Unsupervised Persona Elicitation

Yiding Pei

November 25, 2025

1 Results

We evaluate the performance of different methods on the Global Opinions dataset across three countries: US, DE, and FR.

Table 1: Performance comparison of different methods across three countries (Accuracy).

Country	Zero-shot	Zero-shot (chat)	Many-shot	Gold-label
US	0.5143	0.6393	0.8000	0.7714
DE	0.6230	0.6230	0.7429	0.6429
FR	0.6167	0.6167	0.8143	0.8000

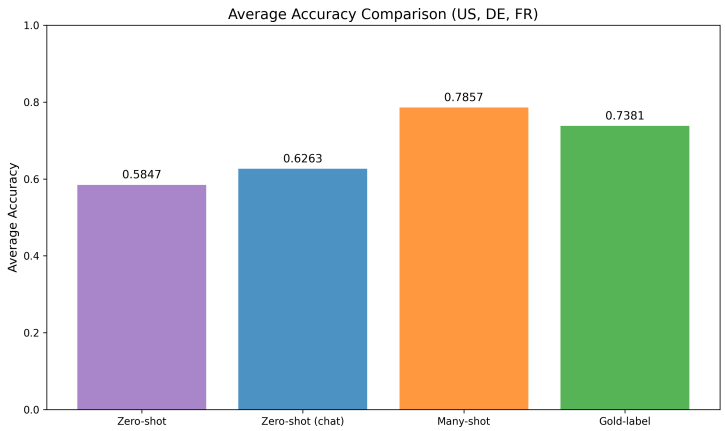

Figure 1: Average accuracy comparison across different methods.

We also analyze the effect of the number of in-context examples on the accuracy for the US dataset.

Table 2: Accuracy vs. Number of In-Context Examples on US dataset.

Num Examples	Many-shot	Gold-label	Random (0.8 acc)
0	0.5143	0.5143	0.5143
10	0.7714	0.7143	0.7286
20	0.7714	0.7143	0.7000
30	0.8286	0.7429	0.6286
40	0.8286	0.7857	0.6857
50	0.8000	0.7714	0.6143

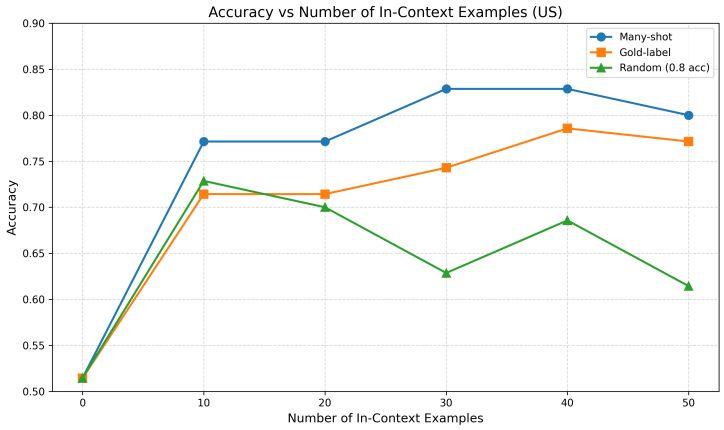

Figure 2: Accuracy vs. Number of In-Context Examples (US).

2 Experimental Setup

2.1 Datasets

We utilize the `Anthropic/llm_global_opinions` dataset, which originally contains approximately 2,560 entries. We filtered the dataset to include only questions with binary answers, resulting in 440 entries. We selected the three most frequent countries for our experiments: Germany (203 entries), the United States (199 entries), and France (197 entries).

For each country, we reserved 70 entries as the validation set, using the remaining data (approximately 120-130 entries) for training. After the unsupervised training process, we obtained approximately 40-60 labeled examples with a labeling accuracy of around 75-80%.

2.2 Models

We used the Llama 3.1 405B Base model for generating labels during the training phase. For testing, we evaluated both the Llama 3.1 405B Base and Instruct

models.

It is important to note that the API used for the Instruct model (Hyperbolic) does not return `logprobs`. Therefore, we employed greedy string matching for evaluation. We excluded examples where the model's output could not be successfully parsed, which accounted for approximately 15% of the test cases.

2.3 Evaluation Metrics (Random Samples)

We compare the ICM-generated labels against random labels with a similar accuracy rate (approximately 0.8). The objective is to determine whether the higher internal consistency of the labels generated by ICM translates to better performance on the test set compared to random labels of equivalent accuracy.

3 Discussion

We observe that the overall trend aligns with the findings reported in the paper for GSM8k and TruthfulQA. However, two specific observations regarding this dataset need further explanation.

First, the performance of gold-label demonstrations is weaker than that of ICM (Many-shot). We assume that this is because many classifications in the Global Opinions dataset have confidence levels around 50%, introducing noise into the gold labels. Unlike datasets such as GSM8k where correctness is verifiable, the "ground truth" in this context is less absolute, making the gold labels potentially less reliable than the self-consistent personas elicited by ICM.

Second, the model achieves strong performance with a very small number of context samples. This suggests that for knowledge activation tasks, the marginal benefit of additional samples diminishes rapidly.