Bauhaus-Universität Weimar
Faculty of Media
Degree Programme Human-Computer Interaction

Do Large Language Models Extrapolate Personas of Dialogue Participants from Context?

Master's Thesis

Daria Zhukova
Born Mar. 2, 2000 in Novosibirsk

Matriculation Number 125603

1. Referee: Prof. Dr. Benno Stein
2. Referee: Jun.-Prof. Dr. Maurice Jakesch

Submission date: July 31, 2025

Declaration

Unless otherwise indicated in the text or references, this thesis is entirely the product of my own scholarly work.

Weimar, July 31, 2025

.....
Daria Zhukova

Abstract

This thesis investigates whether large language models (LLMs) can extrapolate the personas of dialogue participants based solely on their previous turns in a conversation. While prior work on persona modelling often relies on explicit profiles or fine-tuned identity representations, this thesis aims to investigate the zero-shot abilities of pretrained LLMs to infer user behaviour from preceding dialogue turns. Three open-weight LLMs are evaluated in two persona-imitation experiments. The first experiment assesses whether the LLMs can mimic the decision-making process by predicting binary donation decisions and donation amounts in persuasive dialogues, while the second experiment examines the capability of LLMs to generate user utterances that reflect dialogue acts of the original conversation from a task-oriented travel booking dataset. Results show that LLMs perform well when the final decision or utterance is present in the dialogue context but struggle to infer unstated intent, frequently overpredicting affirmative responses and misrepresenting social dialogue acts. Longer dialogue contexts consistently improve alignment with the original speaker, yet extrapolative reasoning remains limited. These findings highlight that current LLMs primarily reproduce surface-level patterns rather than infer latent behavioural traits.

Contents

Chapter 1

Introduction

Recent advances in natural language processing (NLP) and machine learning (ML) have led to the emergence of large language models (LLMs) that can produce remarkably coherent, contextually appropriate, and grammatically accurate text [15] across a wide range of domains. LLMs such as GPT-4 [1] or models from the Mistral [22] and Llama family [47] are trained on large corpora embracing diverse registers of human communication. As a result, LLMs have demonstrated capabilities not only in factual retrieval and reasoning but also in generating persuasive arguments, summarising content, and engaging in multi-turn dialogue that approximates natural human conversation [5, 47]. Their capacity to condition outputs on preceding dialogue turns has made them valuable tools for applications requiring both linguistic fluency and contextual awareness.

Mimicking human behaviour in dialogue entails more than merely producing contextually relevant responses; it involves conveying a consistent style, emotional tone, and communicative intent that together signal aspects of a speaker's persona. In psychological and sociolinguistic theories of interaction, persona emerges through patterns of word choice, affect, rhetorical strategies, and decision-making cues [30, 34]. LLMs, by virtue of their training on diverse conversational data, have the potential to reproduce such stylistic and behavioural traits, thereby creating the impression of a coherent and recognisable personality throughout a dialogue [56]. This capacity is especially salient in settings where the dialogue system must adapt its responses to reflect empathy, assertiveness, or humour, aligning with human expectations of socially intelligent agents.

The ability to emulate a persona in dialogue systems has significant practical implications. In customer service applications, virtual agents that embody specific interpersonal styles, such as warmth or authority, can improve user satisfaction and foster trust [37]. In education, adaptive tutoring systems may

calibrate their language to mirror learners' engagement styles or motivational profiles [13]. Similarly, in entertainment and gaming contexts, the creation of non-player characters whose dialogue consistently reflects distinct personalities enhances immersion and narrative believability [31]. More critically, in mental health and counselling domains, conversational agents are increasingly expected to deliver sensitive support while maintaining a persona perceived as authentic and caring [46].

Computational approaches to persona modelling typically focus on either conditioning models with explicit persona profiles (e.g., pre-specified traits or goals) or fine-tuning on speaker-annotated datasets to induce characteristic styles [55]. Other lines of work investigate the prediction of dialogue acts—labels describing communicative functions like questioning, suggesting, or confirming, given a dialogue's preceding turns [51]. These methods, while effective in constrained domains, often require dedicated persona conditioning inputs or extensive supervised training with persona-labelled examples. Less explored is whether pretrained LLMs, instruction-tuned for general dialogue applications, can extrapolate human behavioral traits purely from observing a dialogue participant's prior utterances, without any explicit persona embeddings.

This thesis seeks to address this gap by investigating whether LLMs are capable of inferring and projecting aspects of human behaviour and persona from the dialogue history alone. Specifically, we ask whether models can leverage prior conversational turns to predict either the outcome of the interaction or the communicative style of the next utterance. None of the existing approaches systematically examine this extrapolative capacity in settings where the persona must be induced implicitly from the evolving dialogue context rather than supplied in advance.

To this end, two experiments that operationalise different facets of persona extrapolation were designed. The first experiment assesses whether an LLM can predict discrete decisions, such as a participant's agreement or refusal, based solely on earlier conversational cues. This experiment uses dialogue histories culminating in decision points to evaluate the model's ability to anticipate binary outcomes. In contrast, the second experiment shifts from outcome prediction to utterance generation plus multi-label classification of dialogue acts that should reflect the same choice of dialogue act than the original speaker. Here, the LLM must infer what kind of utterance (e.g., proposal, elaboration, rejection) a participant is likely to produce next, thus capturing more subtle aspects of behavioural style beyond decision-making.

The results of these studies suggest that pretrained LLMs cannot guess and express human-like personalities well enough based on dialogue context alone. By evaluating their predictive performance on both binary and multi-

label tasks, this work contributes to our understanding of the capacities and limitations of LLMs in modelling human behaviour and informs future design of conversational systems that aim to deliver more authentic, adaptive interactions.

Chapter 2

Related Work

The increasing sophistication of large language models (LLMs) has inspired significant interest in their capacity to emulate certain persona-like dialogue behaviours. These capabilities include not only the generation of syntactically fluent and contextually coherent utterances, but also the ability to simulate stable stylistic traits and decision-making tendencies of specific individuals. Such developments have positioned LLMs as potential surrogates for human participants in dialogue settings, useful in domains ranging from customer service and education to health counselling and virtual assistants.

Recent advances in persona-grounded generation systems highlight the feasibility of this endeavour. For example, Persona-L [46] combines retrieval-augmented generation with prompt conditioning to simulate user-specific dialogue behaviour. By drawing on experiential narratives and structured profiles, it improves contextual grounding and reduces response genericity. Doppelgänger LLM [8] builds a profile-conditioned model trained on survey and dialogue data to imitate individuals' verbal and decision patterns. These approaches demonstrate that, given access to structured persona data or fine-tuning pipelines, LLMs can exhibit high-fidelity imitation of target speakers.

Similarly, user simulation systems have long attempted to model dialogue agents with distinct personalities, especially for testing conversational systems in a reproducible way. Early systems relied on rule-based probabilistic models [41], whereas more recent work uses LLMs to generate synthetic users whose behaviours reflect realistic and goal-aligned persona profiles [44, 38]. These systems show that persona consistency improves perceived naturalness, user trust, and task success in dialogue interactions.

Despite these advances, most models rely on explicit persona information, either structured metadata or fine-tuned identity representations. Far fewer have investigated whether personality traits can be inferred implicitly from the dialogue history alone. This question lies at the core of the present the-

sis: Can pretrained LLMs extrapolate the communicative persona of a human participant purely from prior utterances, without additional conditioning?

2.1 Theoretical Foundations of Persona Modelling in Dialogue Systems

The ability to recognise and imitate a persona from dialogue history draws on theoretical perspectives from sociolinguistics, pragmatics, and personality psychology.

From a sociological standpoint, Goffman's theory of the "presentation of self"[12] frames identity as a social performance in which individuals manage impressions through language and behaviour. This dynamic construction of persona informs how dialogue agents might simulate role-consistent speech. Pragmatics offers further insight through Grice's Cooperative Principle[14], emphasising the role of conversational maxims (quantity, quality, relevance, and manner) in shaping interactional coherence. Variations in adherence to these norms signal pragmatic intent and stylistic traits, which persona-aware systems must recognise and reproduce.

Psycholinguistic studies extend this by linking language use with psychological profiles. Pennebaker et al. [35] and Mairesse and Walker [29] identify correlations between lexical choices and personality traits such as extraversion, conscientiousness, or neuroticism. These findings have guided computational persona modelling using trait-based classification and personality-aware language generation [23, 25].

2.2 Computational Approaches to Persona Modelling

Computational efforts in persona modelling broadly fall into two categories: explicit and implicit methods.

Explicit persona modelling includes systems such as Persona-Chat [53] that condition dialogue generation on textual profiles. These models often rely on memory networks or transformer architectures fine-tuned to replicate user facts or conversational goals. Similarly, survey-conditioned models [8] encode preferences and past responses to guide generation.

Implicit modelling, by contrast, forgoes structured inputs in favour of learning stylistic patterns directly from interactional data. Such systems may leverage contrastive training [25], speaker embeddings [45], or in-domain fine-tuning

to capture linguistic idiosyncrasies.[1] However, even these models typically require speaker-labelled data. In contrast, the current thesis explores persona extrapolation under zero-shot conditions using pretrained LLMs on unmodified dialogue histories.

2.3 Applications of Persona Imitation

Persona-aware dialogue systems have practical relevance across numerous domains. In customer service, maintaining a consistent agent persona supports branding and user rapport [39]. In educational technologies, persona-based tutors adjust communication style to match learners' preferences, enhancing engagement [46]. Mental health applications similarly benefit from empathetic and adaptive dialogue agents [32]. Beyond utility, persona consistency affects trustworthiness and user satisfaction in human-computer interaction more broadly [44, 52].

Nonetheless, the same capacity to imitate a persona has raised ethical concerns. Studies have warned of risks related to privacy leakage, misrepresentation, and consent when systems mimic identifiable communication patterns without oversight [17]. This underscores the need for transparent evaluation methods to validate such systems responsibly.

2.4 Evaluation Methods for Persona Imitation

Evaluating whether an LLM successfully mimics a specific persona requires methods that go beyond general measures of human likeness or linguistic fluency. Unlike task-oriented dialogue evaluation, where task success is paramount, persona imitation focuses on the degree to which a model maintains consistent stylistic, strategic, and cognitive patterns attributable to the original speaker. Although this thesis presents an experiment with both regular dialogues and an experiment with a set of task-specific dialogues.

Existing work has applied both qualitative and quantitative strategies to this problem. Sun et al. [46], for example, combined expert review and thematic analysis to assess whether Persona-L outputs remained faithful to a given persona profile across sessions. Cho et al. [8] employed controlled survey prediction experiments, comparing model-generated responses with actual participant answers. Their approach measured consistency and accuracy at the level of individual preferences, showing that richer conversation histories improved persona fidelity.

[1] Linguistic idiosyncrasy describes the way an individual deviates from standard language use in a way that is unique to them.

2.4.1 Qualitative Evaluation Approaches

Qualitative methods typically involve human judgment. Reviewers may assess:

- **Persona Consistency Metrics** [42]: These track whether generated outputs exhibit stable persona traits over multiple exchanges.
- **Prompt-Based Caricature Evaluation** [7]: This evaluates whether the model exaggerates or distorts a persona over time.
- **Semantic Coherence and Relevancy**: Tools such as G-EVAL [26] or SBERT similarity [49] measure contextual alignment between model and reference responses.

Although insightful, qualitative methods face reproducibility challenges, subjective variance, and limited scalability.

2.4.2 Quantitative Evaluation Approaches

Quantitative metrics offer replicable and objective assessments. These include:

- **Dialogue Act Matching**: Measuring how well the predicted dialogue act (DA) labels align with human-annotated ground truth, using precision (4.2), recall (4.3), F_1 (4.4) [9].
- **Task Success and Efficiency**: In task-oriented settings, completion rate and average number of turns remain central metrics [6].
- **Semantic Similarity Scores**: Metrics like BERTScore [54] and DialogRPT [10] evaluate response appropriateness by comparing with reference responses.

These approaches are particularly well-suited to scalable analysis in large dialogue datasets. In the context of this thesis, DA-level prediction metrics provide a tractable framework to compare LLM-generated responses with original human utterances.

2.5 Positioning of this Thesis

While previous research has demonstrated effective persona imitation using profile conditioning and fine-tuning, the capacity of off-the-shelf LLMs to infer and emulate human behaviour from dialogue history alone remains underexplored. This thesis addresses this gap by evaluating whether pretrained LLMs

can generate utterances that reflect the communicative persona of a dialogue participant, given only the preceding turns in the conversation.

Two experimental paradigms are adopted: one focusing on outcome prediction in binary decision-making scenarios and another on outcome prediction in task-oriented dialogue with multi-label dialogue act classification. By framing persona imitation as a classification task and employing standard quantitative metrics (accuracy (4.1), precision (4.2), recall (4.3), F_1 (4.4)), this study contributes a structured and replicable method to investigate emergent persona modelling capabilities in large-scale language models.

Chapter 3

Problem Statement and Research Questions

A dialogue consists of a sequence of alternating utterances, or *turns*, produced by two participants who co-construct meaning over time. In structured dialogue datasets, each of these utterances serves as a form of ground truth, capturing the authentic linguistic behaviour, communicative intent, and decision-making tendencies of individual speakers. With the recent progress in large language models, a key question arises: can these models imitate one participant's behaviour in a dialogue if provided only with the preceding conversation history?

This thesis investigates whether pretrained LLMs, without specific fine-tuning to this task or external conditioning, can emulate a human participant's utterances in regular and task-oriented dialogues. The problem is not merely one of coherence or semantic appropriateness, but of behavioural alignment—that is, whether an LLM can generate responses that are characteristic of the speaker it is meant to mimic. For example, if the original speaker used unconventional punctuation, capitalised certain expressions, or made a particular booking decision, and the LLM-generated output replicates these traits, then we consider the model to have successfully imitated the speaker's persona within the dialogue. Consequently, this study focuses on aspects such as response formatting, dialogue act classification alignment, and final decision reproduction as core indicators of success.

To investigate this, we selected two datasets of human-human dialogues, each consisting of turn-by-turn interactions between a *user* and an *assistant*. In our experimental setup, we systematically remove the utterances of one speaker, typically the *user* or persuadee role, from selected dialogue segments and prompt the LLM to regenerate those missing utterances using only the remaining dialogue context. The regenerated outputs are then evaluated against

the original, human-produced utterances (Ground Truth). This process forms the basis for measuring persona imitation fidelity.

The significance of this problem extends beyond dialogue research. If left unresolved, LLMs will continue to produce surface-level fluent responses without any behavioural consistency or fidelity to the users they are supposed to emulate, thereby limiting their utility in personalisation-sensitive domains such as adaptive education, long-term healthcare advising, or automated mediation. As Ji et al. (2023) [21] have pointed out, the inability to align LLM outputs with individual behavioural patterns poses risks for trust, interpretability, and human-AI collaboration. Budzianowski et al. (2018) [6] and See et al. (2019) [44] similarly emphasise the need for models to preserve persona coherence and decision intent across multi-turn interactions. These studies call for further research into context-sensitive generation and behavioural alignment, which this thesis directly addresses.

Research Questions This thesis explores the following three research questions, each designed to investigate a different facet of persona imitation from dialogue history:

> **RQ1:** To what extent can a large language model predict a decision of a persona inferred solely from prior dialogue turns?
>
> **RQ2:** How many preceding dialogue turns of a persona are needed for a large language model to predict the persona's decision accurately or generate persona-aligned next utterances?
>
> **RQ3:** To what extent can a large language model generate the next utterance of a persona that accurately aligns with the persona's next action in a dialogue?

These questions are addressed through a combination of dialogue truncation experiments, persona imitation analysis, and quantitative evaluations of dialogue act and decision alignment between human utterances and LLM-generated outputs.

Chapter 4

Experimental Framework

This chapter presents the experimental setup common to both experiments. It describes the selected LLMs: Mistral [22], Gemma 2 [11], and Dolphin-Llama3,[1], their parameter sizes, pre-training data, and fine-tuning regimes. It details the inference settings, including hyperparameters and prompting strategies. The evaluation metrics used to assess model performance: *accuracy*, *precision*, *recall*, F_1 *(macro/micro)*, are defined and justified. This chapter provides the methodological foundation for the two experiments that follow.

4.1 Model Selection

In order to investigate how different architectures and training methodologies impact multi–turn persona extrapolation, three characteristic open-weight LLMs were selected, spanning a range of sizes, pre-training data mixtures, and fine-tuning strategies: Mistral [22], Gemma2 [11], and Dolphin-Llama3.[1]

Mistral Mistral AI's 7 billion-parameter model was developed with high throughput and long-context efficiency in mind [22]. It was pretrained on RefinedWeb—an approximated 1.6 trillion-token corpus, using standard next-token prediction to build broad coverage of web domains. Grouped-Query Attention (GQA) accelerates inference by projecting queries in subgroups, while Sliding-Window Attention (SWA) handles inputs of up to eight thousand tokens with reduced quadratic cost. A rolling-buffer cache further curtails memory growth, and prompt-prefilling with chunked sequences enables extremely long contexts at manageable overhead.

Following pre-training, Mistral-7B-v0.1 was instruction-fine-tuned on publicly released datasets using supervised fine-tuning (SFT), yielding the Instruct-

[1] `https://huggingface.co/cognitivecomputations/dolphin-2.9-llama3-8b`

v0.1 variant available on Hugging Face.[2] On the MMLU benchmark (5-shot), Mistral-7B achieves 62.5% accuracy,[3] substantially outperforming Gemma 2 9B (34.5%) and Dolphin-2.9 Llama3 8B (19.7%). Its instruction-tuned variant further closes the gap with larger systems on MT-Bench [3], demonstrating that a compact, optimised architecture can rival much heavier models while maintaining low inference cost. This superior balance of reasoning performance, context length, and efficiency makes Mistral-7B-v0.1 an essential baseline for these persona-extrapolation experiments.

Gemma 2 Gemma 2 is Google DeepMind's 9 billion-parameter open model designed to push the boundaries of efficient, high-quality language understanding without incurring the resource costs of much larger systems. As detailed in the Gemma 2 technical report, the architecture builds upon standard Transformer blocks but incorporates two key modifications: interleaved local–global attention (inspired by Beltagy et al., 2020 [4]) to better capture both short- and long-range dependencies, and group-query attention (GQA) [2] to halve the compute cost of self-attention while preserving performance.

Unlike conventional next-token prediction, Gemma 2's 2B and 9B variants are pre-trained via knowledge distillation from a much larger teacher model, enabling them to effectively "see" over 50× more data than would otherwise be computationally optimal [19] and thereby approach the capabilities of models two to three times their size.

The 9B model itself is trained on approximately 8 trillion primarily English tokens, striking a balance between scale and accessibility for deployment on modest hardware.

In comprehensive benchmark evaluations, Gemma 2 9B consistently outperforms its peers in the same size class. On the MMLU suite, it achieves a score that rivals or exceeds that of Llama 3 8B[4], and on BigBench Hard, its reasoning accuracy places it among the top open models in its category.

Moreover, the instruction-tuned variant, Gemma-2-9B-it, demonstrates class-leading performance on MT-Bench, surpassing many 13B and even 34B competitors, validating its instruction-following prowess in multi-turn dialogue and complex reasoning tasks. Gemma 2's public release, accompanied by an in-depth report and code, makes it an ideal candidate for preliminary evaluation in any setting where model size and inference speed are at a premium.

[2] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1

[3] https://paperswithcode.com/sota/multi-task-language-understanding-on-mmlu?utm_source=chatgpt.com

[4] https://blog.google/technology/developers/google-gemma-2/

Dolphin-Llama3 Dolphin-2.9 Llama 3 8B is an instruction-tuned descendant of Meta's LLaMA 3 8B, curated and released by Eric Hartford, Lucas Atkins, Fernando Fernandes and the Cognitive Computations team at the Ollama library [5].

The model was fine-tuned on a blend of ShareGPT[6] and UltraChat[7] datasets to endow it with versatile instruction-following, conversational, and coding abilities. Notably, Dolphin-2.9 also introduces initial agentic capabilities and supports function-calling out of the box, allowing it to autonomously interact with external tools or APIs during generation.

Unlike many contemporary LLMs, Dolphin is deliberately kept uncensored; its training pipeline filters out alignment-specific data to maximise compliance with arbitrary prompts. This "raw" behaviour profile provides a unique testbed for probing how unconstrained models extrapolate user personas and biases from dialogue context [16].

Beyond its skill suite, Dolphin-2.9 offers a standard 4,000-token context window and an extended 256,000-token variant, enabling it to maintain coherence over exceptionally long dialogues. In scenarios where persona inference depends on deep, multi-turn context—such as tracking a participant's traits across a novel-length conversation—this extended context capability is invaluable. Although formal benchmark results (e.g., Massive Multitask Language Understanding (MMLU) or BigBench Hard) are not publicly reported for Dolphin, empirical evaluations have shown it matches or exceeds the performance of LLaMA 3's base instruct variant on both general instruction and coding benchmarks. Its combination of high context capacity, tool-enabled interactivity, and unfiltered generation makes Dolphin-2.9 an ideal candidate for this thesis's preliminary evaluation of persona extrapolation in extended conversational settings.

Table 4.1 gives an overview of all the used models, comparing parameter size pre-training and fine-tuning data, as well as fine-tuning methods.

4.2 Setup

LLMs are asked to generate human utterances, while making binary decisions and reacting accordingly to given dialogue personas. The generated outcomes are then evaluated according to human judgements and choices. The hyperparameters are adjusted to ensure that the generated results are appropriate for the task. The next section discusses the hyperparameters that are adjusted

[5] `https://ollama.com/library/dolphin-llama3`

[6] `https://sharegpt.com/`

[7] `https://huggingface.co/datasets/stingning/ultrachat`

Metrics	Mistral-7B-v0.1	Gemma 2 9B	Dolphin-2.9 Llama3 8B
Parameters	7.25 billion	9 billion	8 billion
Release	2023	2024	2024
Pre-training Data	RefinedWeb (approx. 1.6 T tokens)	Google web crawl (approx. 8 T tokens, proprietary)	Meta LLaMA 3 corpora (proprietary)
Fine-tuning Data	Publicly released instruction datasets	Teacher-distilled outputs, public instruction corpora	ShareGPT, UltraChat
Fine-tuning Methods	SFT	Knowledge distillation, SFT	SFT, function-calling support

Table 4.1: Comparison of evaluated Language Models. SFT = Supervised Fine-Tuning.

during the experiments and the process of developing prompts.

4.2.1 Hyperparameter Configuration

Hyperparameters are the external settings that control the language model's behavior during inference. Below are some hyperparameters that are manually modified and influence different aspects of the generated output.

Top_k We employ `top_k = 1`, i.e. greedy decoding, to ensure that at each generation step the model selects the single most likely token. Although greedy decoding is known to limit diversity and can introduce repetition in truly open-ended generation [20], it offers two critical benefits for our experiments. First, it maximises output confidence and determinism, producing identical continuations for identical inputs; this repeatability is essential when quantitatively comparing persona inference accuracy or numeric prediction errors across Mistral-7B, Gemma-2-9B, and Dolphin-2.9 Llama3-8B. Second, by removing sampling randomness, we isolate each model's intrinsic decision-making behavior, rather than conflating it with stochastic variation, thereby improving the statistical power of our evaluation metrics. In sum, while greedy decoding

may underproduce creative diversity, it is the preferred choice for controlled, reproducible experiments focused on model-driven persona extrapolation and numerical reasoning.

Temperature The temperature parameter scales the logits of the model before softmax, controlling the randomness of the sample. At `temperature = 1.0`, logits are used as is, preserving the probability distribution learned by the model. Increasing the temperature (> 1.0) flattens this distribution, increasing diversity but risking nonsense results; lowering the temperature (< 1.0) makes the distribution sharper, making the model more conservative and prone to safe, repeatable markers [36]. A value of 1.0 is often chosen as a neutral compromise that preserves both fluency and some diversity, especially when subsequent tasks (e.g., persona extrapolation) require naturalness without excessive noise.

Maximum Generation Length The `num_predict` parameter controls the maximum number of tokens that the model can produce per generation call. This constraint balances the provision of sufficiently long, coherent continuations and limits computation and latency: longer sequences incur quadratic attention costs and an increased risk of output drift [55]. In this thesis, a length of `num_predict = 128` was chosen to capture complete thought segments without overwhelming subsequent post-processing or the person reading the text.

Context Window Size The `num_ctx` parameter determines how many preceding tokens the model will take into account when predicting each following token, i.e. the context window size. The original Transformer architecture was demonstrated with a context size of up to 512 tokens [48], but modern LLMs typically extend this value (e.g., GPT-3's 2,048). A window of `num_ctx = 2,048` tokens provides a practical balance: it sufficiently captures multi-turn conversation histories or document contexts for consistent persona inference, while keeping the attention cost $O(n^2)$ within GPU/TPU memory limits.

4.2.2 Prompting

In this thesis, prompting plays a critical role as a control interface between the experimental setup and the generative behaviour of the large language models. The main task involves predicting human utterances within a dialogue by replacing a real speaker with a model while keeping the dialogue context unchanged. Since models are inherently tuned to act as helpful assistants, the

risk of breaking the speaker's role, especially when mimicking a human rather than an agent, is not trivial. A precisely engineered prompt is necessary to suppress assistant-like tendencies and induce a persona-consistent reply that aligns with the conversational style of the user, not the system.

Thus, the prompts in this thesis are not simply a formality in generating answers but a constructive constraint vital to role validity and approaching personality extrapolation. The degree to which LLMs are successful in this environment depends not only on their architecture and training but also on how well the cues promote appropriate speaker role selection and contain hallucinations within acceptable limits.

4.3 Evaluation Metrics

The baseline evaluation is essential to establish a reference point for future experiments and enhancements.

Accuracy One of the basic measurements for determining a model's output quality is accuracy. It tells what percentage of the model's predictions out of the total predictions are correct. It is calculated as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{4.1}$$

where:
TP - True Positive,
FP - False Positive,
TN - True Negative,
FN - False Negative.

Precision Precision is the proportion of all the model's positive classifications that are positive in the ground truth. It is mathematically defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.2}$$

Recall Recall calculates how many examples were correctly predicted as positive out of all the actual true values. The recall is also called sensitivity or True positive rate (TPR). Recall always focuses on the actual positives. We use recall whenever the false negative result is important.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.3}$$

Precision improves as false positives decrease, while recall improves when false negatives decrease.

F_1 Score F_1 is a harmonic mean between precision and recall, and we can use the F_1 score when we do not know whether FP is important or FN is important in a certain problem.

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.4}$$

As the F_1 score is more sensitive to data distribution, it's a suitable measure for classification problems on imbalanced datasets. Different variations of the F_1 score exist, such as weighted, macro, and micro F_1, which can be more appropriate for multi-class problems or when one wants to weight the classes differently.

Macro F_1 The macro approach calculates F_1 scores for each class separately and then averages them. This approach assumes that all classes are equally important, but in practice, this is not always the case. The formula for the macro F_1 score is as follows:

$$\text{Macro } F_1 = \frac{1}{N} \sum_{i=1}^{N} F_1^i \tag{4.5}$$

where:

$$F_1^i = 2 \cdot \frac{\text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}$$

Support_i is the number of true instances of class i,
N represents the number of classes in the classification task.

Micro F_1 In the micro approach, the contributions of all classes are summed to calculate the average F_1 score. This method is useful when there is an imbalance of classes in the dataset; smaller classes are given the same weight as larger classes. The formula for the micro F_1 score can be derived as follows:

$$\text{Micro } F_1 = 2 \cdot \frac{\text{Precision}_{\text{micro}} \cdot \text{Recall}_{\text{micro}}}{\text{Precision}_{\text{micro}} + \text{Recall}_{\text{micro}}} \tag{4.6}$$

where

$$\text{Precision}_{\text{micro}} = \frac{\sum_{i=1}^{N} \text{True Positives}_i}{\sum_{i=1}^{N} (\text{True Positives}_i + \text{False Positives}_i)}$$

and

$$\text{Recall}_{\text{micro}} = \frac{\sum_{i=1}^{N} \text{True Positives}_i}{\sum_{i=1}^{N} (\text{True Positives}_i + \text{False Negatives}_i)}$$

This method is useful when the overall classification performance is more important than the performance of individual classes.

Thus, these metrics together provide a measure of model performance under both balanced and unbalanced conditions. Accuracy provides a rough overview of overall correctness, while precision and recall separate the costs of false positives and false negatives. The F_1 score combines these two dimensions into a single score, with its macro and micro versions adapting to multi-class and class imbalance scenarios. By reporting all of these values together, we not only ensure comparability with previous work but also gain a deeper understanding of where each model succeeds or fails, whether it is in terms of common or rare classes or prioritising safety (minimising false positives) over sensitivity (minimising false negatives).

Chapter 5

Experiment I: Decision Making

The first experiment explores the capability of three large language models: Mistral [22], Gemma 2 [11], and Dolphin-Llama3[1] to predict binary decisions (Yes / No) regarding donation behaviour in response to dialogues about charity [50]. The primary objective is to assess whether language models, given prior dialogue history, can predict an individual's decision to donate and the potential donation amount. The experiment simulates persuasive dialogue scenarios where the language model plays the role of a persuadee, engaging with a persuader. The assistant must decide whether to donate to a charity based on the conversation history and indicate how much they would donate. The research evaluates each model's capability for sentiment understanding, decision-making, and numerical prediction based on context.

5.1 Dataset

The data used in this study originates from the 'Persuasion For Good Corpus' [50], a richly annotated dataset of human-human dialogues designed to study persuasive strategies in donation contexts. The dialogues focus on charitable contributions to 'Save the Children',[2] an international non-profit organisation, and were collected through structured online interactions using the ParlAI platform[3] on Amazon Mechanical Turk.[4]

The corpus consists of 1,017 dialogues involving 1,285 unique participants, with each participant randomly assigned the role of either *persuader* or *persuadee*. Table 5.1 provides a view to one of the dialogues from the corpus. Conversations were conducted via text and involved real monetary stakes:

[1] `https://huggingface.co/cognitivecomputations/dolphin-2.9-llama3-8b`
[2] `https://www.savethechildren.org/`
[3] `https://parl.ai/`
[4] `https://www.mturk.com/`

Table 5.1: An example dialogue from the "Persuasion for good" dataset with the persuadee's decision to make a donation - a so-called "positive" dialogue.

Turn	Context	Speaker
0	Hello how are you today?	Persuader
...	...	...
12	You can donate the $0.30 form this hit every penny helps.	Persuader
13	Okay, sounds good. If you don't mind, I'll donate half, so how about $0.15?	User
14	Okay that is great. We appreciate your donation. Knowing that you are willing to help says a lot.	Persuader
...	...	...
19	Hey thanks. You too!	User

participants could choose to donate part of their task earnings to the charity. Conversations were required to have at least 10 turns, but the average was slightly higher at 10.43 turns per dialogue, with a mean utterance length of 19.36 words. The vocabulary used across the dataset spanned 8,141 unique tokens.

Each dialogue was labelled with a unique identifier, and both participants' roles were explicitly annotated. The dataset also records whether a donation occurred and the amount donated. On average, participants chose to donate $0.35, drawn from their payment.

Participant-level statistics reveal a notable asymmetry in linguistic behaviour: persuaders averaged 22.96 words per utterance, whereas persuadees averaged 15.65 words. Donation behaviour was similarly skewed, with 42% of persuaders and 54% of persuadees choosing to donate. These behavioural and linguistic asymmetries provide a rich basis for modelling persuasion dynamics.

The preprocessing involved arranging the user's and persuader's utterances into the required prompt structure for the LLM. During this stage, the data was divided into two equal sets of 125 dialog histories: positive and negative dialogues.

> *Positive dialogues* are dialogues, where the outcome of the dialogue is a donation of a known sum of dollars.
>
> *Negative dialogues* are dialogues in which the persuadee declines to donate.

These dialogues were extracted and stored in CSV files to facilitate easy access and use in subsequent analysis stages. The model was provided with the

context history of each dialogue, and the expected outcome (whether the assistant would decide to donate or not) was the primary label for prediction. Apart from this binary decision, the model was asked to tell how much money it wants to donate in each case.

5.1.1 Dialogue Context Segmentation

Each dialogue is divided into several turns, where participants either play the role of the *Persuader* (assistant) or the *Persuadee* (user). The statements of the persuader and the persuadee alternate in turn. The dialogue data was preprocessed to make it suitable for input into the LLM. From 1,017 dialogues of varying quality [50], 250 dialogues of more than 20 utterances in length and with a clearly made decision were manually selected.

Among the selected dialogues, it was found that the average number of utterances where the decision was made is thirteen. That is, before turn thirteen in the whole dialogue, the person being asked, most often gave his/her specific answer about the decision to donate.

Each dialogue history was then truncated to seven, eleven, and all given turns of a conversation for a given experiment. That way, using eleven utterances in total means that the dialogue is truncated one user utterance before the average decision-making turn. The seventh utterance was a below-average marker, and a complete dialogue history - all turns, was needed to benchmark the model's ability to extract (rather than predict) decision information from the dialogue. Equally, truncating all the dialogues was necessary to keep the length of the input data constant and to ensure that the model can process the entire dialogue within its context window.

5.2 Experimental Results and Analysis

For each dialogue, the model generated a binary response based on the conversation context. The models were instructed to answer with either "Yes" (indicating a decision to donate) or "No" (indicating a decision not to donate).

Beyond predicting binary outcomes, the models were also tasked with predicting a numerical donation amount. The models were expected to interpret the dialogue and, if they predicted a donation, include a dollar amount in their response. For dialogues where a donation was predicted, the numerical amount was extracted from the model response and recorded into a CSV result file for later analysis.

Each of the 250 cases was manually read to choose the dialogues during pre-processing, allowing for a human assessment of how closely the responses of

the large-language models matched the ground truth. For the study, dialogues with a definite persuasion of decision were chosen. The test file additionally contained the most intriguing cases with exceptional dialogue participant behaviour.

Frequently, the models would identify current conversational topics and react appropriately. In some cases, the model imitated the user's hesitation and took a long time to determine how much to donate when the user was truly undecided about whether or not to make a donation to the fund. Additionally, there were a few notable instances where the persuadee was adamantly opposed to children and thus refused to help them. The individual in this conversation also frequently used foul language, and all three LLMs were able to replicate this situation.

Despite the instructions, the models experienced hallucinations, and such responses were then considered "undecided" or "NA" - no answer. Hallucinations came in all shapes and sizes. For example, generating not only the user's last reply, but also the next part of the conversation. Or the model generated completely irrelevant responses that did not correspond to the context, but worked with some subtopic of the dialogue. Here, an example could have been *"love for children"* instead of *"decision to donate money to a children's fund"*. Sometimes the models simply did not give a specific solution in one utterance. All responses similar to these situations were considered "NA".

Since the three chosen LLMs have different underlying architectures and training datasets, the study aimed to find out if LLMs are, in general, able to predict dialogue history outcomes or not.

5.2.1 Binary Classification Metrics

An initial comparative analysis of accuracy (4.1), precision (4.2), and recall (4.3) revealed a consistent relationship between dialogue length and prediction reliability. Mistral demonstrated high levels of recall in all contexts, reaching 99% when evaluated with full dialogue contexts, although its accuracy remained lower in shorter contexts (Figure 5.1). For example, with only seven dialogue turns, Mistral's recall was 91.8% and accuracy was 56.2%, indicating a pronounced tendency to classify undecided or negative instances as positive.

In contrast, Gemma 2 showed the most balanced performance across dialogue lengths. It achieved an accuracy of 89.4% and a recall of 97.0% in the "all turns" scenario, while maintaining more conservative metrics at earlier turns: 64.0% accuracy and 84.0% recall at turn 7 (Figure 5.2). Dolphin-Llama3 showed comparable results, slightly outperforming Gemma 2 in overall accuracy (91.2%) in the "all turns" condition, but exhibited reduced recall in the short context (68.5%) (Figure 5.3).

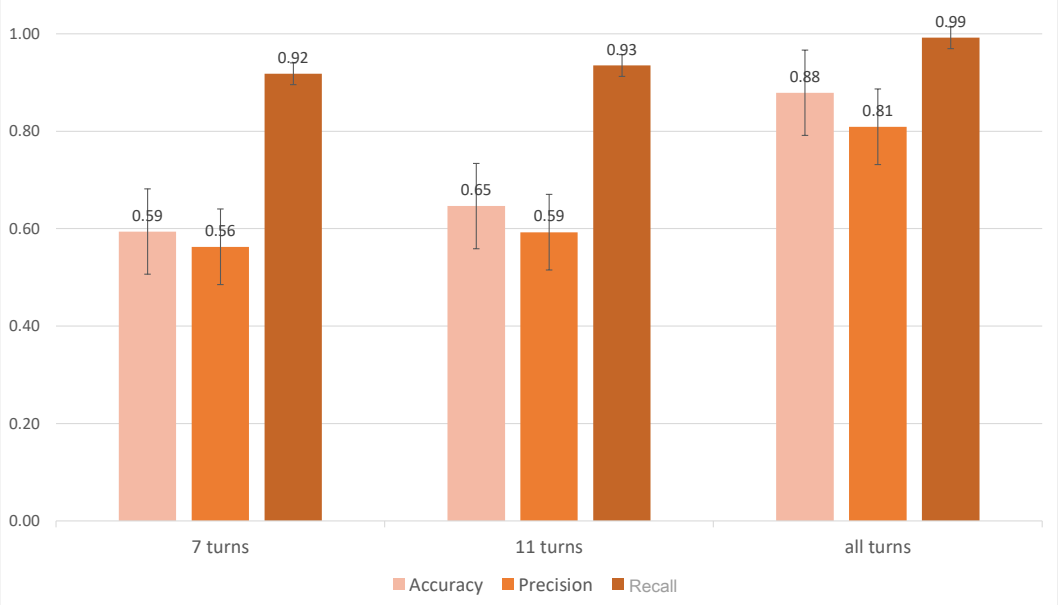

Figure 5.1: Accuracy, precision, and recall for binary decisions made by Mistral.

A particularly noteworthy trend across all three models is that recall is consistently higher than precision—across all dialogue lengths and model configurations. This pattern suggests that models more frequently opt to predict a donation will occur ("yes"), even in cases where the correct label is "no". For instance, Mistral at 7 turns shows a recall of 91.8% but a precision of only 59.0%, and Gemma 2 shows 84.0% recall vs. 58.0% precision (Figures 5.1, 5.2). While this behaviour might be interpreted as a form of charitable optimism, it raises concerns about the calibration of LLMs in moral or value-laden decision-making tasks, where overcommitment may be inappropriate or misleading.

The recall–precision gap is noticeably smaller for Dolphin-Llama3 (recall = 68.5%, precision = 63.0%), suggesting a somewhat more restrained decision-making profile under low-information conditions (see Figure 5.3). One potential explanation for this reduced affirmative bias may lie in its relatively uncensored architecture. As Dolphin-Llama3 lacks extensive guardrailings, its predictions may be less skewed by ethical alignment fine-tuning and thus more conservative or uncertain in ambiguous contexts.

Importantly, the comparison between dialogue conditions shows that while performance improves as more dialogue turns are introduced, the accuracy before the final decision (e.g., after 7 or 11 turns) remains only marginally better than chance. Specifically, the accuracies at 7 turns are: Mistral (56.2%), Gemma 2 (64.0%), and Dolphin-Llama3 (61.0%) seen in Figures 5.1-5.3. These

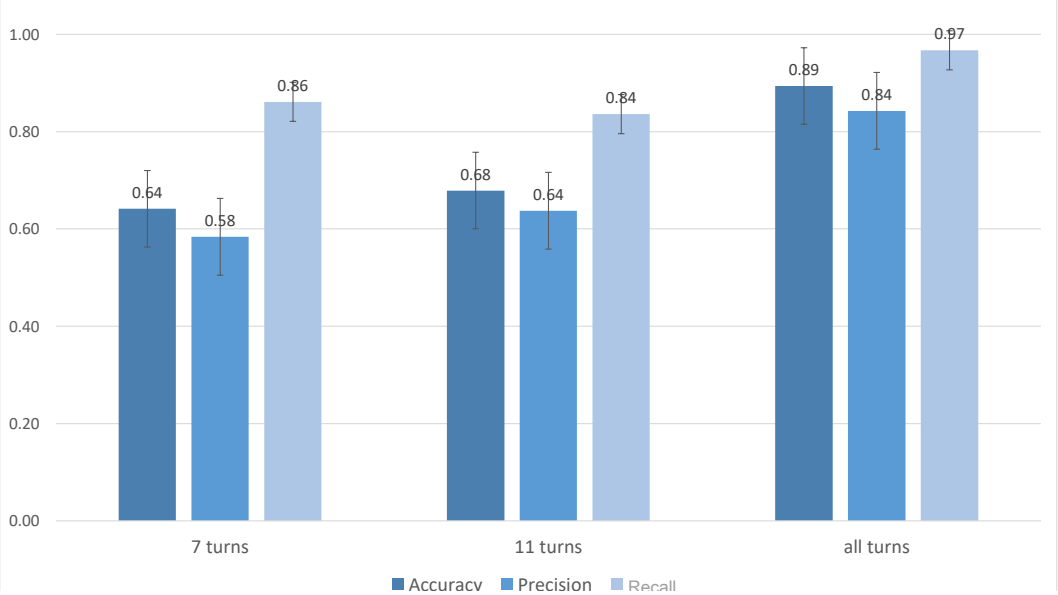

Figure 5.2: Accuracy, precision, and recall for binary decisions made by Gemma 2.

numbers suggest that, absent an explicit decision statement, the models struggle to extrapolate the user's final stance from prior conversational context. This strongly undermines the hypothesis that LLMs are capable of inferring latent decision intent in these dialogues.

Thus, while "all turns" settings showcase each model's capacity to correctly recover and reproduce known outcomes, they do not convincingly demonstrate extrapolative reasoning. In essence, the task becomes one of memory and paraphrasing rather than inference. The critical evidence lies in the performance prior to the decision utterance: all three models behave more like random guessers with a positive bias, rather than sophisticated agents capable of reading between the lines.

5.2.2 Distribution of Generated Decisions

Analysis of the generated decision distributions showed systematic deviations from the balanced truth (50% yes, 50% no). All three models consistently overproduced affirmative responses. The Mistral model, especially in the seven-turn condition, produced almost 199 'yes' predictions out of 250, highlighting its tendency to positively classify when context is lacking (see Figure 5.4).

Gemma 2, although also biased towards 'yes', produced a more tempered distribution: 149 "yes" and 63 "no" responses, with 38 marked as NA (Figure 5.5). This reflects a more stable decoding mechanism and slightly lower affirmative bias.

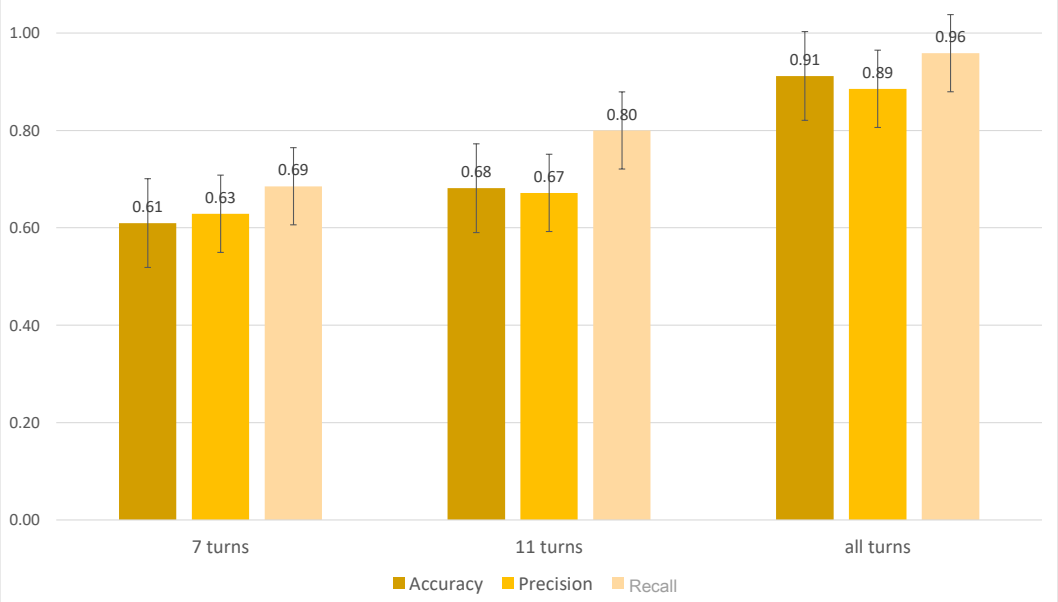

Figure 5.3: Accuracy, precision, and recall for binary decisions made by Dolphin-llama3.

Dolphin-Llama3 displayed the most conservative judgment and the highest uncertainty: 97 "yes", 67 "no", and a significant 86 NA responses at 7 turns (see Figure 5.6). This reinforces the earlier interpretation that its lower recall but higher precision stems from a greater reluctance to commit to a decision without sufficient evidence, an arguably more cautious and realistic behaviour under information scarcity.

Overall, these distributions align with the earlier metric analysis: Mistral is heavily affirmative, Gemma 2 is moderate, and Dolphin-Llama3 is the most hesitant. The balance of yes/no predictions becomes more even as dialogue length increases, but initial decision distributions are systematically skewed.

5.2.3 Analysis of Predicted Donation Amounts

The quantitative discrepancies between predicted and actual donation amounts were very large, confirming the assumption that even with improved classification accuracy, numerical estimates are still prone to overestimation. Mistral, in particular, consistently overpredicted donation amounts in early context windows. More often than not, Mistral would reply, *'Yes, $50,'* which meant a *'positive'* decision and a willingness to donate $50 to the fund. Such answers were too frequent in the case of seven dialogue turns, and it is likely that such a decision is related to the model's pre-training data. In the seven-turn

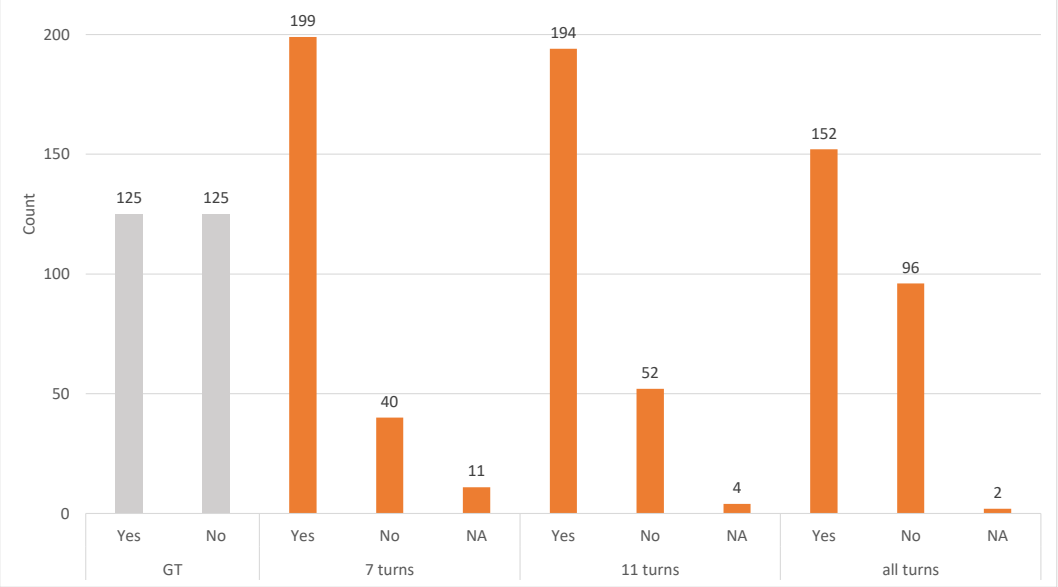

Figure 5.4: Mistral generated "Yes", "No" and hallucinated (NA) answers compared to Ground Truth (GT).

condition, prior to the explicit decision being made, the model's total predicted donation reached $5,544.20, nearly seven times the ground truth value of $805.36 (see graph in Fig. 5.7). This inflation aligns with the model's strong tendency to predict affirmative donation decisions (199 "yes" vs. 40 "no" in the same condition) (as shown in Fig. 5.1).

Gemma 2 produced a more calibrated numerical response across all conditions. In the full-dialogue setting, its predicted total ($788.49) closely approximated the ground truth, and even in shorter contexts, its overestimations were relatively moderate (e.g., $2,330.40 at 7 turns) (see graph in Fig. 5.7). Dolphin-Llama3 showed the most conservative estimation behavior. Despite high rates of indecision (86 NA responses at 7 turns), its predicted donation amounts remained below those of the other two models, culminating in a final total of $537.92 in the full-dialogue case.

Importantly, these numerical outcomes must be interpreted alongside the earlier classification results. The observed overestimation in donation amounts, especially in early turns, is not merely a result of random numerical error. Rather, it reflects a systematic consequence of the affirmative bias in the models' binary classifications. Since donation values are only predicted when the model decides to donate, an overproduction of "yes" responses inherently inflates the cumulative amount. This is particularly evident in Mistral's behaviour: the combination of high recall and low precision (recall = 91.8%, precision = 59.0% at 7 turns) suggests a high false-positive rate, which di-

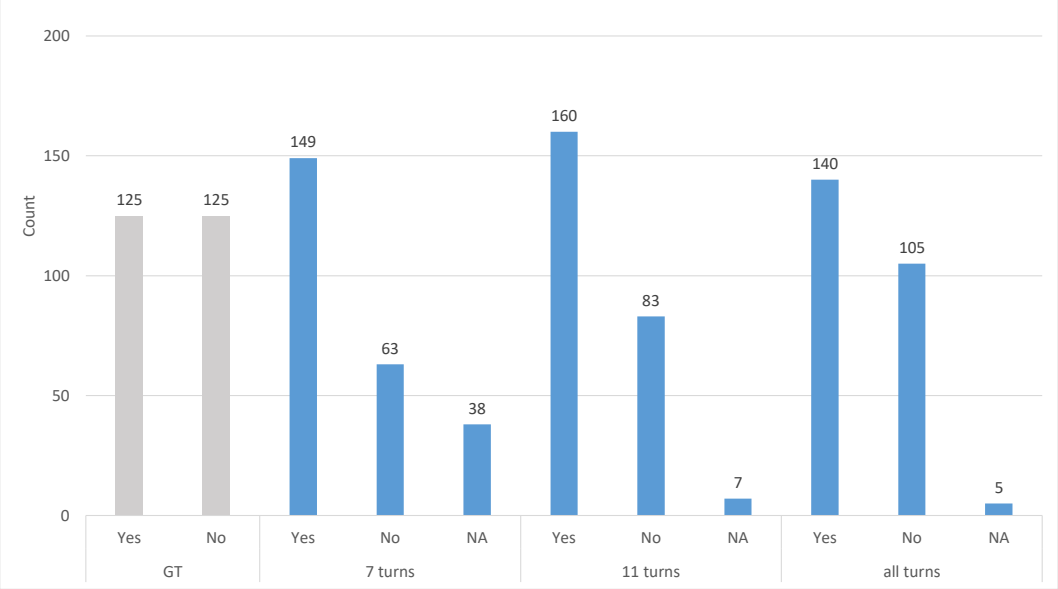

Figure 5.5: Gemma 2 generated "Yes", "No" and hallucinated (NA) answers compared to Ground Truth (GT).

rectly translates into unrealistically high donation sums.

These patterns further support the interpretation that LLMs do not extrapolate nuanced decision intent from partial context but rather rely on priors or optimism biases when information is incomplete. The inflation in monetary predictions, especially in the absence of explicit decision turns, offers an additional signal that models substitute context-sensitive reasoning with a general affirmative default.

The bar graph in Figure 5.7 illustrates this effect: donation predictions consistently converge toward the ground truth as more dialogue turns are provided. Yet prior to the actual decision utterance, all models, especially Mistral, grossly overestimate, underscoring the risk of relying on LLMs for decision forecasting in ambiguous scenarios.

5.2.4 Comparative Discussion

The comparative evaluation of Mistral, Gemma 2, and Dolphin-Llama3 yields several important insights into the behaviour of current large language models (LLMs) in the task of predicting binary donation decisions and corresponding amounts. Despite the shared prompt structure, deterministic decoding (top-k = 1), and identical test inputs, model behaviour varied substantially across metrics and dialogue lengths.

A consistent and central finding is that dialogue length strongly correlates

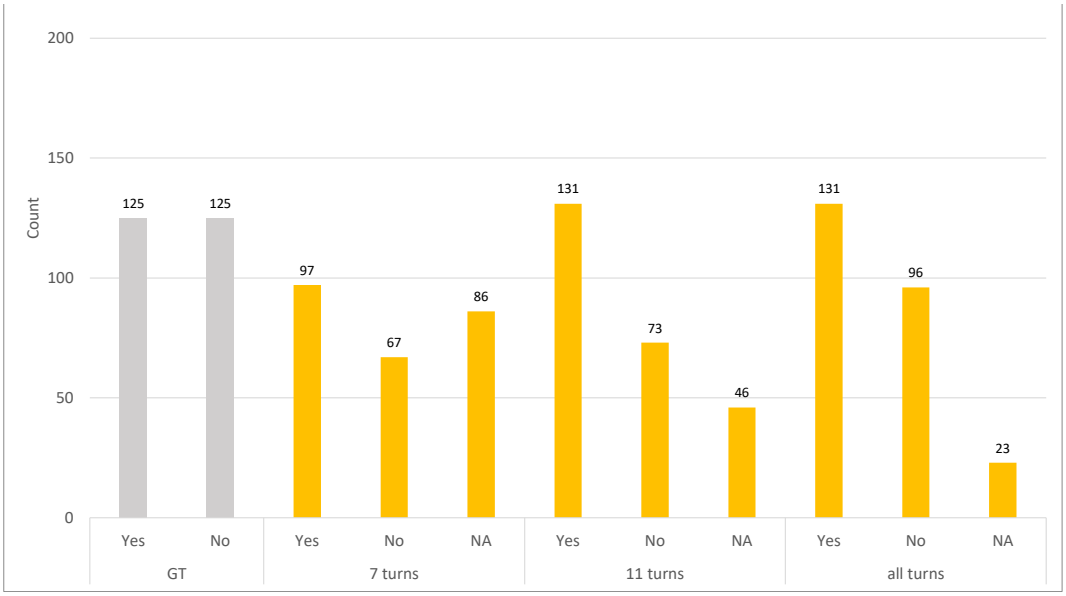

Figure 5.6: Dolphin-llama3 generated "Yes", "No" and hallucinated (NA) answers compared to Ground Truth (GT).

with prediction quality. Across all three models, accuracy and recall improved markedly as additional conversational context was provided. However, this improvement primarily reflects the inclusion of the actual decision utterance in the "all turns" setting. Once the decision is revealed in the dialogue, the classification task becomes one of surface-level reproduction rather than extrapolation. This is evidenced by near-perfect recall and high accuracy in the "all turns" condition (e.g., Mistral recall = 99%, accuracy = 88%), contrasting sharply with near-chance accuracy at 7 turns (Mistral: 56.2%, Gemma 2: 64.0%, Dolphin-Llama3: 61.0%) seen in Figures 5.1-5.3. These results underline the critical limitation: none of the models demonstrates reliable extrapolative capacity when the decision has not yet been stated.

In terms of classification behaviour, all models display a systematic bias toward affirmative ("yes") decisions. This is most extreme in Mistral, which generated 199 "yes" responses out of 250 examples in the 7-turn setting (Figure 5.4). This bias is also reflected in the consistently higher recall than precision observed across all dialogue lengths and models. The fact that the recall–precision gap is largest for Mistral and smallest for Dolphin-Llama3 (e.g., recall = 68.5%, precision = 63.0%) suggests differences in decision conservatism and guardrailing. Dolphin-Llama3, being uncensored and less alignment-tuned, tends to produce more uncertain (NA) responses and fewer false positives—a pattern suggesting greater caution in low-context conditions, albeit at the cost of coverage (Figure 5.6).

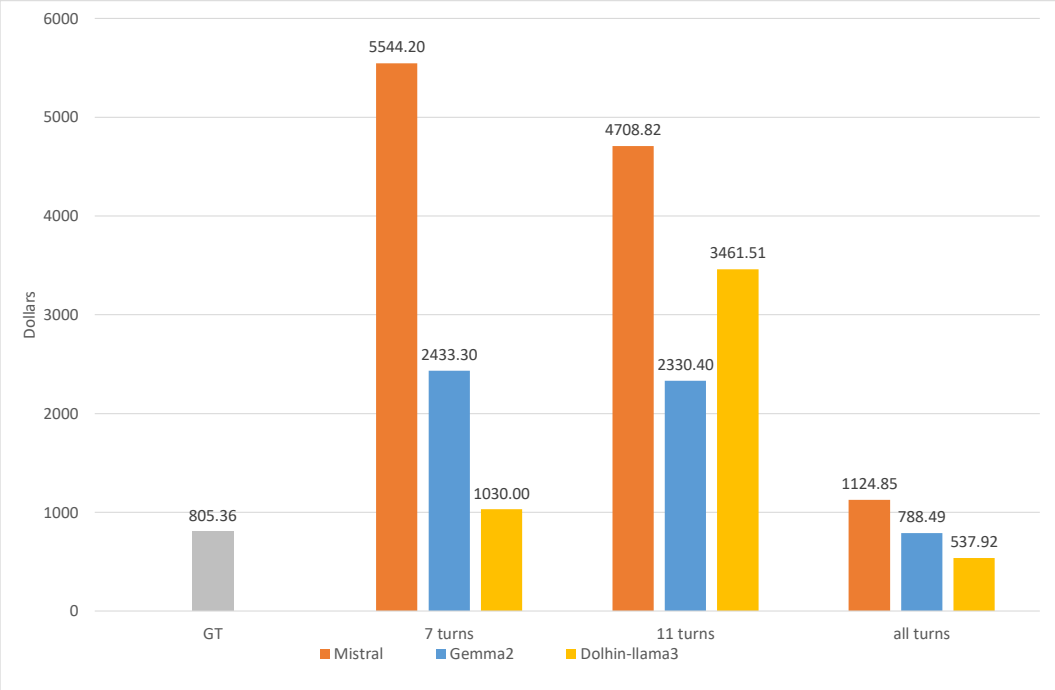

Figure 5.7: Sum of all LLM-generated donations compared to Ground Truth (GT).

Gemma 2 consistently struck a middle ground. Its classification behaviour was more balanced in terms of "yes"/"no" distributions, and it maintained stable performance across metrics. Notably, it produced some of the most accurate donation amount estimates, including a near-match to the ground truth in the all-turns condition ($788.49 vs. $805.36) in Figure 5.7. Mistral, by contrast, massively overestimated donation amounts in low-context settings (e.g., $5,544.20 at 7 turns), reinforcing the interpretation that its high recall comes at the expense of calibration. Since donation amounts are only generated when a "yes" decision is made, affirmative bias directly inflates total donation sums, revealing the tangible consequences of model imprecision.

Despite differences in magnitude, all three models demonstrate a convergence pattern: as dialogue length increases, classification metrics improve, NA outputs decrease, and donation predictions align more closely with the ground truth. However, this convergence is driven more by the presence of explicit signals than by latent inference ability. The models do not extrapolate unstated intent effectively; rather, they react to explicit cues already available in the full dialogue.

In summary:

- Mistral is confident and aggressive in predicting "yes," resulting in high recall and heavily inflated donation estimates in early turns.

- Gemma 2 offers the most balanced performance overall, with relatively good accuracy, modest overestimation, and robust donation value predictions.
- Dolphin-Llama3 is conservative and less biased, with higher NA rates and the smallest recall–precision gap, potentially due to the absence of restrictive fine-tuning.

These findings call into question the assumption that current LLMs are capable of modelling latent user intent based on incomplete dialogue. While they perform well in reproducing decisions after the fact, their extrapolative capabilities in real-time prediction remain weak. This has important implications for the deployment of LLMs in social forecasting, recommendation systems, or decision support applications, where inference from incomplete or early-stage interactions is often essential.

Chapter 6

Experiment II: User Behaviour

Experiment II was performed on the Frames dataset [9]—a multi-label collection of human-human dialogues where one of two participants is acting as a *Wizard-of-Oz* [40] assistant. The goal of the experiment was to anticipate the last or middle human response using large language models, taking the histories of the dialogues into account. For this purpose, firstly, the data preparation was performed; secondly, the dialogues were optimally divided into a training and a test dataset; next, a human response generator was developed using LLMs; and finally, since we could not find a suitable classifier for dialogue acts used by the authors of the dataset, we developed and trained our classifier for 20 dialogue acts (see Table 6.2), to classify the generated utterances.

6.1 Dataset

In this experiment, the Frames dataset [9] is employed as the primary source of conversational data. The Frames corpus was designed to capture complex decision-making and comparison behaviours within a goal-oriented travel domain. Dialogues were collected in a controlled Wizard-of-Oz setting [40] over 20 days, involving 12 rotating participants who alternated daily between the roles of "user" and "wizard" [9].

A total of 1,369 task-oriented human–human dialogues was recorded, yielding 19,986 turns (an average of approximately 14.6 turns per dialogue). Each turn corresponds to a single Slack message, and turns strictly alternate between user and wizard.

Turn- and Act-Level Statistics

- **Dialogue lengths** ranged from 3 to 43 turns, with a median of 14 and a mean of 14.6 turns per dialogue.

Table 6.1: An example dialogue from the "Frames" dataset.

Turn	Context	Speaker	Dialogue Act
0	Hi I'd like to go to Caprica from Busan, between Sunday August 21, 2016 and Wednesday August 31, 2016	User	*inform*, *greeting*
1	And what would be your maximum budget for this trip?	Wizard	*request*
2	Actually it's unlimited for this trip	User	*inform*
...	...	...	...
10	That sounds great. 1:00 am return on Sunday August 28th is very early in the morning...	User	*inform*
11	Would you like to book this package?	Wizard	*suggest*
12	Yes I would. Thanks for your help.	User	*affirm*, *inform*, *thankyou*

- **Annotation rate**: 75% of turns carry exactly one dialogue-act label, while 25% carry two or more labels; unannotated turns are rare (e.g. when a user's request cannot be fulfilled by the database).
- **Dialogue acts**: 20 distinct act types were used, including *inform*, *request*, *offer*, *request_compare* and *switch_frame* (see Table 6.2). The most frequent act, *inform*, occurs approximately 10,200 times; *offer* and *request* appear roughly 4,700 and 8,000 times, respectively (as stated in Figure 2 in [9]).

Table 6.1 illustrates a representative excerpt from the Frames corpus. In each dialogue, turns strictly alternate between user and wizard:

- **User utterances** typically express high-level booking goals (e.g., *'I'd like to book a trip to Atlantis from Caprica on August 13, 2016 for 8 adults, with a budget of $1 700.'*), or they register requests for more information (e.g., *'Can you tell me which of these resorts offers free Wi-Fi?'*).
- **Wizard utterances** describe available packages retrieved from the database, propose suggestions when no direct match exists (e.g., *'Unfortunately, we have no trips matching those constraints. Would you like a slightly higher budget?'*), or offer comparative summaries (e.g., *'The cheapest available flight to New York is $1 947.14; if you're flexible on dates, I can show you other options.'*).

Table 6.2: List of dialogue acts in the annotation of "Frames" dataset

Dialogue Act	Speaker	Description
inform	User/Wizard	Inform a slot value
offer	Wizard	Offer a package to the user
request	User/Wizard	Ask for the value of a particular slot
switch_frame	User	Switch to a frame
suggest	Wizard	Suggest a slot value or package that does not match the user's constraints
no_result	Wizard	Tell the user that the database returned no results
thankyou	User/Wizard	Thank the other speaker
sorry	Wizard	Apologize to the user
greeting	User/Wizard	Greet the other speaker
affirm	User/Wizard	Affirm something said by the other speaker
negate	User/Wizard	Negate something said by the other speaker
confirm	User/Wizard	Ask the other speaker to confirm a given slot value
moreinfo	User	Ask for more information on a given set of results
goodbye	User/Wizard	Say goodbye to the other speaker
request_alts	User	Ask for other possibilities
request_compare	User	Ask the wizard to compare packages
hearmore	Wizard	Ask the user if she'd like to hear more about a given package
you_are_welcome	Wizard	Tell the user she is welcome
canthelp	Wizard	Tell the user you cannot answer her request
reject	Wizard	Tell the user you did not understand what she meant

In order to leverage the Frames corpus for both multi-label dialogue-act classification and LLM–driven response generation, a systematic data-preparation pipeline was designed. This pipeline encompasses three main phases:

1. Parsing the original JSON dialogues into a clean tabular format;
2. Applying 5-fold split with stratified label distributions ensured through iterative stratification to derive an optimal 80%/20% train–test partition;
3. Truncating the held-out (test) dialogues to serve as a context base for LLM response generation. The procedure is described in 6.1.2

With a clean, tabular CSV in hand, the next step was to partition the data in a way that preserves its multi-label characteristics.

6.1.1 Cross-Validation and Iterative Stratification

Standard random partitioning often fails to preserve the delicate balance of matching labels and multi-label assignment that occurs in a multi-label environment. To evaluate classification performance and to avoid potential sampling bias in a multi-label setting, the cleaned dialogues were partitioned into training and test sets via a 5-fold split combined with Iterative Stratification [43], which has been shown to preserve label co-occurrences more faithfully than standard stratified sampling. Conventional stratified partitioning methods preserve the marginal frequencies of individual labels but often fail to preserve the complex co-occurrence patterns characteristic of multi-label data. Iterative stratification eliminates this limitation by distributing both individual labels and their co-occurrences across folds in an iterative manner, thereby ensuring that each layer approximates the overall distribution of labels. The implementation in the skmultilearn package (IterativeStratification) was used to generate five candidate 80/20 splits on the set of final user turns.

For each candidate split, the average absolute difference in per-label frequency between the train and test sets was computed over the 20 DA labels. The fold with the smallest average disparity was chosen, yielding:

> **A training dataset**, consisting of 1101 dialogues (approximately 80% of 1369)
>
> And **a test dataset**, consisting of 268 held-out dialogues (approximately 20%)

Figure 6.1 illustrates the label counts in the chosen fold with the well ballanced proportions. This exact split was further used to train a custom classifier for dialogue acts and to generate LLM-based user responses from the test dataset. The developed dialogue act classifier was used to prove that the responses generated by the three large language models truly mimic the personas in these specific dialogues. This was achieved by comparing the ground truth labels (from the unmodified user responses), the reclassified original user turns, and the classified generated answers.

Having obtained an optimal train–test split, the test dialogues were then restructured for response-generation experiments.

6.1.2 Truncating Test Dialogues

While the selected 20% test dataset serves as a reliable set for evaluating multi-label classification, the part of the experiment related to response generation requires a slight modification, in which each deferred dialogue ends with a statement from the assistant, in this case, a wizard. To explore the capabilities

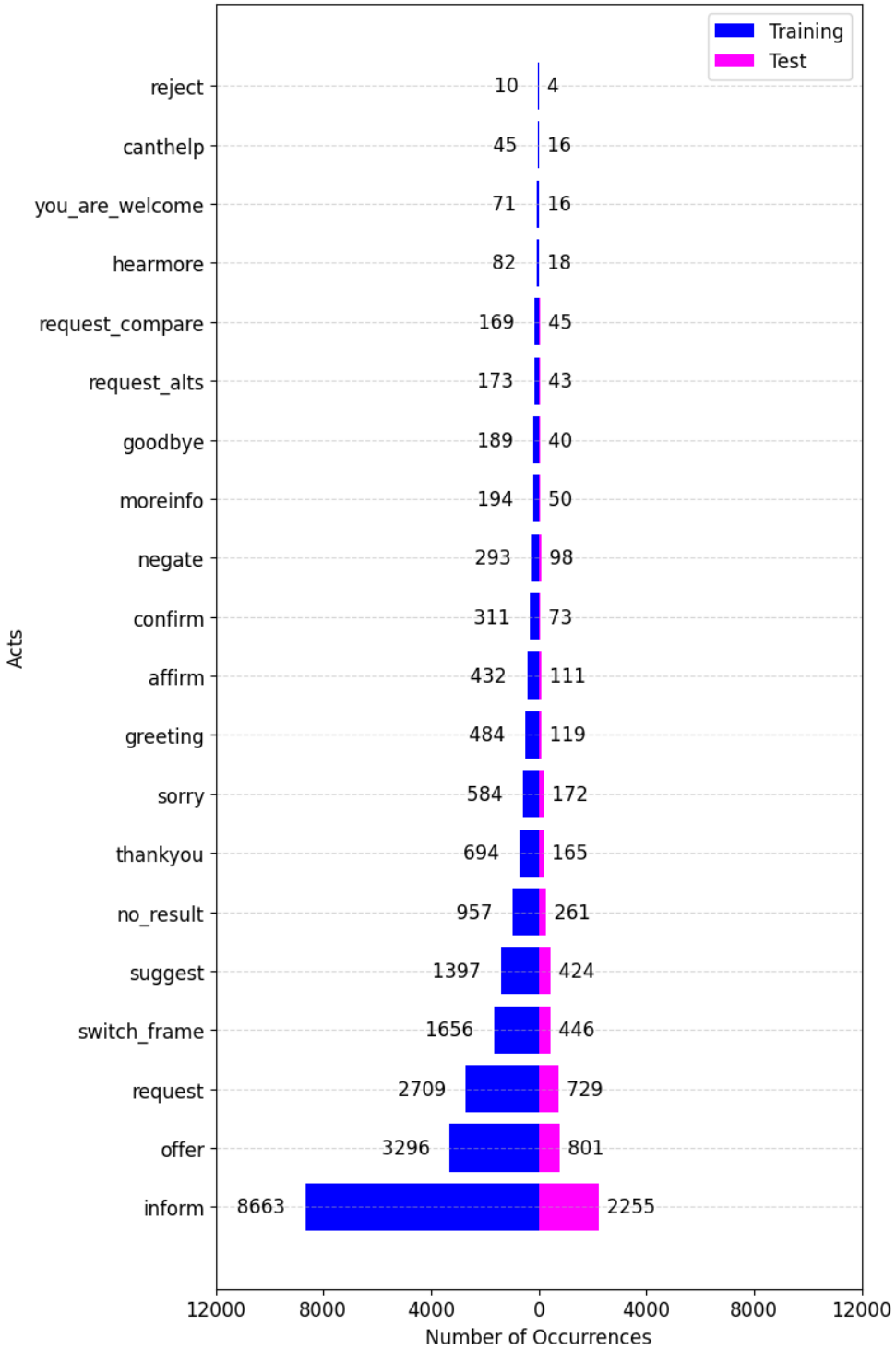

Figure 6.1: A balanced distribution of Frames DS dialogues with multi-label annotation into train and test files.

of language models not only at the end of the context but also in the middle, the dialogues were carefully truncated to their first 6 turns (3 user–assistant exchanges) or first 10 turns (5 user–assistant exchanges), necessarily ending with the wizard's responses, to generate user responses on top of them. Any dialogue shorter than 6 turns was discarded to ensure that there were at least 3 pairs of 'user-assistant' for context. In each remaining dialogue, the last user turn was removed, resulting in a set of test dialogues ready for user utterance generation. Consequently, three versions of the file were created for generation: with 6 turns, with 10, and with entire dialogues. Each dialogue in these files ends with an assistant line, providing exactly the context needed to generate the missing user response based on LLM.

6.2 Evaluation Procedure

In order to assess both the fidelity of dialogue-act prediction and the capacity of large language models to emulate dialogue personas' responses, two complementary systems were developed: an LLM-based user-utterance generator, and a custom RoBERTa classifier for multi-label dialogue acts.

6.2.1 User-Turn Generation

Essentially, the generator processes each dialogue history, consisting of all the utterances involved, as a single contextual string, to which a default system prompt is added, defining the user's task in the conversation. This prompt suppresses initial assistant framing and anchors the model to the user's perspective, encouraging behaviour reconstruction based solely on context.

```
Each dialogue reflects an interaction between a travel agent and
a customer discussing flights, hotels, or other aspects of a trip.
Your task is to be in the role of the customer, for example, to
find the best option for your trip.  Using the dialogue history,
finish the dialogue with one line as you see fit on behalf of
the customer (user).  Derive the result.
```

6.2.2 RoBERTa based Classifier

Because no ready-made classifier was available that supported the twenty dialogue-act labels of interest, a bespoke classifier based on RoBERTa [27] was developed. The overall workflow comprised three stages: converting the dialogue CSV into a training dataset, fine-tuning a pre-trained RoBERTa encoder for multi-label classification, and packaging the resulting model for inference.

The training data were drawn from the filtered, stratified partition of the Frames corpus (see Figure 6.1). Each row contains a user or a wizard utterance plus its associated list of dialogue acts. These lists were transformed into binary indicator vectors using a standard `MultiLabelBinarizer` (from `scikit-learn` [33]), ensuring that every example is represented by a fixed-length label vector.

A `RobertaForSequenceClassification` head was attached to the base RoBERTa encoder. The classifier's output dimension was set to the number of dialogue acts (20). During training, a binary cross-entropy loss with logits was computed against each example's multi-label vector, permitting multiple acts to be predicted per utterance.

Fine-tuning was orchestrated via the Hugging Face `Trainer` API. Batches of tokenized utterances were processed on a GPU, using `AdamW` [28] optimisation and a linear warm-up schedule. Mixed-precision training was enabled to accelerate convergence. Evaluation on the held-out fold was performed at each epoch, from 1 to 30, to monitor macro-F_1 (4.5) performance and to prevent over-fitting.

We fine-tuned a RoBERTa-based multi-label classifier for dialogue-act tagging over 7,440 gradient steps, logging training and validation metrics every 248 steps. During the first 1,000 steps, training loss plunged from about 0.30 to 0.10 and training F_1 rose from 0.14 to roughly 0.40, after which loss declined more slowly to approximately 0.005 by the final checkpoint and F_1 asymptotically approached 0.93 (see Figures A.8, A.9). Validation loss mirrored this trajectory, dropping from 0.13 to 0.08 and then back up to 0.12, while validation F_1 climbed steadily to around 0.58 (Figures A.10 and A.11). The consistent proximity of training and validation curves throughout suggests minimal overfitting and indicates that most gains occur before 4,000 steps, beyond which returns diminish. Taken together, these dynamics demonstrate that our fine-tuning regimen yields rapid convergence, strong generalisation (mid-50% F_1 on the test data), and a clear window for early stopping to optimise compute without sacrificing performance. In a case with multi-label classifications with 20 labels, the result of $F_1 = 0.58$ is sufficient enough.

By leveraging a pre-trained language model and a straightforward multi-label head, supplemented with balanced training data from our optimal split, we obtained a robust predictor whose performance underpins our subsequent analyses of both human and generated utterances.

6.3 Experimental Results and Analysis

For each dialogue at each level: 6, 10, and all utterances, the three models generated the latest deleted utterance from the User. The answers consisted

of one or more sentences, usually more than one. As stated in the prompt, the models generated only the response itself, without the corresponding dialogue act labels. A developed classifier was used to classify the generated dialogue responses as well as reclassify the ground truth utterances from the original dialogues.

Across the created bar charts, which isolate LLM-generated user turns labeled by RoBERTa classifier results (green colour in the following Figures: A.1 - A.7, 6.2 and 6.3), a clear pattern emerges: all models disproportionately produce the *inform* and *switch_frame* acts at the expense of social or closing signals, but Mistral exhibits the most balanced behaviour. For Gemma 2, *inform* counts exceed the yellow (GT) distribution by nearly 200% at 6 turns and climb further with 10 turns or full contexts, while *switch_frame* also skyrockets and *thankyou*, *goodbye*, and *affirm* vanish entirely. Dolphin–LLaMA3 follows a similar, if slightly tempered, trend, with marginally lower peaks in both *inform* and *switch_frame*. In contrast, Mistral not only reaches the highest *inform* over-generation (+253% at ten turns) but also preserves modest volumes of closures (*thankyou* approx. 22 occurrences, *goodbye* approx. 3, *affirm* approx. 3) even in the 6-turn setting. Moreover, increasing context from 6 to 10 turns yields a substantial jump in nuanced acts, most notably *suggest* and courteous labels, whereas extending to the full dialogue yields only marginal further gains. Taken together, these findings indicate that Mistral, when provided with approximately ten preceding turns and more, produces the most contextually appropriate and persona-consistent user replies, striking the best balance between necessary informational content and natural conversational closure.

This section presents a detailed comparison between the original Frames dialogues, the classifier's re-labelling of those same turns, and the labels assigned to LLM-generated user replies. We focus in particular on two illustrative cases: (1) Gemma 2 generation after truncating each dialogue to 6 turns (Figure 6.2), and (2) Mistral generation on full dialogues (Figure 6.3). These cases correspond respectively to the lowest and highest observed classification performance in our macro- and micro-F_1 results (4.5, 4.6) (Figure 6.4 and 6.5).

In Figure 6.2, the bars show the distribution of the ten most frequent acts in the held-out test set. The orange bars reveal that our RoBERTa classifier often under-predicts these acts even on yellow user turns: for instance, *thankyou*, *no_result*, *sorry*, and *goodbye* are almost entirely missed (drops of over 90%), while *inform* is over-labelled by 169%. When applied to Gemma 2's generated replies (green), its labels *inform* and *switch_frame* are at dramatically inflated rates (+239% and +828% respectively), while core acts such as *goodbye* and *you_are_welcome* vanish completely. This severe distortion of the original act distribution corresponds directly to Gemma 2's lowest macro-F_1 (0.59)

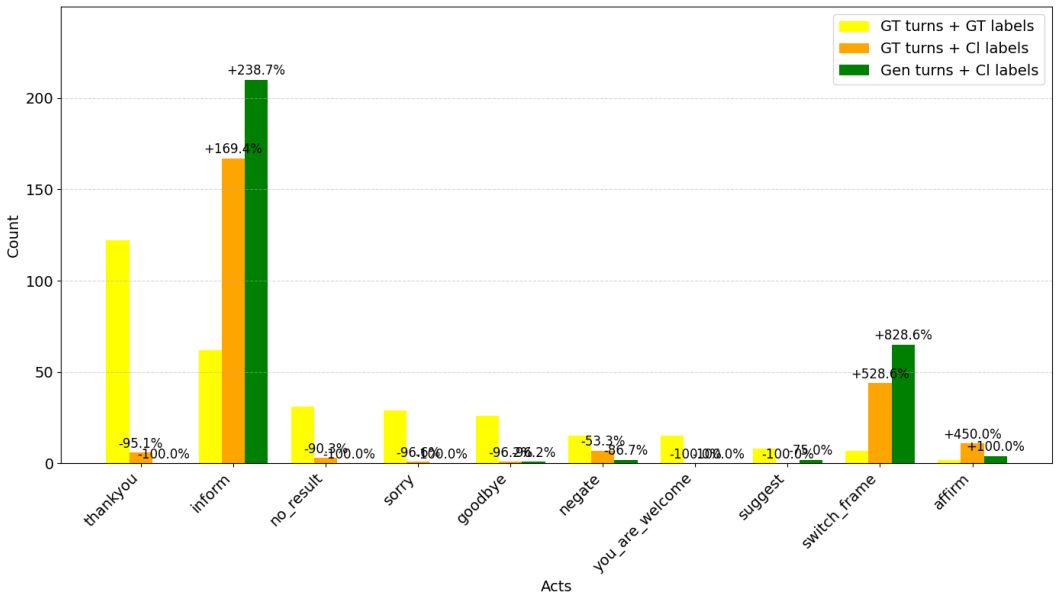

Figure 6.2: Comparison of dialogue act labels **after 6 turns**: yellow - original utterances (ground truth) and corresponding original labels from the corpus (ground truth); orange - original removed utterances from the corpus (ground truth) and corresponding labels classified by the classifier; and green - utterances generated by **Gemma 2** and corresponding labels classified by the classifier.

and micro-F_1 (0.70) scores among all settings (see Figures 6.4 and 6.5).

By contrast, Figure 6.3 shows that Mistral's generated turns yield a closer match to the original distribution. The classifier's recollection on yellow turns (orange) deviates modestly (e.g. *thankyou* –5.7%, *goodbye* –11.5%), while on generated turns (green) the largest deviations are confined to secondary acts (*inform* +234%, *suggest* +413%). Importantly, fundamental acts like *thankyou*, *no_result*, and *sorry* remain present in non-zero counts. This fidelity is reflected in Mistral's leading macro-F_1 of 0.66 and micro-F_1 of 0.77 (see Figures 6.4 and 6.5).

Figures 6.4 and 6.5 summarise generation performance across all models (Mistral, Gemma 2, Dolphin-llama3) and truncation lengths (6, 10, all utterances). Both metrics increase monotonically with longer context windows, with Mistral consistently outperforming the alternatives. The gap between Gemma 2 and Mistral is largest at 6 turns (macro: 0.59 vs. 0.63; micro: 0.70 vs. 0.74), underscoring the impact of context length on dialogue turns generation.

The sharp contrast between Gemma 2 at 6 turns and Mistral with full context illustrates two key findings:

1. insufficient dialogue history leads to severely distorted act distributions,

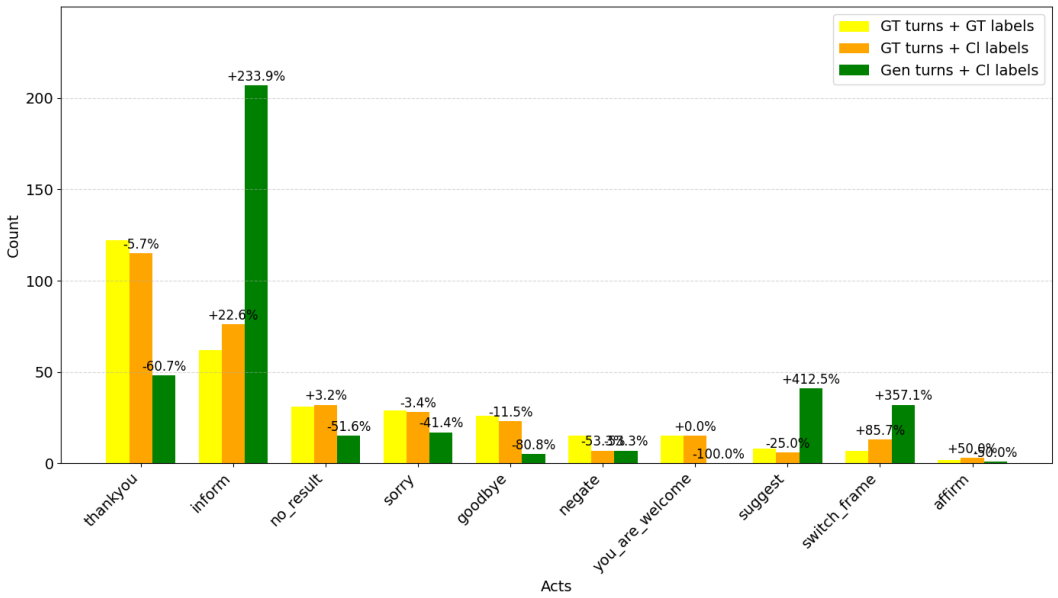

Figure 6.3: Comparison of dialogue act labels **after all turns**: yellow - original utterances (ground truth) and corresponding original labels from the corpus (ground truth); orange - original removed utterances from the corpus (ground truth) and corresponding labels classified by the classifier; and green - utterances generated by **Mistral** and corresponding labels classified by the classifier.

even when using a pre-trained LLM;

2. and giving an LLM full access to prior turns substantially improves its ability to generate utterances that preserve the functional roles seen in the original Frames corpus.

These results validate our hybrid evaluation pipeline, combining RoBERTa classification of generated text with the traditional F_1 metric, as an effective probe of persona-mimicking in task-oriented dialogues.

Analysis of the Dolphin–LLaMA3 series (Figures A.3, A.4, A.5) reveals consistent under-prediction of courtesy acts in generated replies. For example, even on full contexts, *thankyou*, *sorry*, and *goodbye* counts drop by over 55% compared to ground truth, indicating that LLMs de-prioritised these social signals. Conversely, the *switch_frame* act inflates by +843% (6-turn) to +328% (all-turns), highlighting the model's tendency to propose new decision tracks rather than finalise a choice. This mirrors observations that LLMs are mainly functioning to inform, suggest solutions and "switch frames" to follow the conversation.

Moreover, across all LLMs and truncation lengths, certain acts, *greeting*, *suggest*, *you_are_welcome*, *offer*, and *request_alts* remain absent in both

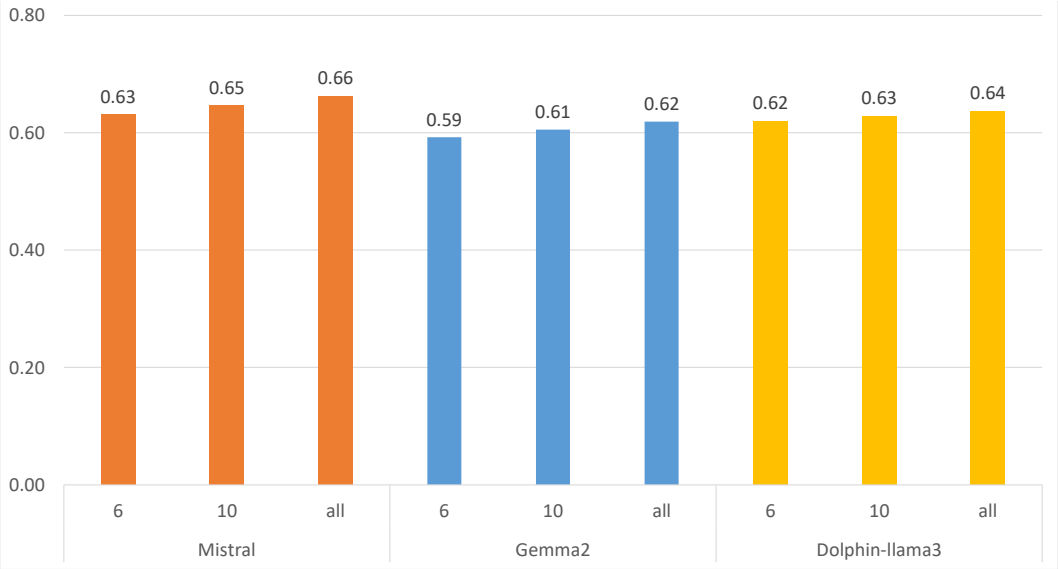

Figure 6.4: Macro F_1 for generated results of all LLMs after 6, 10 and all utterances.

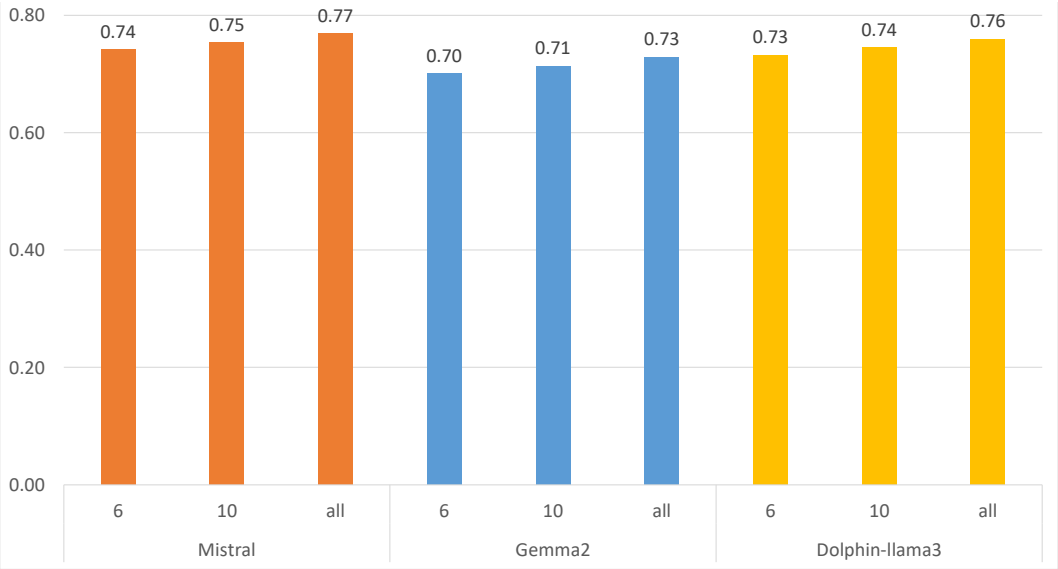

Figure 6.5: Micro F_1 for generated results of all LLMs after 6, 10 and all utterances.

ground-truth and generated user replies, reflecting the fact that users naturally do not perform these functions in this task-oriented domain [9]. However, the generated responses often contain *suggest* — an action reserved for the 'wizard' role. A large number of "suggestions" were found in the generated LLM responses, so the generated 'user' begins to behave a bit like a 'wizard'.

Macro- and micro-F_1 trends (Figures 6.4 and 6.5) show that even truncated

contexts (6 turns) allow for reasonable utterance generation (F_1 up to 0.63/0.74 for Mistral), indicating that LLMs can infer the final user turn structure with moderate accuracy given limited history. However, the poorest performance of Gemma 2 at 6 turns (F_1 0.59/0.70) demonstrates that model capability and context length jointly drive prediction quality. Comparing 6- vs. 10-turn splits across models, F_1 increases by 3–5 points on average, suggesting that approx. 8–10 turns are required before asymptotic gains are observed. Beyond 10 turns, improvements taper off, implying that additional context yields diminishing returns for dialogue turns generation.

The inflated *switch_frame* and *inform* acts in generated replies indicate that LLMs excel at providing information and proposing alternative frames, yet they struggle to produce acts denoting commitment (e.g. *affirm*, *goodbye*). Consequently, although LLMs reliably mimic the informational behaviour of users, and even exceed the norm in producing informative sentences as responses, they are less capable of accurately generating a suitable final statement about decision-making that signifies final confirmation of a reservation. Such expressions would be ideal in "all turns" scenarios.

Chapter 7

Discussion

This chapter discusses the key findings of the two experiments conducted in this thesis and places them in the context of existing research on large language models for multi-turn dialogue tasks. The discussion is structured into three parts. Section 7.1 analyses the factors influencing model effectiveness across both experiments. Section 7.2 explores recurrent failure patterns and their implications for understanding LLM behaviour. Section 7.3 integrates findings from both experiments to derive broader conclusions about persona extrapolation and decision prediction in dialogue systems.

7.1 Factors Influencing Model Effectiveness

Both experiments confirm that dialogue history length is a decisive factor in improving LLM performance. In Experiment I, Dolphin-Llama3 achieved an accuracy of 91.2% when provided with full context, compared to only 61.0% under a 7-turn constraint. Similarly, Mistral's recall increased from 91.8% to 99.1% with the inclusion of all prior turns. These results clearly show that deeper context allows models to better capture decision-relevant cues and user intent.

Experiment II, which studied user-turn generation in travel-booking dialogues from the Frames dataset, reinforces this conclusion. Both macro-F_1 and micro-F_1 scores improved consistently as more context was provided. For example, Mistral's macro-F_1 increased from 0.63 to 0.66, and Dolphin-Llama3's micro-F_1 rose from 0.73 to 0.76. The application of iterative stratification ensured balanced label distributions across folds, reducing variance and increasing robustness of evaluation.

Model size also played a crucial role. The three evaluated models: Mistral-7B-v0.1 (7.25B parameters), Gemma 2 (9B parameters), and Dolphin-2.9-llama (8B parameters), differ not only in size but also in their pre-training

and fine-tuning regimes (Table 4.1). These differences contributed to the observed variations in performance.

Larger models, such as Gemma2 and Dolphin-Llama3, generally achieved higher accuracy and recall in Experiment I when sufficient dialogue history was provided. Gemma 2, the largest model, achieved a balanced trade-off between precision and recall in the all-turns condition (accuracy 89.4%, recall 96.7%). Dolphin-Llama3, fine-tuned on ShareGPT and UltraChat data, achieved the highest overall accuracy (91.2%) with full context. Mistral, despite being the smallest model, delivered comparatively balanced dialogue-act distributions in Experiment II and retained some social and closing acts even with shorter contexts.

However, larger models also amplified biases. Gemma 2 frequently overproduced *inform* and *switch_frame* acts in Experiment II, with *inform* counts exceeding ground truth by nearly 200% at six turns. Mistral, with its smaller size and less aggressive instruction tuning, displayed the most balanced DA outputs despite having the highest relative *inform* overgeneration (+253% at 10 turns).

These findings align with scaling-law literature, showing that increasing parameter count improves general predictive power but does not inherently reduce biases or improve alignment. In Experiment I, larger models achieved higher recall and accuracy with extended context, but in Experiment II, they tended to amplify dominant patterns from training data at the expense of socially appropriate or closing acts. Effective fine-tuning and calibration remain necessary to achieve balanced, contextually faithful outputs.

Prompting strategy further influenced outcomes. Structured prompts that explicitly framed the model's task (e.g., decision-focused generation in Experiment I and task-specific framing in Experiment II) led to more contextually coherent outputs. This aligns with evidence that LLMs rely heavily on task framing to access relevant internal representations.

7.2 Prediction Failures and Their Causes

Despite the benefits of larger context and structured prompting, both experiments revealed systematic prediction failures.

In Experiment I, all three models exhibited a consistent bias towards affirmative donation predictions, overproducing "yes" responses relative to the ground truth distribution (50% yes / 50% no). Mistral was the most prone to this bias, producing 199 affirmative responses out of 250 in the 7-turn setting. Moreover, its predicted donation sums were heavily inflated, with a total of $5,544.20 compared to the ground truth of $805.36. These errors suggest

that models rely excessively on surface-level linguistic markers of agreement or empathy, mistaking them for commitment signals in the absence of explicit decision cues.

Experiment II revealed analogous issues in structured prediction and DA classification. Across the bar charts isolating LLM-generated user turns labeled by the RoBERTa classifier (green bars in Figures A.1–A.7, 6.2 and 6.3), a distinct pattern emerges: all models disproportionately produce *inform* and *switch_frame* acts while failing to generate socially oriented or closing acts such as *thankyou*, *goodbye*, and *affirm*. All models struggle with context-sensitive dialogue act allocation, particularly for compound or nuanced acts that depend on pragmatic inference. These results show that while LLMs are excellent at generating fluent utterances, they still cannot reliably map conversational context to discrete decision outcomes or structured denotations. Their reliance on shallow heuristics leads to systematic errors when important indicators of intent are implicitly expressed.

7.3 Cross-Experiment Reflections

Comparing the two experiments reveals convergent trends and complementary insights. Both emphasise the importance of rich conversational context, demonstrating that dialogue history provides important cues for resolving ambiguity. However, both also show that larger context alone is insufficient for high-fidelity predictions, model biases persist, and outputs remain prone to hallucinations and misclassifications.

Experiment I underscores that generative LLMs tend to overcommit to positive decisions and overestimate numerical values when context is insufficiently constraining. Experiment II complements this by revealing that zero-shot LLMs underperform on structured tasks requiring precise label attribution, further supporting the need for fine-tuning and calibration.

The combination of these findings suggests that persona extrapolation and decision prediction in dialogues require models that integrate both generative flexibility and structured reasoning. Task-specific fine-tuning, calibration techniques, and possibly hybrid architectures combining symbolic reasoning with LLM generation appear necessary to achieve robust performance in any conversational setting.

Chapter 8

Conclusion

The conclusion summarises the key findings, explicitly answers the three research questions, and discusses the implications of the results. It outlines the main limitations of the study, such as dataset constraints and reliance on automated evaluation metrics, before identifying open research questions. The chapter concludes by suggesting future directions, including task-specific fine-tuning, hybrid architectures that combine structured reasoning with generative models, and human-in-the-loop evaluation for assessing persona alignment and naturalness of generated utterances.

This thesis investigated the ability of large language models (LLMs) to predict users' utterances in multi-turn dialogues, knowing only their context, with a focus on exhibited personality traits in the generated utterances. Two experiments were conducted to address the research questions: Experiment I examined LLMs' capacity to generate binary donation decisions and corresponding donation amounts in persuasive dialogues. In contrast, Experiment II assessed the models' ability to produce responses by simulating human utterances on the travel booking task in a dialogue with labelled dialogue acts. The research questions guiding this work were:

> **RQ1:** To what extent can a large language model predict a decision of a persona inferred solely from prior dialogue turns?
>
> **RQ2:** How many preceding dialogue turns of a persona are needed for a large language model to predict the persona's decision accurately or generate persona-aligned next utterances?
>
> **RQ3:** To what extent can a large language model generate the next utterance of a persona that accurately aligns with the persona's next action in a dialogue?

8.1 Key Findings

The findings of Experiment I demonstrate that LLMs can generate plausible persona-aligned utterances when provided with sufficient dialogue history. Accuracy and recall improved markedly with longer context; for example, Dolphin-Llama3 achieved an accuracy of 91.2% with full dialogue context compared to only 61.0% with seven turns. However, all models exhibited a consistent affirmative bias in donation predictions and tended to overestimate donation amounts, especially in shorter contexts. These results address **RQ1** by showing that while LLMs can partially predict a persona's decision based on prior dialogue turns, their outputs are strongly biased and overfitted to affirmative outcomes, reflecting their training as helpful assistants rather than neutral predictors.

Regarding **RQ2**, both experiments emphasise the importance of dialogue history. The largest gains in accuracy and contextual relevance with both datasets were observed when the context was increased from 6-7 to 10-11 antecedent turns, and only a slight improvement was observed when further extended to full dialogue. This is since both the 6-7 and 10-11 conversational turns typically ended before the dialogue decision in both datasets. But 10-11 utterances statistically end just before the persona's committed decision (Ground Truth) in the dialogue, and in the case of "all turns", the decision is already known to the LLM because it is included in the context. From this, we can conclude that the approximately 10-11 preceding turns provide sufficient context for LLMs to produce more coherent and reflexive utterances concerning the persona in the situations in "Persuasion for Good" and "Frames" datasets. Mistral, in particular, generated the most balanced results with this length of context. This model had the largest gap in the coherence of generated utterances with the persona's responses between 6-7 and the other two context truncation methods in both experiments.

The results of the experiments give us reason to believe that the person participating in the dialogue should already be on the strut to the decision made, so that the LLM replacing him or her in that dialogue would be able to respond with as much resemblance to that person as possible.

RQ3 examined whether LLMs can generate final-turn utterances that accurately reflect the persona's decisive action. Experiment I showed that, although LLMs frequently produced syntactically plausible decision statements, their semantic correspondence to the ground truth remained limited. Dolphin-Llama3 displayed the highest number of irrelevant outputs (up to 86 of 250 cases in the 7-turn condition), while Mistral achieved the highest recall for affirmative outcomes but exhibited inflated affirmative predictions. Gemma 2 proved to be the most balanced, reaching an accuracy of 89.4% with full con-

text and producing more realistic donation predictions. Experiment II further demonstrated that zero-shot LLMs underperformed on structured dialogue-act generation (micro-$F_1 < 0.76$), reinforcing that accurate generation of final-turn utterances, particularly those reflecting explicit decisions, remains challenging without task-specific fine-tuning.

8.2 Limitations and Open Questions

Several limitations affect the interpretation of these findings. First, the experiments evaluated only three LLMs (Mistral, Gemma 2, and Dolphin-Llama3) under specific prompting configurations. Different models, prompting strategies, or fine-tuning approaches may yield different results. Second, the evaluation of generated utterances relied on a RoBERTa-based classifier for DA annotation, which, while robust, introduces its sources of error. Finally, the study focused on specific dialogue datasets, persuasive donation dialogues and the Frames corpus, which may limit generalizability to other dialogue domains.

Open questions remain regarding how to mitigate overprediction biases, particularly in decision-oriented tasks, and how best to calibrate LLM outputs to better reflect ground truth distributions. It also remains unclear how persona traits influence decision generation in more complex conversational scenarios and whether models can be systematically adapted to better align with human decision-making styles.

8.3 Future Work

As an example of future research on this topic, we can present fine-tuning selected large language models specifically to simulate each user in each dialogue (as in both datasets, different numbers of participants were involved in recording the data). Another option for future research is the integration of hybrid architectures that combine generative capabilities with symbolic reasoning or structured planning modules. Such an option could improve both accuracy and interpretability. Surely, more dialogue datasets can be searched for or created for the same purposes. This would allow us to extend the experiments to include domains with more diverse user goals, allowing us to evaluate generalizability.

An additional direction for future research involves incorporating human judgment as part of the evaluation pipeline. While this thesis primarily relied on quantitative metrics, such as accuracy, F_1-scores, and distributional comparisons, these measures cannot fully capture the perceived naturalness, coherence, or persona-consistency of generated utterances. Human evaluators could

assess whether the LLM-generated responses appropriately reflect the inferred user persona, conversational intent, and decision-making style. Such evaluations could be conducted through structured annotation tasks, pairwise preference tests, or Likert-scale ratings [24] of attributes like politeness, decisiveness, or alignment with the preceding context. Integrating human judgment would provide richer insights into the qualitative aspects of LLM performance and help validate whether models are genuinely mimicking user behaviour, rather than merely achieving high scores on automated metrics.

Moreover, calibration techniques, such as post-hoc adjustments to predicted probabilities [18], could be explored to reduce overcommitment to affirmative decisions and overestimation of numerical values. Persona modelling remains a promising direction: future work could examine how explicit persona conditioning impacts decision prediction fidelity and conversational naturalness. Finally, reinforcement learning with human feedback (RLHF) could be employed to optimise LLM behaviour for goal-oriented dialogues that require both content accuracy and persona-consistent style.

Acknowledgements

I would like to express my sincere gratitude to Marcel Gohsen for his invaluable guidance throughout this thesis. His insightful feedback, advice and experience have been instrumental in shaping this research.

Appendix A

Experiment II Results

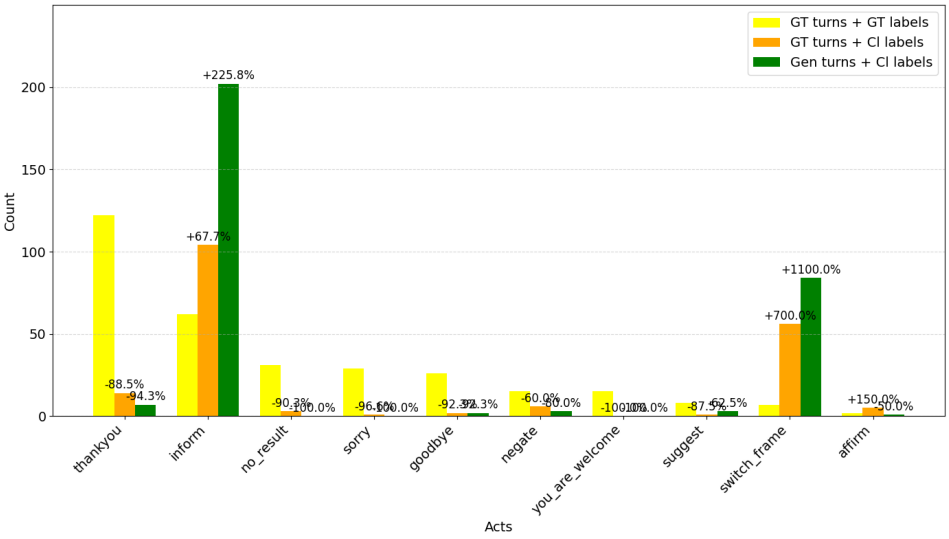

Figure A.1: Comparison of dialogue act labels **after 10 turns**: yellow - original utterances (ground truth) and corresponding original labels from the corpus (ground truth); orange - original removed utterances from the corpus (ground truth) and corresponding labels classified by the classifier; and green - utterances generated by **Gemma 2** and corresponding labels classified by the classifier.

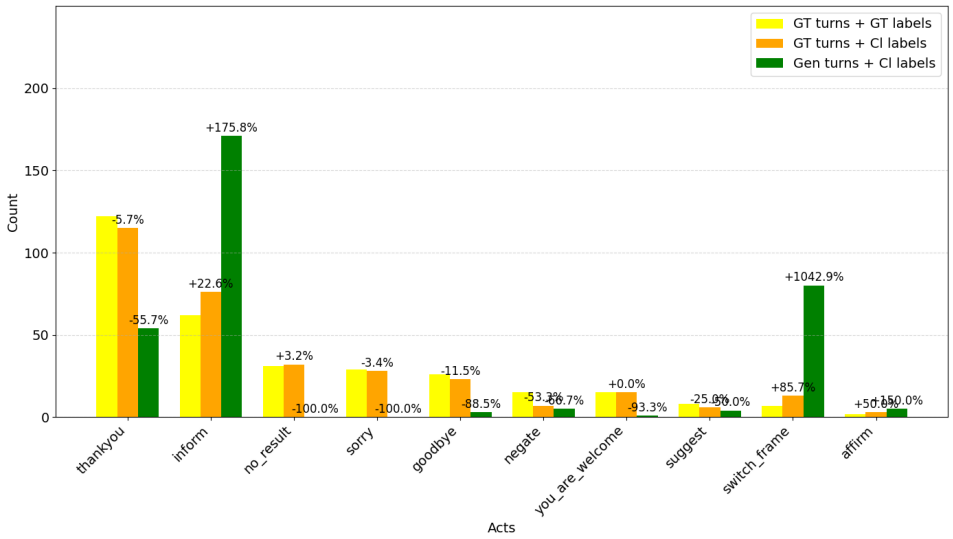

Figure A.2: Comparison of dialogue act labels **after all turns**: yellow - original utterances (ground truth) and corresponding original labels from the corpus (ground truth); orange - original removed utterances from the corpus (ground truth) and corresponding labels classified by the classifier; and green - utterances generated by **Gemma 2** and corresponding labels classified by the classifier.

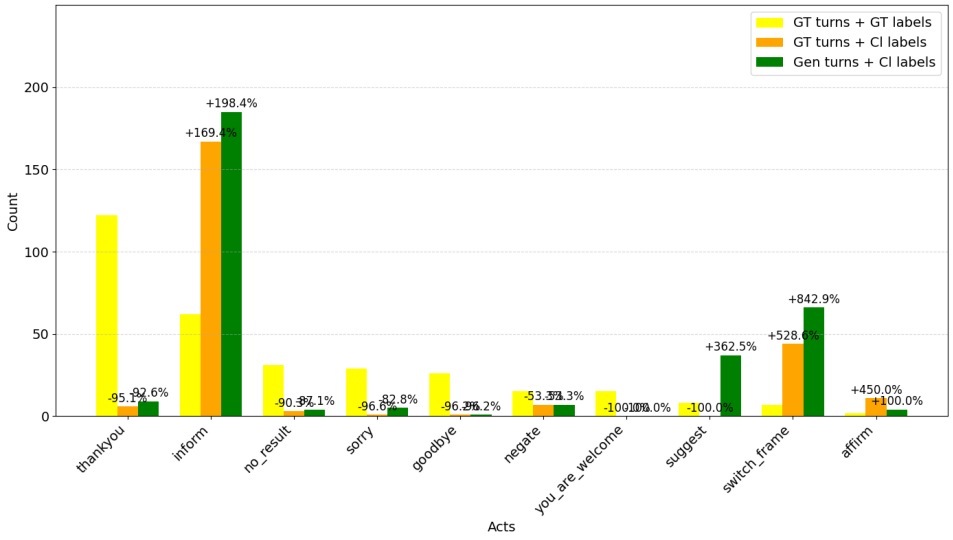

Figure A.3: Comparison of dialogue act labels **after 6 turns**: yellow - original utterances (ground truth) and corresponding original labels from the corpus (ground truth); orange - original removed utterances from the corpus (ground truth) and corresponding labels classified by the classifier; and green - utterances generated by **Dolphin-Llama3** and corresponding labels classified by the classifier.

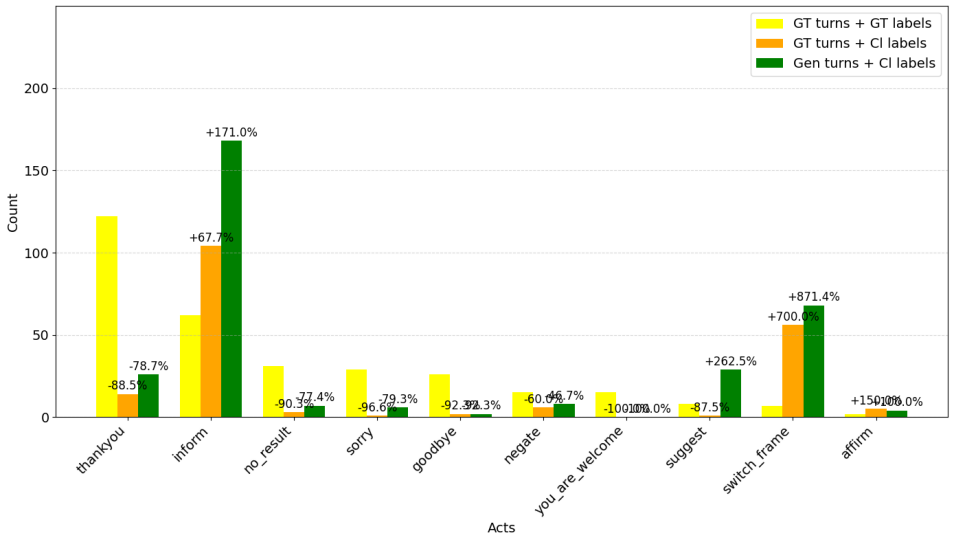

Figure A.4: Comparison of dialogue act labels **after 10 turns**: yellow - original utterances (ground truth) and corresponding original labels from the corpus (ground truth); orange - original removed utterances from the corpus (ground truth) and corresponding labels classified by the classifier; and green - utterances generated by **Dolphin-Llama3** and corresponding labels classified by the classifier.

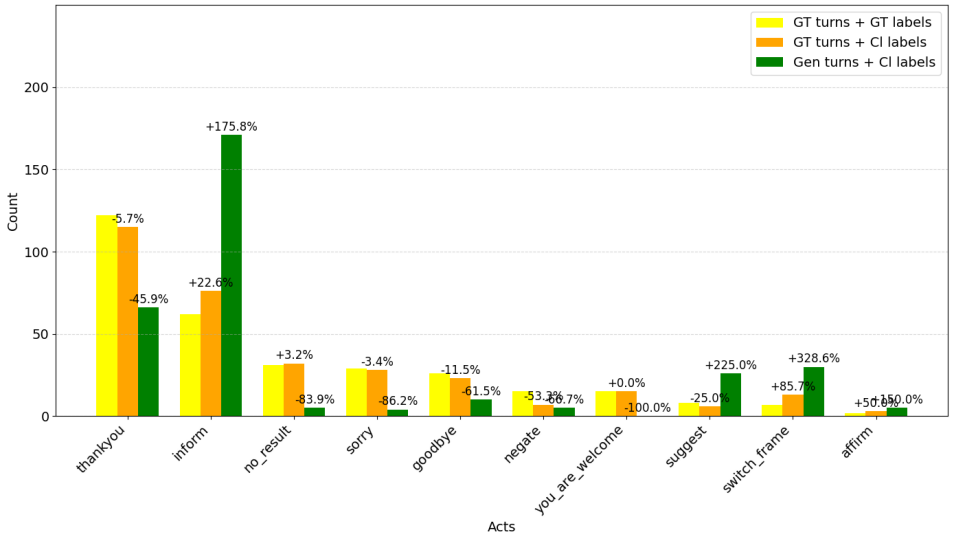

Figure A.5: Comparison of dialogue act labels **after all turns**: yellow - original utterances (ground truth) and corresponding original labels from the corpus (ground truth); orange - original removed utterances from the corpus (ground truth) and corresponding labels classified by the classifier; and green - utterances generated by **Dolphin-Llama3** and corresponding labels classified by the classifier.

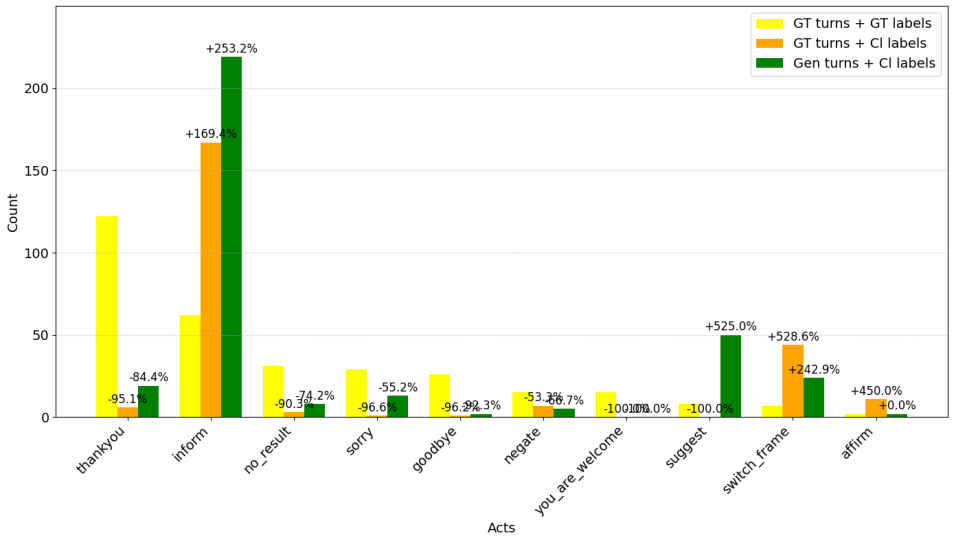

Figure A.6: Comparison of dialogue act labels **after 6 turns**: yellow - original utterances (ground truth) and corresponding original labels from the corpus (ground truth); orange - original removed utterances from the corpus (ground truth) and corresponding labels classified by the classifier; and green - utterances generated by **Mistral** and corresponding labels classified by the classifier.

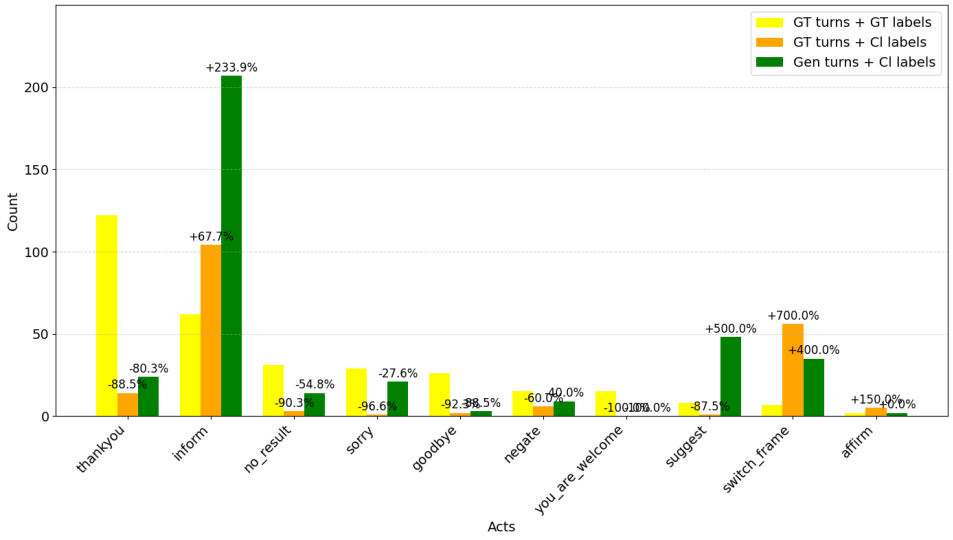

Figure A.7: Comparison of dialogue act labels **after 10 turns**: yellow - original utterances (ground truth) and corresponding original labels from the corpus (ground truth); orange - original removed utterances from the corpus (ground truth) and corresponding labels classified by the classifier; and green - utterances generated by **Mistral** and corresponding labels classified by the classifier.

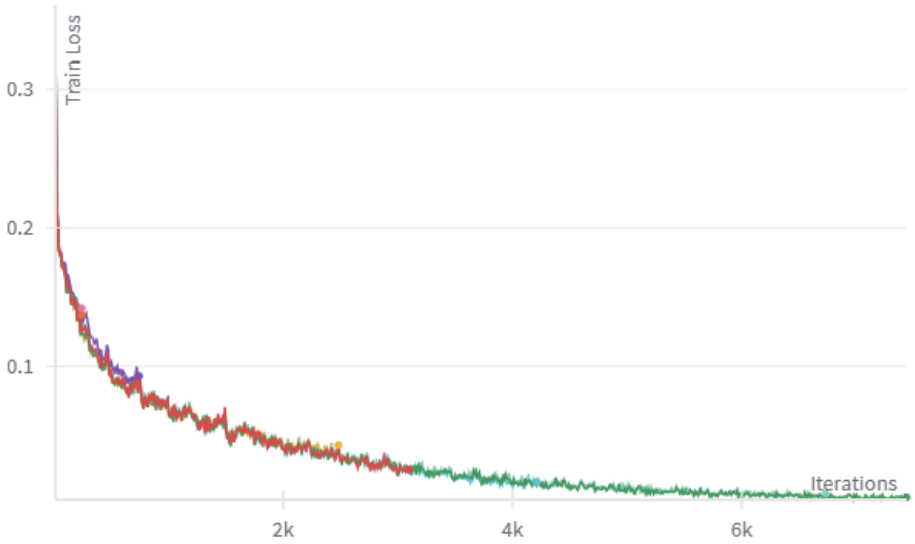

Figure A.8: Custom RoBERTa Classifier training loss; each colour represents an instance (model) of Hugging Face's RobertaForSequenceClassification, from 10 to 7440 global training steps.

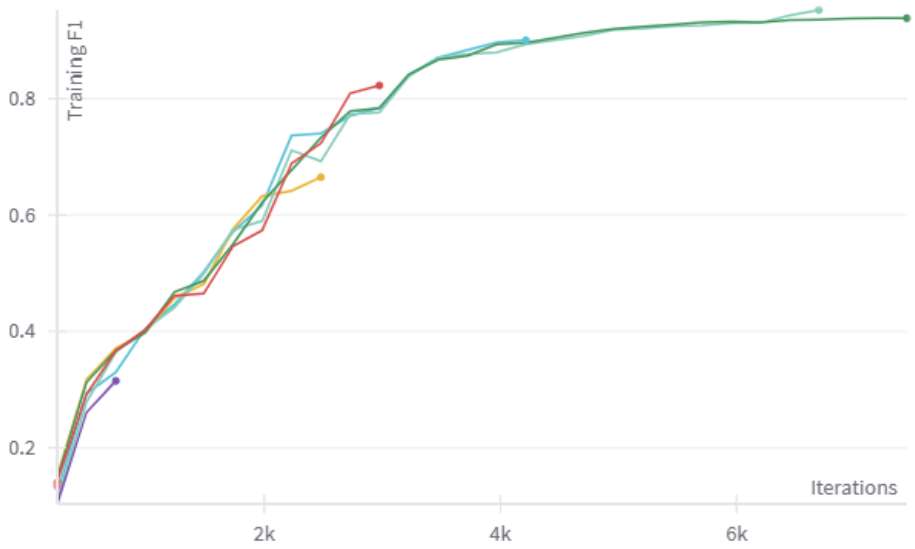

Figure A.9: Custom RoBERTa Classifier training F_1; each colour represents an instance (model) of Hugging Face's RobertaForSequenceClassification, from 10 to 7440 global training steps.

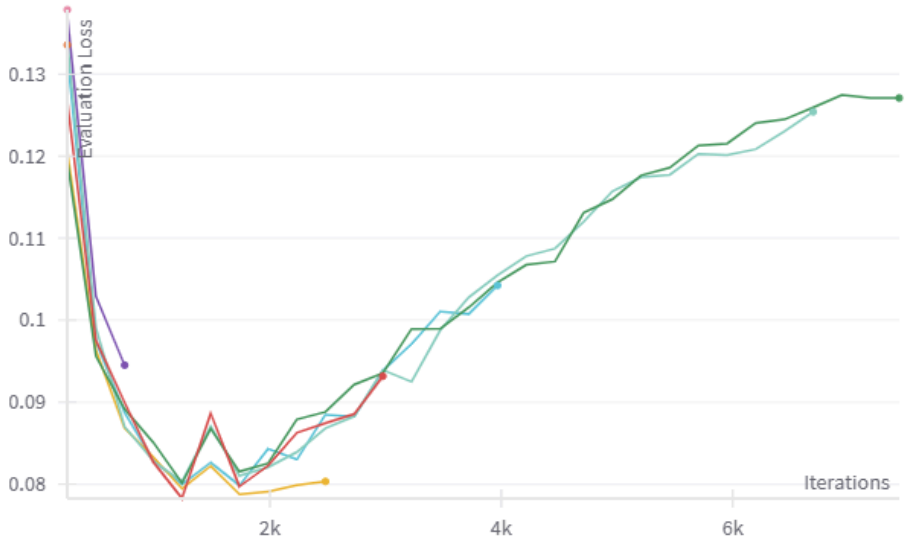

Figure A.10: Custom RoBERTa Classifier evaluation loss; each colour represents an instance (model) of Hugging Face's RobertaForSequenceClassification, from 10 to 7440 global training steps.

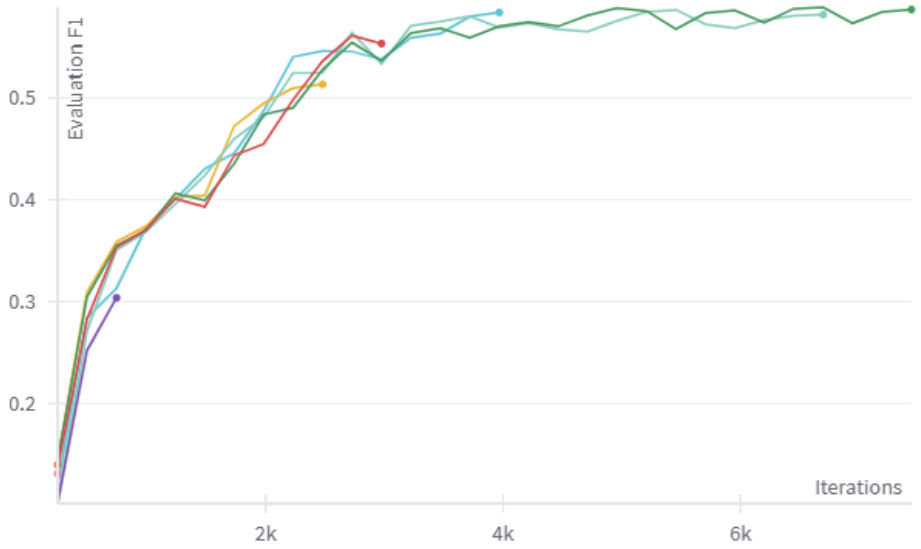

Figure A.11: Custom RoBERTa Classifier evaluation F_1; each colour represents an instance (model) of Hugging Face's RobertaForSequenceClassification, from 10 to 7440 global training steps.

Bibliography

[1] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Ainslie, J., Lee-Thorp, J., de Jong, M., Zemlyanskiy, Y., Lebrón, F., and Sanghai, S. Gqa: Training generalized multi-query transformer models from multi-head checkpoints, 2023.

[3] Bai, G., Liu, J., Bu, X., He, Y., Liu, J., Zhou, Z., Lin, Z., Su, W., Ge, T., Zheng, B., and Ouyang, W. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2024), Association for Computational Linguistics, p. 7421–7454.

[4] Beltagy, I., Peters, M. E., and Cohan, A. Longformer: The long-document transformer, 2020.

[5] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems 33* (2020), 1877–1901.

[6] Budzianowski, P., Wen, T., Tseng, B., Casanueva, I., Ultes, S., Ramadan, O., and Gasic, M. Multiwoz - A large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *CoRR abs/1810.00278* (2018).

[7] Cheng, M., Piccardi, T., and Yang, D. Compost: Characterizing and evaluating caricature in llm simulations, 2023.

[8] Cho, S., Kim, J., and Kim, J. H. LLM-based doppelgänger models: Leveraging synthetic dat for human-like responses in survey simulations. *IEEE Access 12* (2024), 178917–178927.

[9] El Asri, L., Schulz, H., Sharma, S., Zumer, J., Harris, J., Fine, E., Mehrotra, R., and Suleman, K. Frames: a corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue* (Stroudsburg, PA, USA, 2017), Association for Computational Linguistics.

[10] Gao, X., Zhang, Y., Galley, M., Brockett, C., and Dolan, B. Dialogue response ranking training with large-scale human feedback data, 2020.

[11] Gemma Team, Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., and et al. Gemma 2: Improving open language models at a practical size, 2024.

[12] Goffman, E. *The Presentation of Self in Everyday Life*. Anchor Books, 1959.

[13] Graesser, A. C., Hu, X., Nye, B. D., VanLehn, K., Kumar, R., Heffernan, C., Heffernan, N., Woolf, B., Olney, A. M., Rus, V., et al. Electronixtutor: an intelligent tutoring system with multiple learning resources for electronics. *International journal of STEM education 5*, 1 (2018), 15.

[14] Grice, H. P. Logic and conversation. In *Syntax and Semantics 3: Speech Acts*, P. Cole and J. Morgan, Eds. Academic Press, 1975, pp. 41–58.

[15] Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., Wu, J., Mirjalili, S., et al. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea preprints 1*, 3 (2023), 1–26.

[16] Hartford, E., Atkins, L., Neto, F. F., and Golchinfar, D. Spectrum: Targeted training on signal to noise ratio, 2024.

[17] He, F., Zhu, T., Ye, D., Liu, B., Zhou, W., and Yu, P. S. The emerged security and privacy of llm agent: A survey with case studies, 2024.

[18] Hekler, A., Brinker, T. J., and Buettner, F. Test time augmentation meets post-hoc calibration: uncertainty quantification under real-world conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2023), vol. 37, pp. 14856–14864.

[19] Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., and Sifre, L. Training compute-optimal large language models, 2022.

[20] Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. The curious case of neural text degeneration, 2020.

[21] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. Survey of hallucination in natural language generation. *ACM computing surveys 55*, 12 (2023), 1–38.

[22] Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7b, 2023.

[23] John, O. P., and Srivastava, S. The big five trait taxonomy: History, measurement, and theoretical perspectives. In *Handbook of Personality: Theory and Research*, L. A. Pervin and O. P. John, Eds. Guilford Press, 1999, pp. 102–138.

[24] Joshi, A., Kale, S., Chandel, S., and Pal, D. K. Likert scale: Explored and explained. *British journal of applied science & technology 7*, 4 (2015), 396.

[25] Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B. A persona-based neural conversation model. In *Proceedings of ACL* (2016), pp. 994–1003.

[26] Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. G-eval: Nlg evaluation using gpt-4 with better human alignment, 2023.

[27] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach, 2019.

[28] Loshchilov, I., and Hutter, F. Decoupled weight decay regularization, 2019.

[29] Mairesse, F., and Walker, M. A. Using linguistic cues for the automatic recognition of personality in conversation and text. In *Journal of Artificial Intelligence Research* (2007), vol. 30, pp. 457–500.

[30] Mairesse, F., and Walker, M. A. Towards personality-based user adaptation: psychologically informed stylistic language generation. *User Model. User-adapt Interact. 20*, 3 (Aug. 2010), 227–278.

[31] Marincioni, A., Miltiadous, M., Zacharia, K., Heemskerk, R., Doukeris, G., Preuss, M., and Barbero, G. The effect of LLM-based NPC emotional states on player emotions: An analysis of interactive game play. In *2024 IEEE Conference on Games (CoG)* (Aug. 2024), IEEE, pp. 1–6.

[32] Milne-Ives, M., de Cock, C., Lim, E., Shehadeh, M. H., de Pennington, N., Mole, G., Normando, E., and Meinert, E. The effectiveness of artificial intelligence conversational agents in health care: Systematic review. *J. Med. Internet Res. 22*, 10 (Oct. 2020), e20346.

[33] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research 12* (2011), 2825–2830.

[34] Pennebaker, J. W., Mehl, M. R., and Niederhoffer, K. G. Psychological aspects of natural language. use: our words, our selves. *Annu. Rev. Psychol. 54*, 1 (2003), 547–577.

[35] Pennebaker, J. W., Mehl, M. R., and Niederhoffer, K. G. Psychological aspects of natural language use: Our words, our selves. *Trends in Cognitive Sciences 7*, 12 (2003), 547–551.

[36] Renze, M., and Guven, E. The effect of sampling temperature on problem solving in large language models, 2024.

[37] Rizwan, M., Carlsson, L., and Loni, M. Personabot: Bringing customer personas to life with llms and rag, 2025.

[38] Rojas-Barahona, L. M. Talking to machines: do you read me?, 2024.

[39] Rome, S., Chen, T., Tang, R., Zhou, L., and Ture, F. "ask me anything": How comcast uses LLMs to assist agents in real time. In

Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (New York, NY, USA, July 2024), ACM, pp. 2827–2831.

[40] Salber, D., and Coutaz, J. Applying the wizard of oz technique to the study of multimodal systems. In *Lecture Notes in Computer Science*, Lecture notes in computer science. Springer Berlin Heidelberg, Berlin, Heidelberg, 1993, pp. 219–230.

[41] SCHATZMANN, J., WEILHAMMER, K., STUTTLE, M., and YOUNG, S. A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. *The Knowledge Engineering Review 21*, 2 (2006), 97–126.

[42] Schuller, A., Janssen, D., Blumenröther, J., Probst, T. M., Schmidt, M., and Kumar, C. Generating personas using llms and assessing their viability. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2024), CHI EA '24, Association for Computing Machinery.

[43] Sechidis, K., Tsoumakas, G., and Vlahavas, I. On the stratification of multi-label data. In *Joint European conference on machine learning and knowledge discovery in databases* (2011), Springer, pp. 145–158.

[44] See, A., Roller, S., Kiela, D., and Weston, J. What makes a good conversation? how controllable attributes affect human judgments. *CoRR abs/1902.08654* (2019).

[45] Shuster, K., Boureau, Y.-L., and Weston, J. Dialogue natural language inference. *arXiv preprint arXiv:2004.04954* (2020).

[46] Sun, L., Qin, T., Hu, A., Zhang, J., Lin, S., Chen, J., Ali, M., and Prpa, M. Persona-L has entered the chat: Leveraging LLMs and ability-based framework for personas of people with complex needs. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, Apr. 2025), ACM, pp. 1–31.

[47] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., and et al. Llama 2: Open foundation and fine-tuned chat models, 2023.

[48] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2023.

[49] Venkatesh Sharma, K., Ayiluri, P. R., Betala, R., Jagdish Kumar, P., and Shirisha Reddy, K. Enhancing query relevance: leveraging sbert and cosine similarity for optimal information retrieval. *International Journal of Speech Technology 27*, 3 (2024), 753–763.

[50] Wang, X., Shi, W., Kim, R., Oh, Y., Yang, S., Zhang, J., and Yu, Z. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence, Italy, July 2019), A. Korhonen, D. Traum, and L. Màrquez, Eds., Association for Computational Linguistics, pp. 5635–5649.

[51] Wu, C., Wu, F., An, M., Huang, J., Huang, Y., and Xie, X. Neural news recommendation with attentive multi-view learning, 2019.

[52] Wu, T.-L., Kottur, S., Madotto, A., Azab, M., Rodriguez, P., Damavandi, B., Peng, N., and Moon, S. SIMMC-VR: A task-oriented multimodal dialog dataset with situated and immersive VR streams. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Toronto, Canada, July 2023), A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds., Association for Computational Linguistics, pp. 6273–6291.

[53] Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of ACL* (2018), pp. 2204–2213.

[54] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. Bertscore: Evaluating text generation with bert, 2020.

[55] Zhang, Y., Wang, S., Liu, J., Yu, S., Dong, Z., Liu, S., Liu, X., and Zhu, E. DLEFT-MKC: Dynamic late fusion multiple kernel clustering with robust tensor learning via min-max optimization. In *The Thirteenth International Conference on Learning Representations* (2025).

[56] Zhou, H., Zheng, C., Huang, K., Huang, M., and Zhu, X. Kdconv: A chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation, 2020.