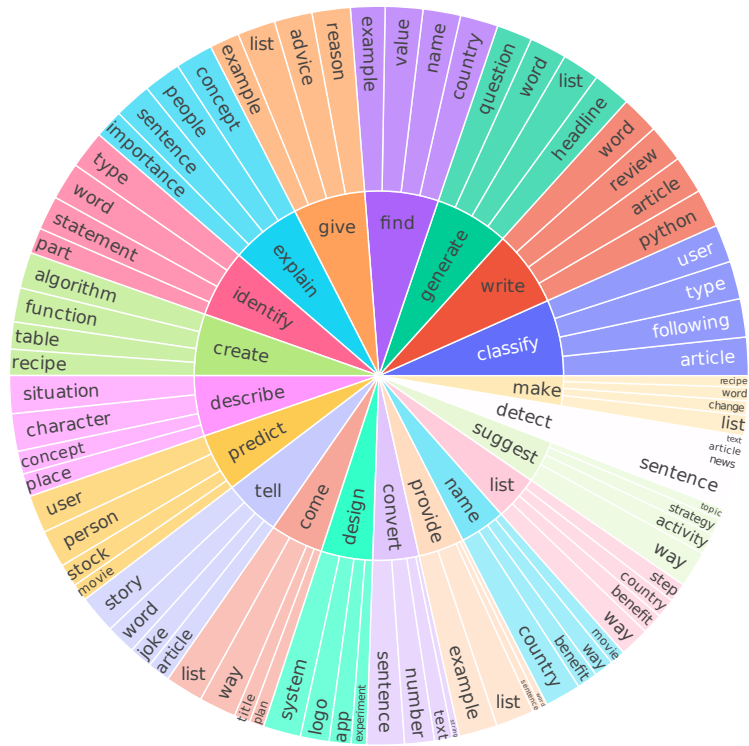
Figure 13: Verb-noun distributions of IFT-train.

18

Figure 14: Verb-noun distributions of IFT-test.

19