Short Analysis Report

Explaining the 55% Agreement Between GPT-Based Classification and Reference Labels

1. Overview

In this study, a large language model (GPT-3.5) was used to classify open-ended user queries into a predefined categorization scheme consisting of 13 action codes (D.I–OTHER).
The model's output was compared against an existing manually annotated reference dataset. The resulting exact match rate was **55.4%**.

At first glance, this accuracy may appear low. However, a closer inspection of the mismatches reveals that many disagreements arise from **conceptual ambiguity and subjective interpretation**, rather than clear classification errors.

---

2. Nature of the Classification Task

The task involves intent classification of short, open-ended user queries. Such queries are often:

- very brief,
- ambiguous,
- and multi-intent by nature.

Unlike traditional text classification benchmarks, this task does not involve clearly separable classes, long contextual documents, or objectively verifiable labels.
Instead, the categories represent **abstract learning actions**, which already require interpretation by human annotators.

---

3. Conceptual Overlap Between Categories (with Examples)

An analysis of individual mismatches shows that many disagreements occur between **closely related categories**.

Example 1: D.I vs. S.S

Entry ID	Gold Label	GPT Label
17000008	S.S	D.I

This query asks for an explanation of a concept.
While the gold annotation classifies this as **Seeking information (S.S)**, the model interprets it as **problem identification (D.I)**.
Both interpretations are plausible, illustrating the semantic overlap between these categories.

Example 2: E.RF vs. D.I

Entry ID	Gold Label	GPT Label
17000006	E.RF	D.I

Here, the user asks to solve tasks based on provided material.
The gold label **E.RF (Reformatting/Reworking)** assumes the intent is to transform or apply existing content, whereas the model interprets the request as defining a task to be solved (**D.I**).
This mismatch reflects different assumptions about the *primary intent*.

Example 3: E.RV vs. E.O

Entry ID	Gold Label	GPT Label
17000027	E.RV	E.O

The user asks for clarification or checking of content.
This can be interpreted as **Review (E.RV)** or as **Organising/structuring information (E.O)**.
The model's choice differs from the gold label, but remains semantically close.

---

4. Subjectivity of the Reference Labels

The reference dataset represents a **single annotation perspective**:

- Labels were assigned manually.
- No inter-annotator agreement or uncertainty scores are available.

- The labels therefore cannot be considered absolute ground truth.

As a result, disagreements between the model and the reference labels may reflect:

- different but reasonable interpretations,
- borderline cases between categories,
- or implicit assumptions made by the original annotators.

This inherently limits the maximum achievable agreement, even for an ideal classifier.

---

5. Conclusion

The observed **55.4% agreement** does not indicate poor model performance.
Instead, it reflects the intrinsic difficulty of intent classification for short, ambiguous user queries and the conceptual overlap between several categories in the taxonomy.

A manual inspection of mismatches shows that many disagreements occur between **semantically adjacent categories**, where multiple labels are defensible. Consequently, the reported accuracy should be interpreted as a **conservative lower bound** on the model's actual usefulness.

Overall, the results demonstrate that **prompt-based LLM classification is a viable approach** for categorizing open-ended queries in mixed-methods research, while also highlighting the importance of careful category design and qualitative error analysis alongside quantitative metrics.