Part I

The first stage of the project involved manually annotating a natural conversation for dialogue acts. Although the annotation scheme was generally clear and intuitive, several utterances exposed the difficulty of assigning a single functional label to speech in spontaneous interaction. The overall inter-annotator agreement was high (κ = 0.83), indicating that both annotators understood the categories consistently. However, the points of disagreement reveal that the most challenging aspect of the task lies not in identifying grammatical form, but in inferring pragmatic intention.

Short, seemingly straightforward responses were particularly problematic. For example, the utterance **"yeah true↓ true man."** appeared at first glance to be a simple acknowledgement, but the added lexical emphasis (*true man*) can also be interpreted as conveying a stance or claim, which prompted one annotator to label it as an **ACKNOWLEDGEMENT** and the other as a **STATEMENT**. Similar ambiguity was evident in emotionally charged utterances such as **"it was nice feelings↓ I swear."** One annotator viewed this as a **STATEMENT** describing an experience, while the other categorized it as **EXPRESSIVE**, focusing on the affective element introduced by *I swear*.

Another area of divergence emerged around interrogatives. The utterance **"What did you post?"** was labeled as a **QUESTION** by one annotator but as a **DIRECTIVE** by the other, since the request implicitly directs the interlocutor to retrieve and share information rather than simply asking for clarification. Such examples illustrate how grammatical form does not necessarily determine discourse function.

Overall, the disagreements demonstrate that annotators must rely on interpretive reasoning, especially when utterances serve dual purposes. This suggests that pragmatic annotation requires attention not only to words and syntax, but also to conversational context, prosody, and speaker intent, making it a cognitively demanding task despite a seemingly simple label taxonomy.

Error Analysis Table

Utterance	Annotator A	Annotator B	Reason for Disagreement
"yeah true↓ true man."	ACKNOWLEDGEMENT	STATEMENT	Agreement token vs. evaluative stance

shared_memory_transcript_updated

“What did you post?”	QUESTION	DIRECTIVE	Interrogative syntax vs. request for action shared_memory_transcript_updated
“it was nice feelings↓ I swear”	STATEMENT	EXPRESSIVE	Description of event vs. affective emphasis shared_memory_transcript_updated
“I think more! Man.”	STATEMENT	EXPRESSIVE	Opinion-giving vs. emotional emphasis shared_memory_transcript_updated

Part III – Reflection

The final stage of the project examined whether a locally running LLM could perform dialogue-act annotation with a level of consistency comparable to human annotators. After several iterations of prompt refinement, the model was able to produce complete annotations in the required format, demonstrating that prompt engineering can effectively control structural output. However, structural compliance does not equate to interpretive competence, and the comparison between human and machine decisions made this distinction evident.

The inter-annotator agreement between the two human annotators remained high (κ = **0.83**), confirming that the annotation scheme itself is learnable and that humans apply the categories in a reasonably consistent manner. In contrast, the agreement between humans and the LLM was substantially lower. The κ score for **Human 1 vs. the LLM** was **0.147**, and **Human 2 vs. the LLM** yielded **0.155**. Although both values are slightly higher than random performance and therefore an improvement over earlier iterations, they still fall within the range of *slight agreement*. This indicates that the model does not reliably infer the communicative intent underlying each utterance.

A qualitative review of the LLM's output showed recurring errors. The model frequently assigned **STATEMENT** labels to short acknowledgements, treated expressive utterances as informational, and failed to recognize when questions were functioning as directives. These error patterns suggest that the model processes utterances primarily through lexical and syntactic similarity rather than pragmatic interpretation. In other words, it identifies what is said but not what is being done with language.

Overall, the results suggest that while the LLM can reproduce annotation procedures, it does not possess the inferential or contextual sensitivity required for pragmatic judgment.

Automatic annotation may serve as a useful preliminary tool, but human expertise remains essential when communicative intent is central to the analysis.

The progressive improvements observed during prompt refinement show that the model is responsive to clearer instructions and more explicit decision rules. It is therefore plausible that further iterations such as adding more detailed examples, incorporating few-shot demonstrations, or constraining the model with additional meta-prompts could result in higher agreement with human annotators. Due to time constraints, I was unable to fully explore these advanced refinement strategies, but the preliminary improvements suggest that additional experimentation may significantly enhance the model's performance.