Unsupervised Persona Elicitation

Author Name
Affiliation
Address
email@domain.com

November 24, 2025

Abstract

The abstract should summarize the contents of the paper. It should briefly state the problem of persona elicitation in Large Language Models (LLMs), the limitations of current supervised approaches, and introduce the proposed unsupervised framework. Mention the key methodology and the main experimental results.

1 Introduction

Introduce the context of Large Language Models and their capability to adopt specific personas.

- **Motivation**: Why is persona elicitation important? (e.g., for evaluating alignment, simulating diverse populations, etc.)
- **Problem Statement**: Discuss the challenges with existing supervised methods (e.g., need for labeled data, bias in prompts).
- **Contribution**: Briefly outline the contributions of this work, specifically the unsupervised approach.

2 Related Work

Discuss relevant literature in the following areas:

- **Persona Adoption in LLMs**: How LLMs are currently prompted to take on personas.
- **Unsupervised Learning in NLP**: Relevant unsupervised techniques.
- **Evaluation of LLM Opinions**: Previous work on datasets like Global Opinions or TruthfulQA.

3 Methodology

Detail the proposed unsupervised persona elicitation framework.

3.1 Problem Formulation

Formally define the task. Let M be the model, D be the dataset, etc.

3.2 Unsupervised Elicitation Framework

Explain the core mechanism. How does the method discover or elicit personas without explicit labels?

- **Step 1**: Description of the first step (e.g., clustering, latent variable modeling).
- **Step 2**: Description of the refinement or generation process.

3.3 Algorithm

If applicable, include pseudocode or a step-by-step description of the algorithm.

4 Experimental Setup

Describe how the experiments were conducted.

4.1 Datasets

Describe the testbeds used (e.g., Global Opinions).

- **Global Opinions**: Description of the dataset, size, and characteristics.
- **Other Datasets**: If any.

4.2 Models

List the LLMs used in the experiments (e.g., Meta-Llama).

4.3 Baselines

Describe the baseline methods compared against (e.g., Zero-shot, Few-shot with random labels, etc.).

4.4 Evaluation Metrics

Define the metrics used to evaluate performance (e.g., Accuracy, Consistency, Agreement).

5 Results

Present the experimental findings.

5.1 Main Results

Compare the proposed unsupervised method against baselines. Use tables and figures.

Table 1: Main performance comparison on Global Opinions.

Method	Accuracy	Consistency	Metric 3
Baseline 1	0.00	0.00	0.00
Baseline 2	0.00	0.00	0.00
Ours	**0.00**	**0.00**	**0.00**

5.2 Ablation Studies

Analyze the contribution of different components of the proposed method.

6 Discussion

Interpret the results. Why does the unsupervised method work? What are the limitations?

- **Analysis of Learned Personas**: Qualitative examples of elicited personas.
- **Limitations**: Where does the method fail?

7 Conclusion

Summarize the paper and suggest future research directions.

References