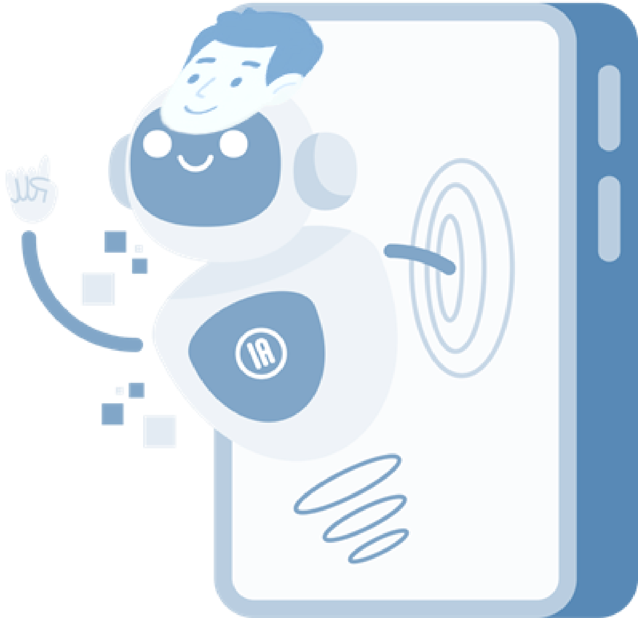
Bauhaus-Universität Weimar
Faculty of Media
Human-Computer Interaction M. Sc.

Do Large Language Models Extrapolate Personas from Dialogue Context?

Daria Zhukova, 125603

Master's Thesis Defense

Prof. Dr. Benno Stein
Jun.-Prof. Dr. Maurice Jakesch



Roadmap

Motivation

Research questions

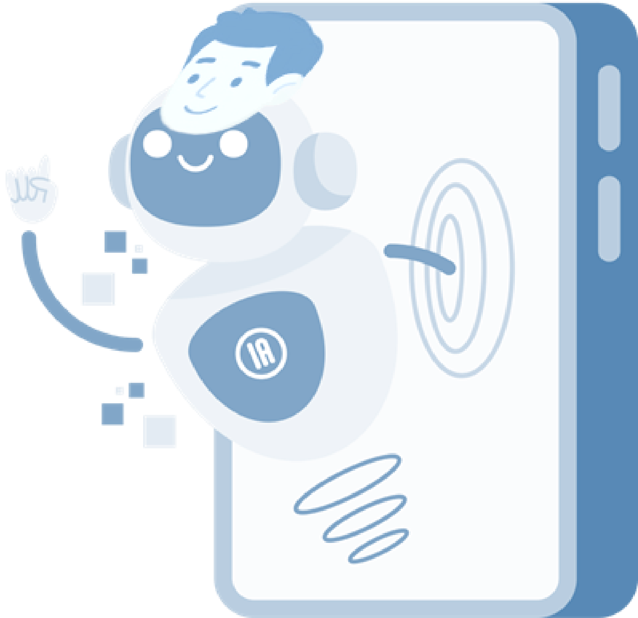
Experimental design

Key findings

Limitations & Future Work

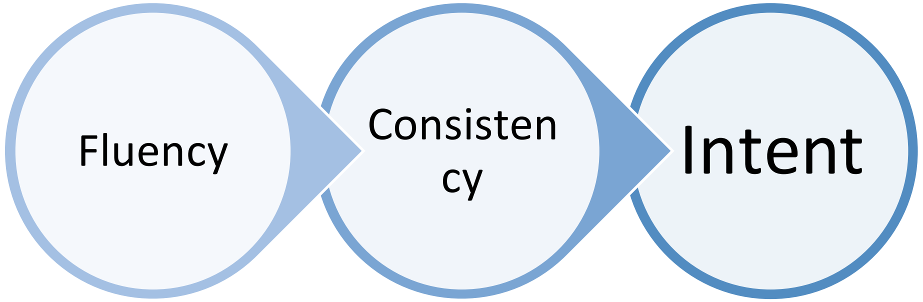
Conclusion

Q&A



Persona Mimicry

A **persona** is the *consistent style, tone, and communicative intent* through which a speaker reveals their identity, expressed in *patterns of word choice, affect, rhetorical strategies, and decision-making cues*.



- ❑ **Customer service**
- ❑ **Education**
- ❑ **Mental health**
- ❑ **General human-computer interaction**:

  Persona consistency enhances trustworthiness and satisfaction.

Research Questions

RQ1: is it possible to predict a persona’s decision from prior turns only?

RQ2: how much context does it take?

RQ3: can LLMs generate the next utterances aligned with persona’s actions?

Study Design — Two complementary experiments

Experiment I

Prediction of binary decisions in donation to the charity fund and the amount of money in the persuasive dialogues

Experiment II

Next-turn generation in travel domain dialogues, and then automatic dialogue-act matching, using a custom classifier

Models

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

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

Inference & Metrics

top_k = 1	
temperature = 1.0	
max_length = 128	
context window = 2048	

Accuracy
overall correct predictions

Precision

$$Precision = \frac{TP}{TP + FP}$$

correct positives among predicted positives

Recall

$$Recall = \frac{TP}{TP + FN}$$

correct positives among actual positives

F1 (macro)
average of per-class F1s

DA Alignment
match rate of predicted vs. GT dialogue acts (Experiment II)

F1 (micro)
global TP/FP/FN aggregation

Experiment I

Design

- Dialogues truncated to different lengths (e.g., 7 turns, 11 turns, full history)
- Final persuadee decision removed → LLM asked to predict:
  - Binary decision: donate or not
  - Donation amount (numeric value)
- Setup isolates whether LLMs can extrapolate **implicit decision-making cues** from dialogue context

Dataset I
Persuasion for Good

Dialogues
1017

Roles
- Persuader
- Persuadee

Participants
1285

Vocabulary
8141
Unique tokens

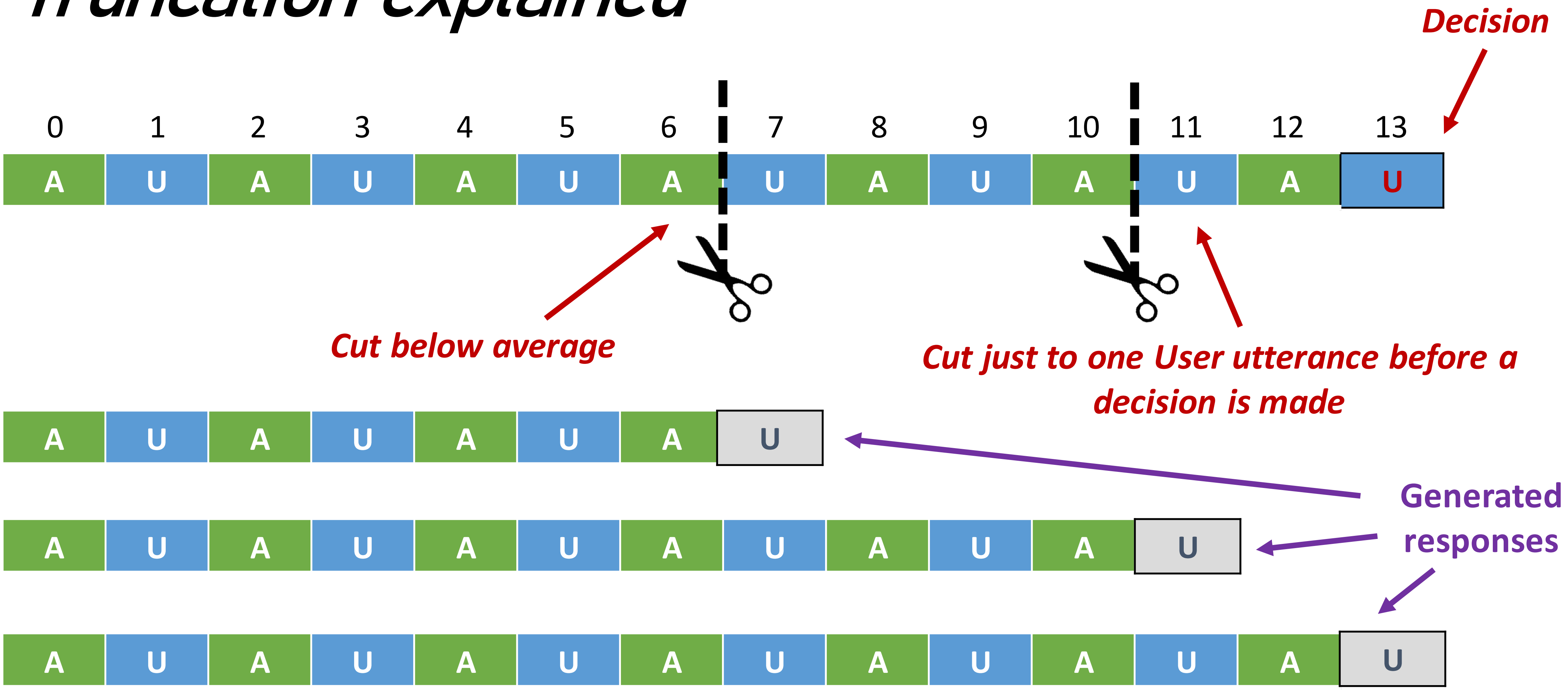
Turns per Dialogue
10,43

Utterance Length
19,36 Aords

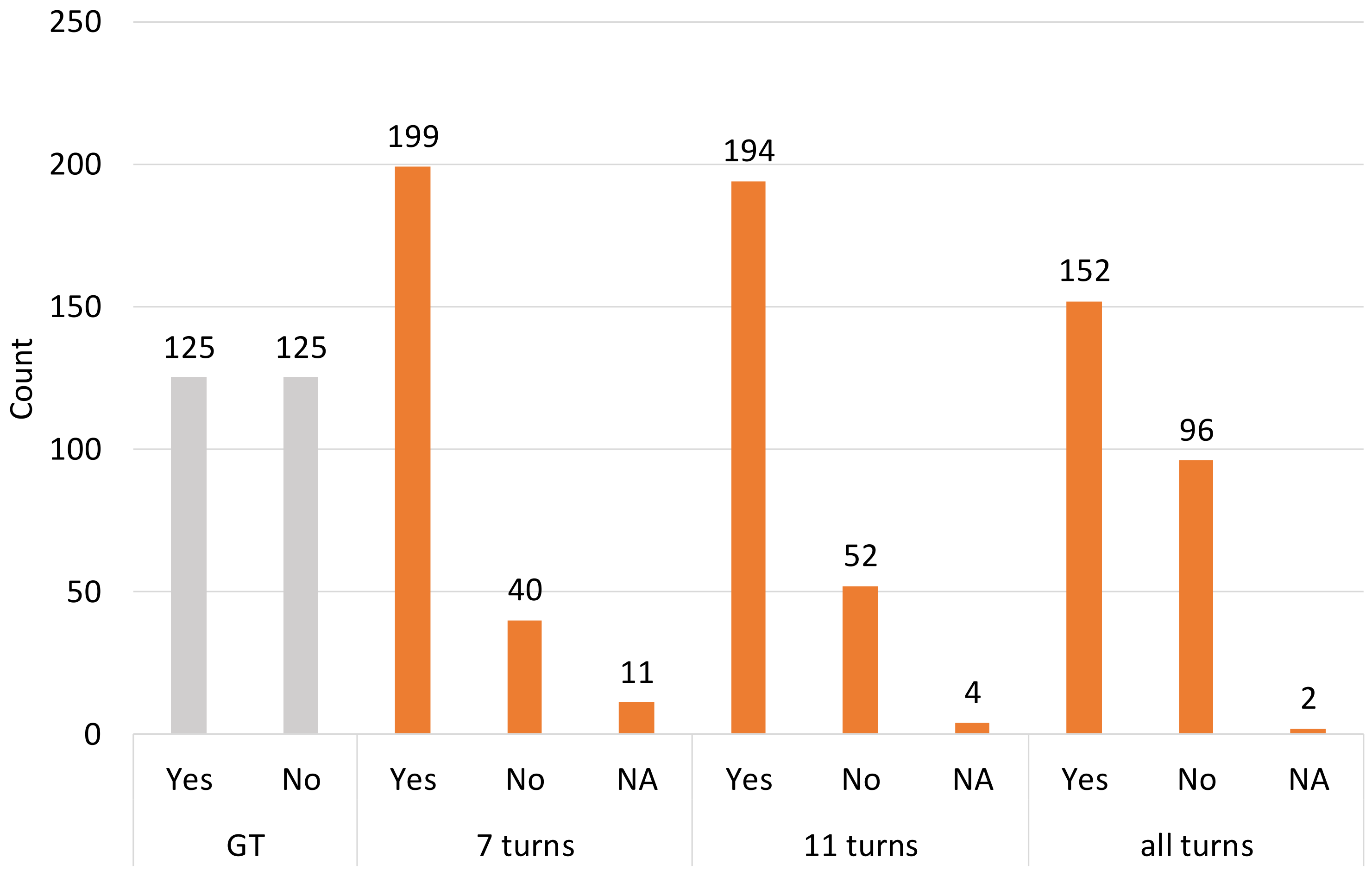
Donations
Average $0,35

Used Dialogues
125 Positive
125 Negative

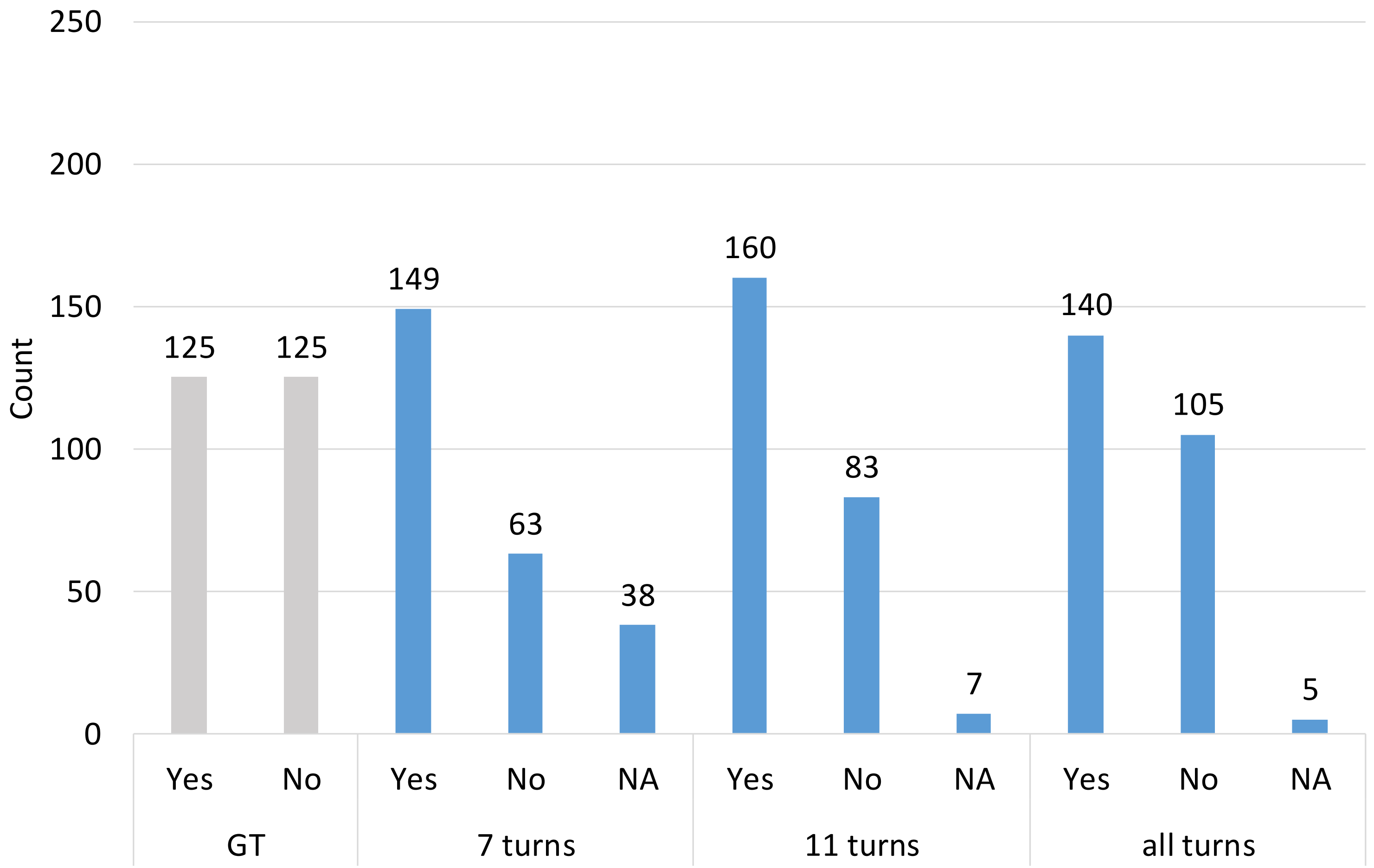
Experiment I Setup
Truncation explained



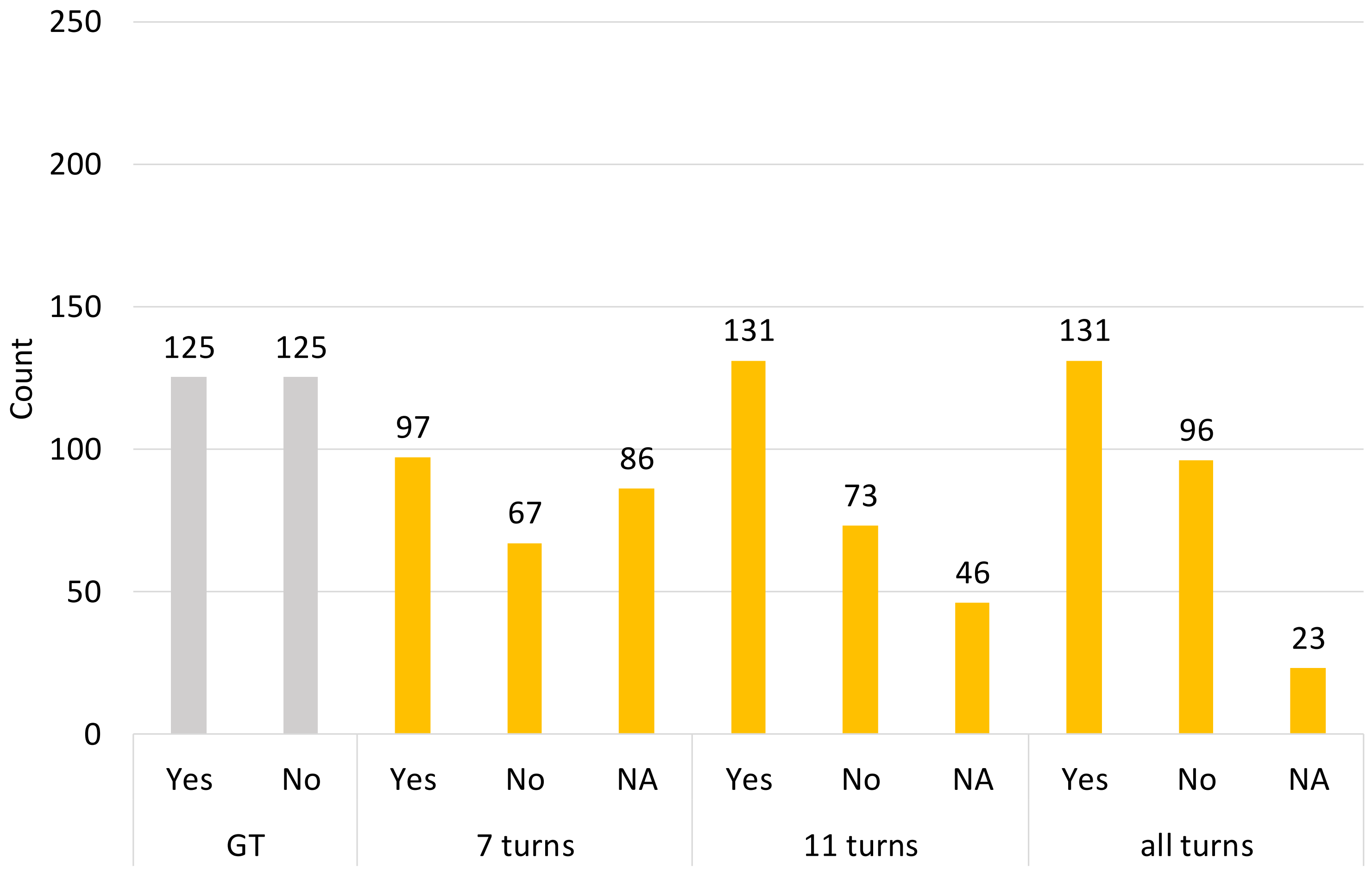
Experiment I: Yes/No Decision
Mistral



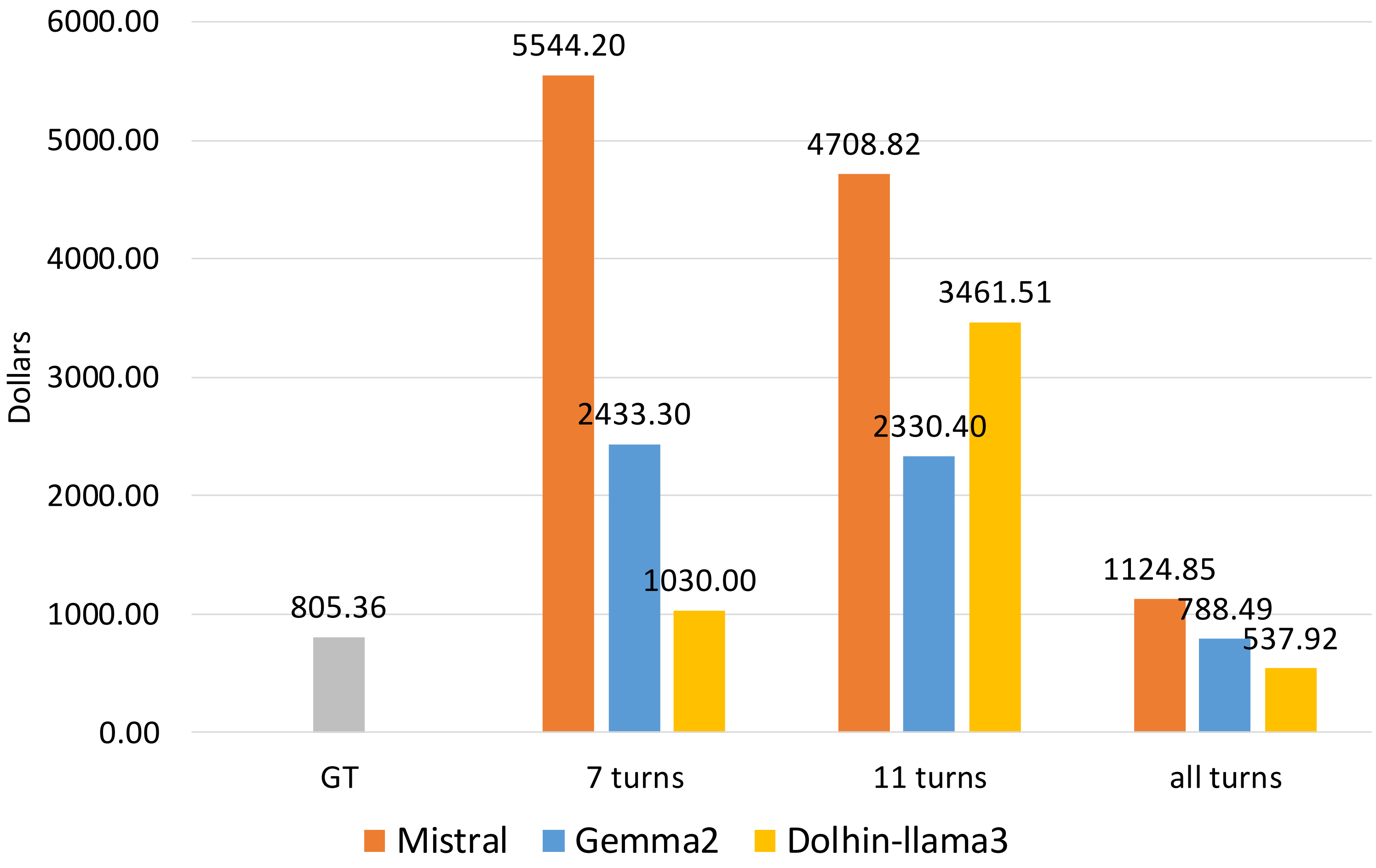
Experiment I: Yes/No Decision
Gemma 2



Experiment I: Yes/No Decision
Dolphin-Llama 3



Experiment I: Donation amounts reveal bias



Experiment II

Design

- Dialogues truncated to different lengths (e.g., 6 turns, 10 turns, full history)
- LLMs were asked to generate the missing user utterance, using only the dialogue history
- A custom RoBERTa-based multi-label Dialogue Act classifier was developed precisely for the 20 labels used in the dataset
- Generated utterances were classified with a developed DA classifier
- The Ground Truth Das were compared to Generated DAs

Dataset II
Frames

Dialogues
1369

Roles
- Wizard
- User

Turns
19986

Turns per Dialogue
14,6

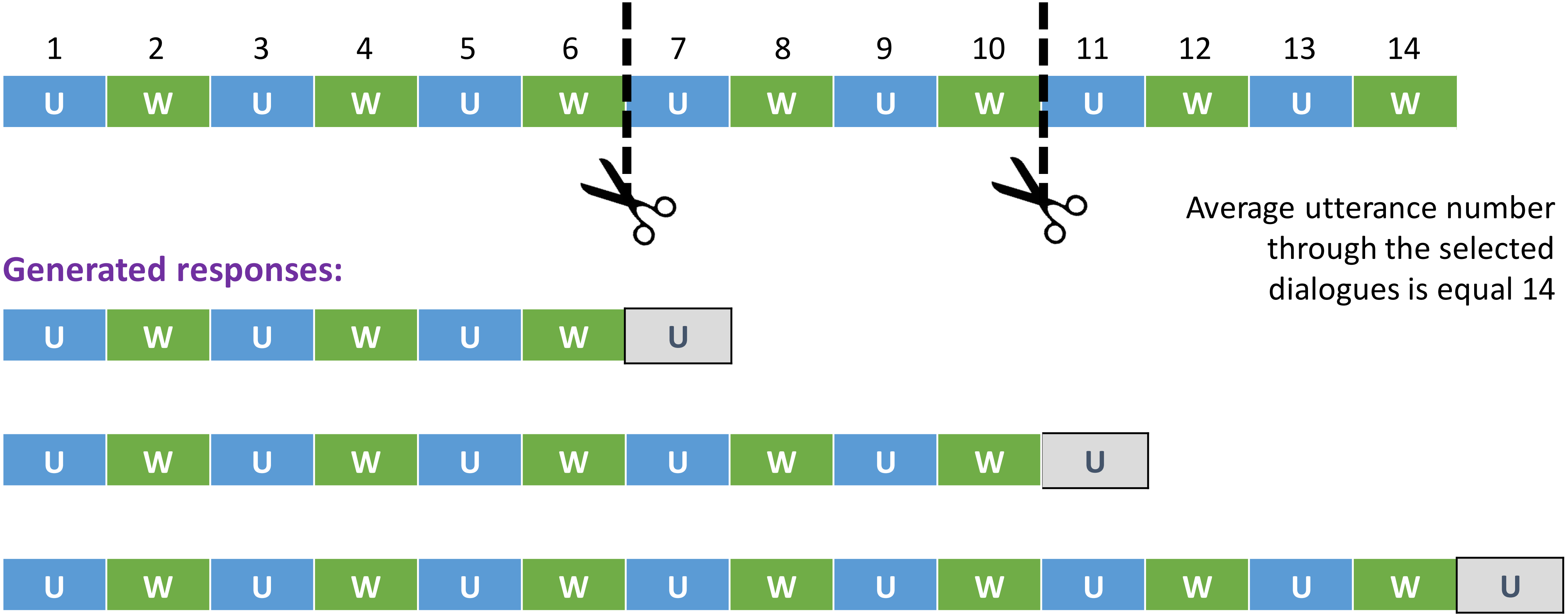
Dialogue Length
3 – 43 turns

Dialog Acts
20

Annotation
75%: 1 DA
25%: >1 DA

Used Dialogues
1101 Train
268 test

Experiment II Setup
Generating the user's next turn

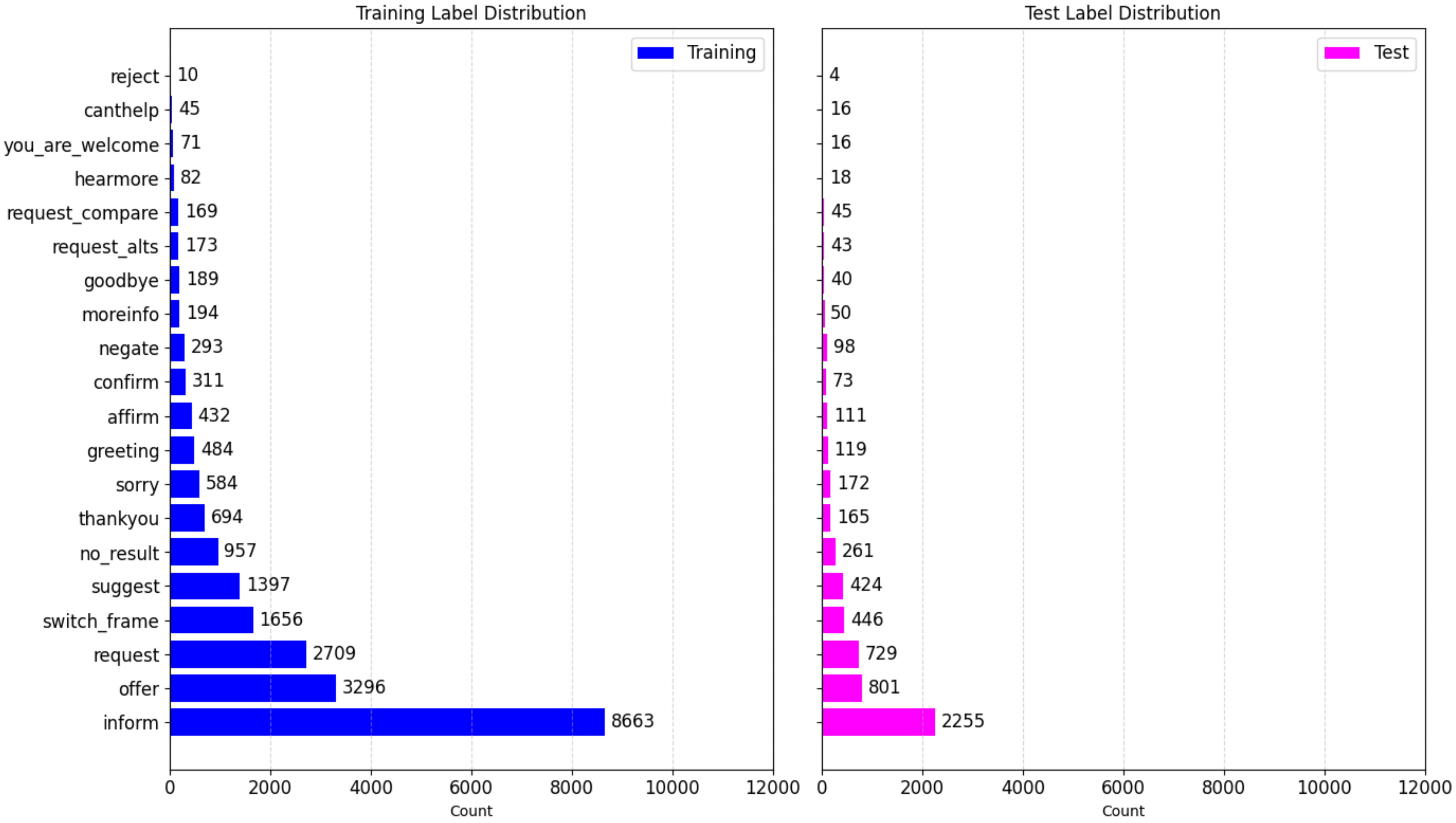


Average utterance number through the selected dialogues is equal 14

Dialogue Acts

Dialogue Act	Speaker	Description
inform	User/Wizard	Inform a slot value
thankyou	User/Wizard	Thank the other speaker
no_result	Wizard	Tell the user that the database returned no results
sorry	Wizard	Apologize to the user
goodbye	User/Wizard	Say goodbye to the other speaker
negate	User/Wizard	Negate something said by the other speaker
you_are_welcome	Wizard	Tell the user they are welcome
suggest	Wizard	Suggest a slot value or package that does not match the user's constraints
switch_frame	User	Change the topic
affirm	User/Wizard	Affirm something said by the other speaker

Dialogue Act alignment after pre-processing *Train VS Test*



Dialogue Act Classifier

Why create a custom classifier?

- Frames dataset → 20 dialogue act labels
- No suitable classifier available

How was it built?

- RoBERTa encoder + multi-label head
- Binary cross-entropy loss
- Iterative stratification → balanced labels

Performance & Role:

- F1: Train 0.93 / Validation ~0.58
- Classifies both human & LLM utterances
- Backbone for Experiment II evaluation

suggest

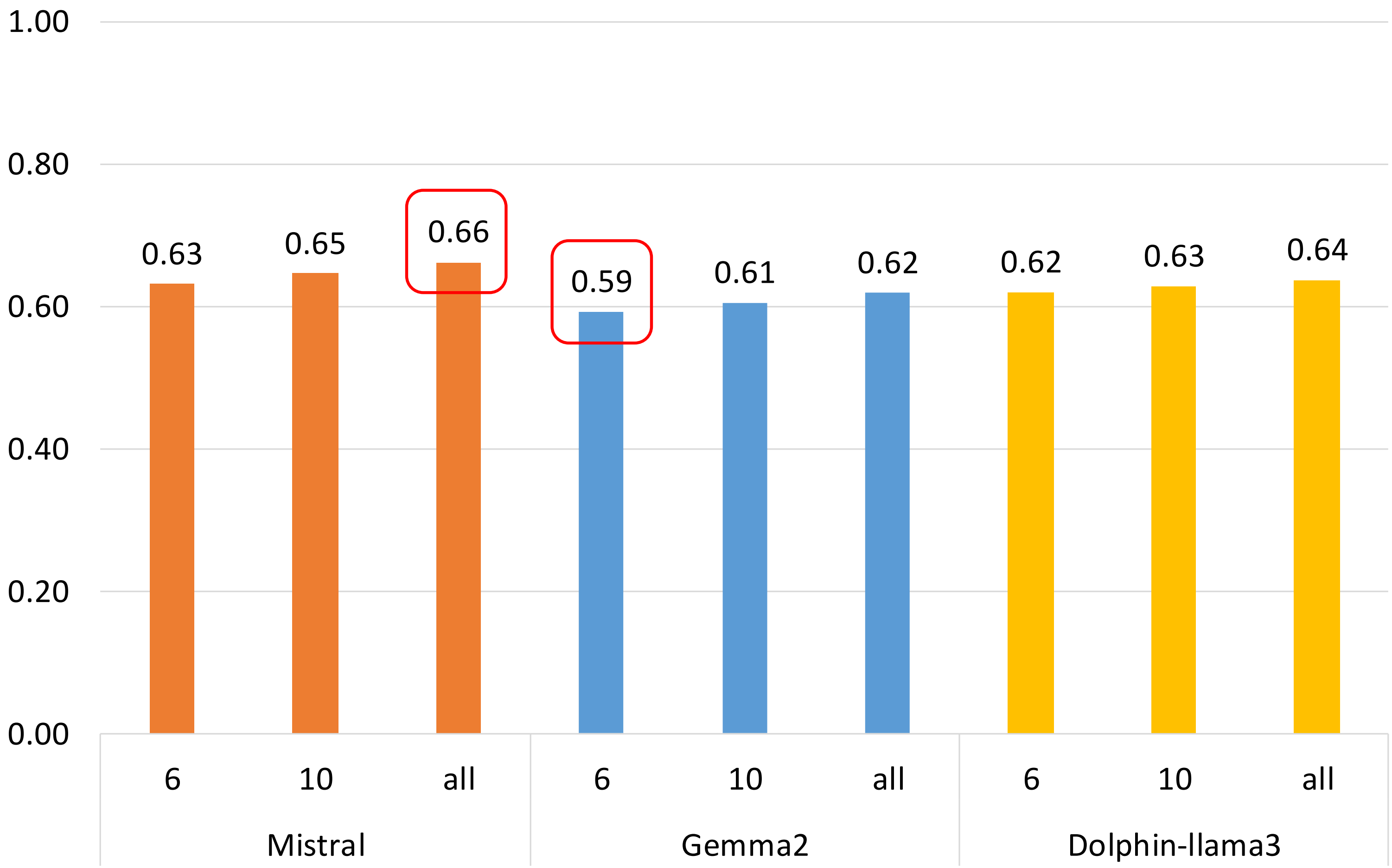
inform

affirm

no_result

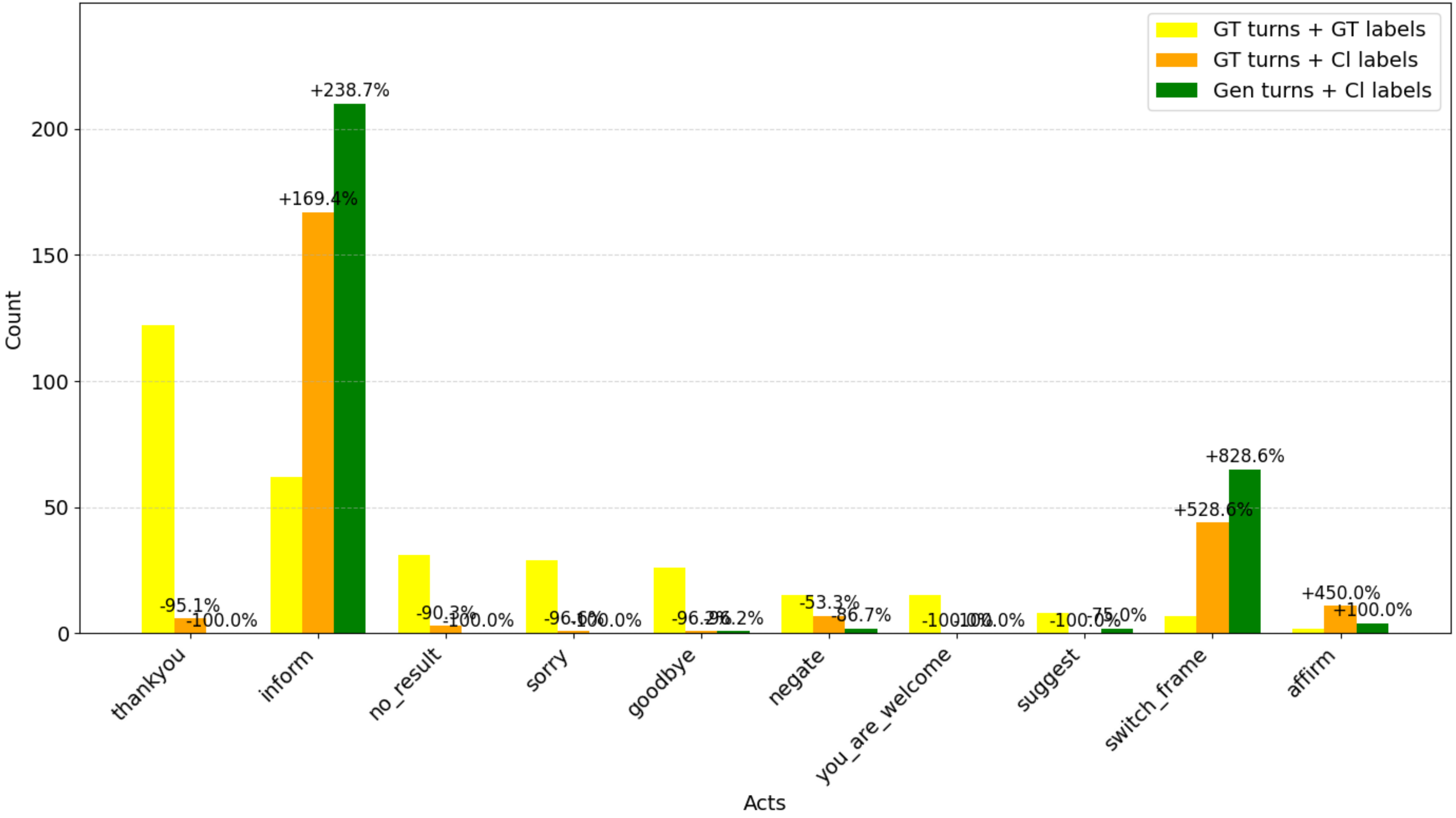
goodbye

F1-Macro for generated results of all LLMs after 6, 10 and all utterances



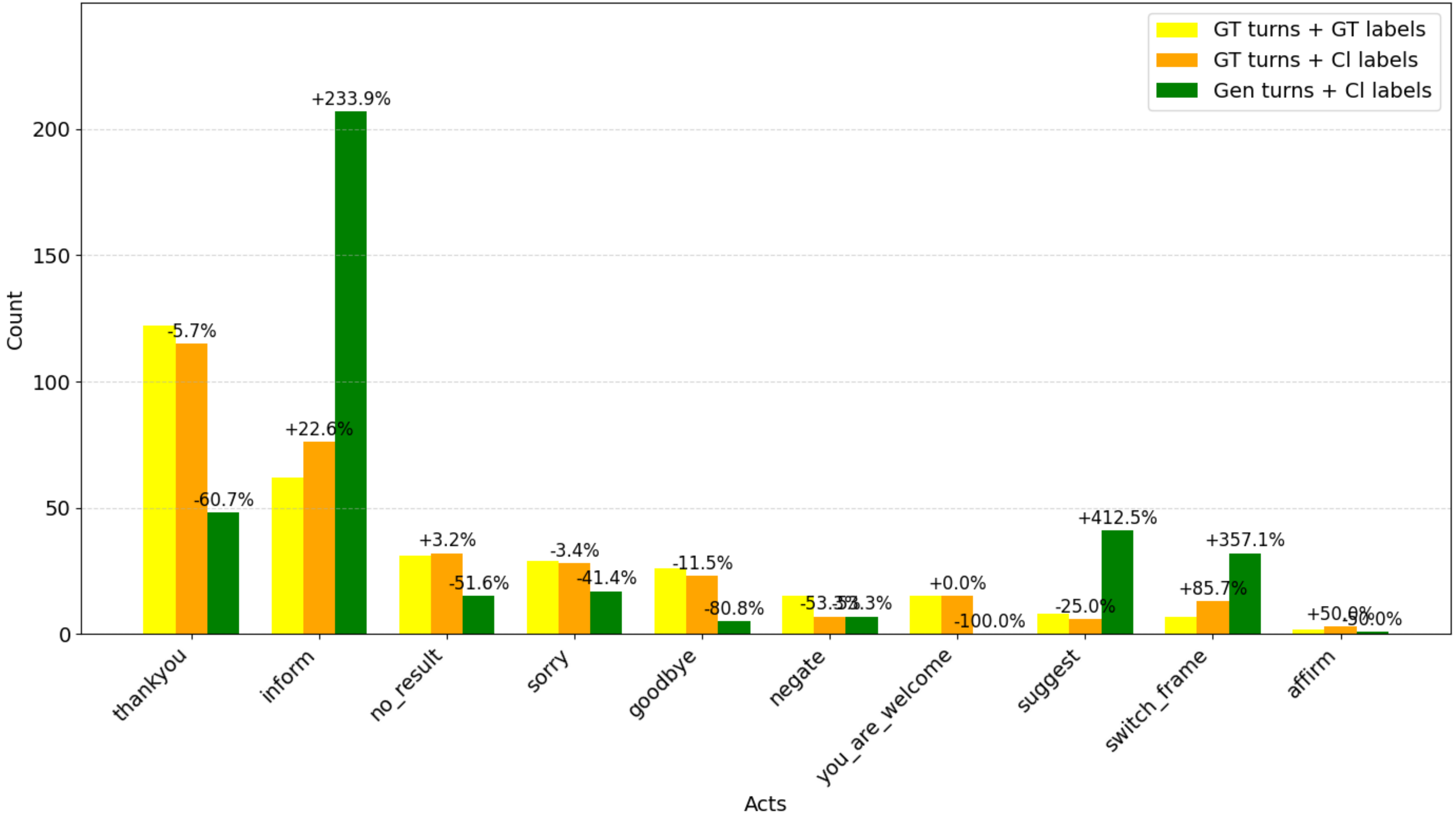
Experiment II Results

Gemma 2, after 6 turns



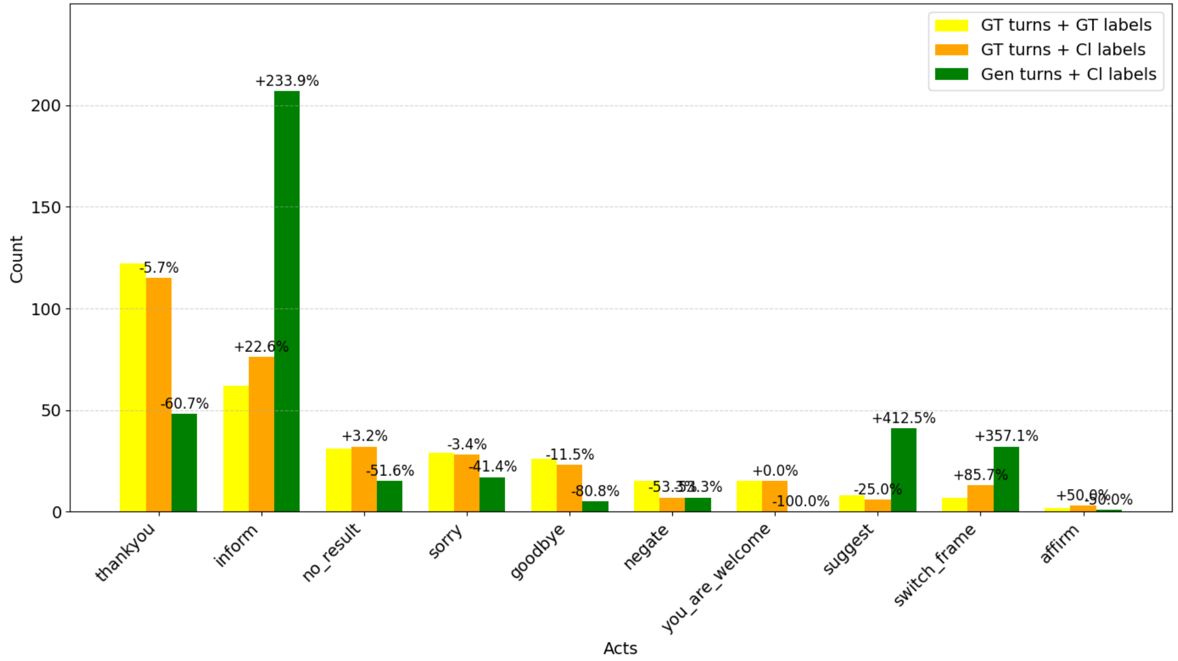
Experiment II Results

Mistral, after all turns



Where models fail

- Affirmative bias + Donation amount inflation ("Yes" answers in Expriment I)
- Dialogue Act over/under-shoot (Experiment II)



Under-produce

Over-produce

thankyou	*inform*	*no_result*	*sorry*	*goodbye*	*nagate*	*you_are_ welcome*	*suggest*	*switch_ frame*	*affirm*

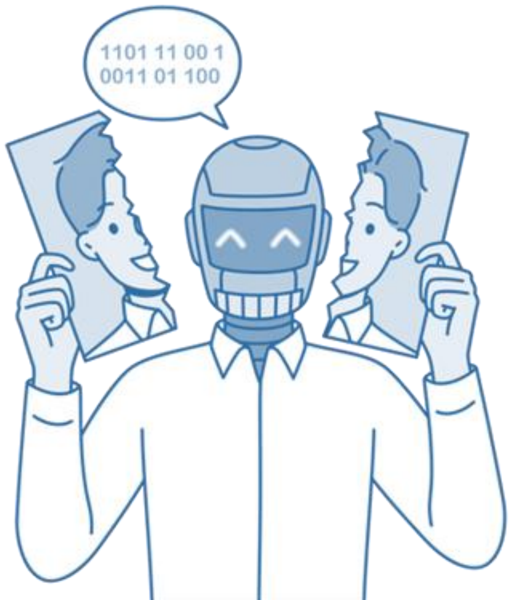
Key Findings

Can LLMs predict a persona's decision from dialogue history?

- Partially: accuracy improves with longer context;
- But → strong *affirmative bias*, inflated donation amounts, unsuitable dialogue acts that don't match the dialogue history.

How much history is needed?

- Approximately 10-11 turns are sufficient for coherent, somewhat persona-aligned outputs.



Can LLMs generate next utterances aligned with persona's actions?

- Syntax: plausible;
- Semantics: limited alignment with ground truth;
- Experiment II → F1 < 0.66; missing social acts like *thankyou* and *goodbye*.

Limitations & Future Work

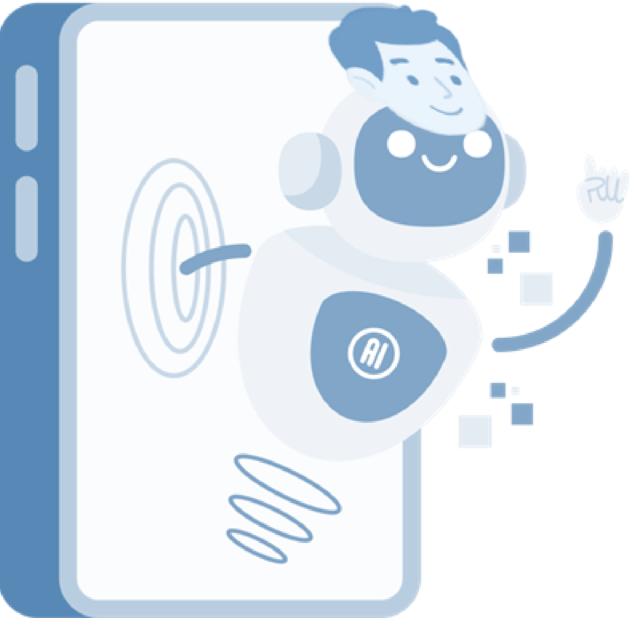
- Open-weight mid-size models only
- Reliance on automated DA labelling
- Only two domains (donations, travel)

- Task-specific fine-tuning and calibration
- Hybrid symbolic-neural approaches for reasoning
- Richer human-in-the-loop evaluation of persona alignment

Conclusion

Current LLMs mostly echo surface patterns;

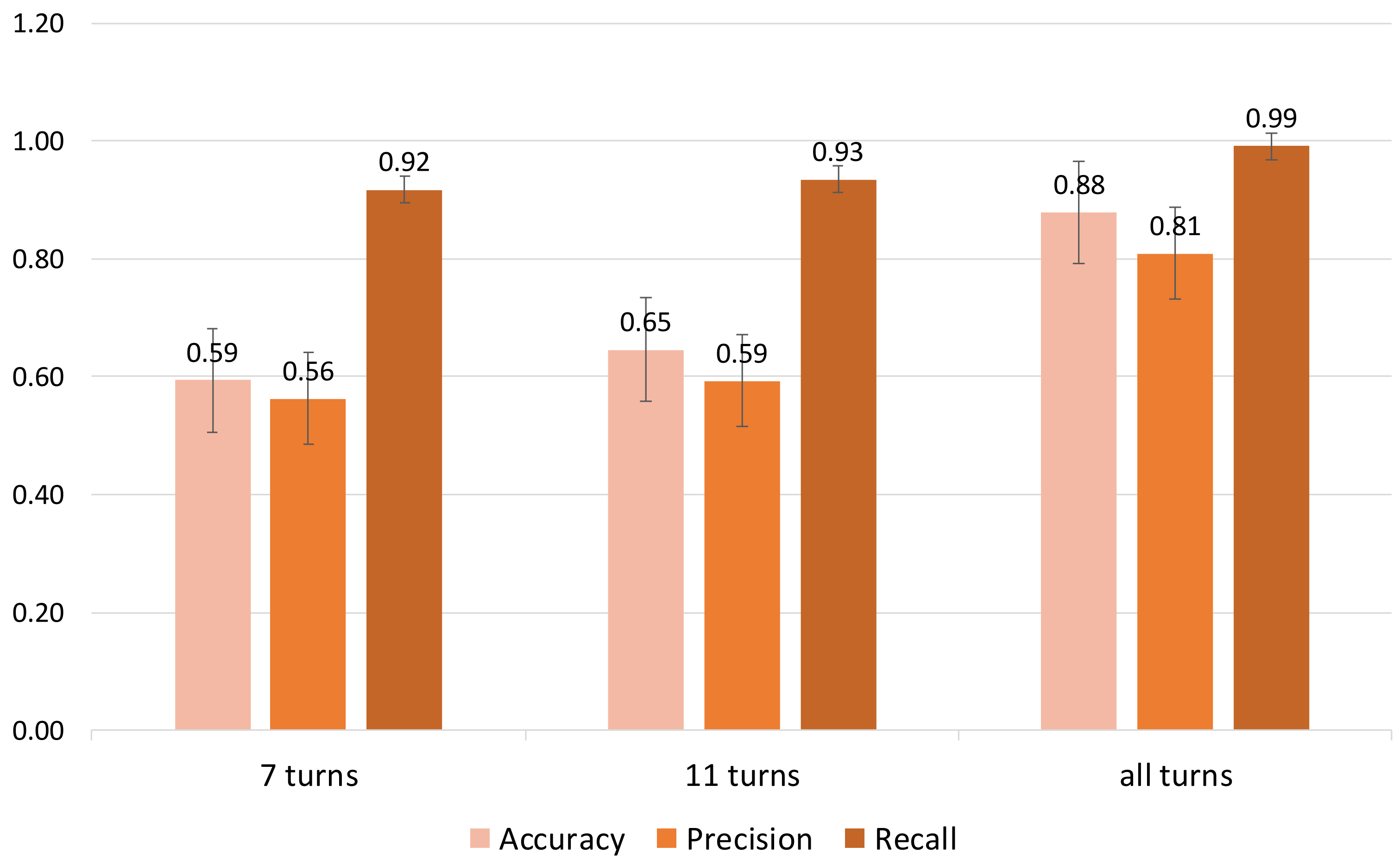
Robust persona alignment will require explicit preference elicitation, calibration, and structured (hybrid) control.



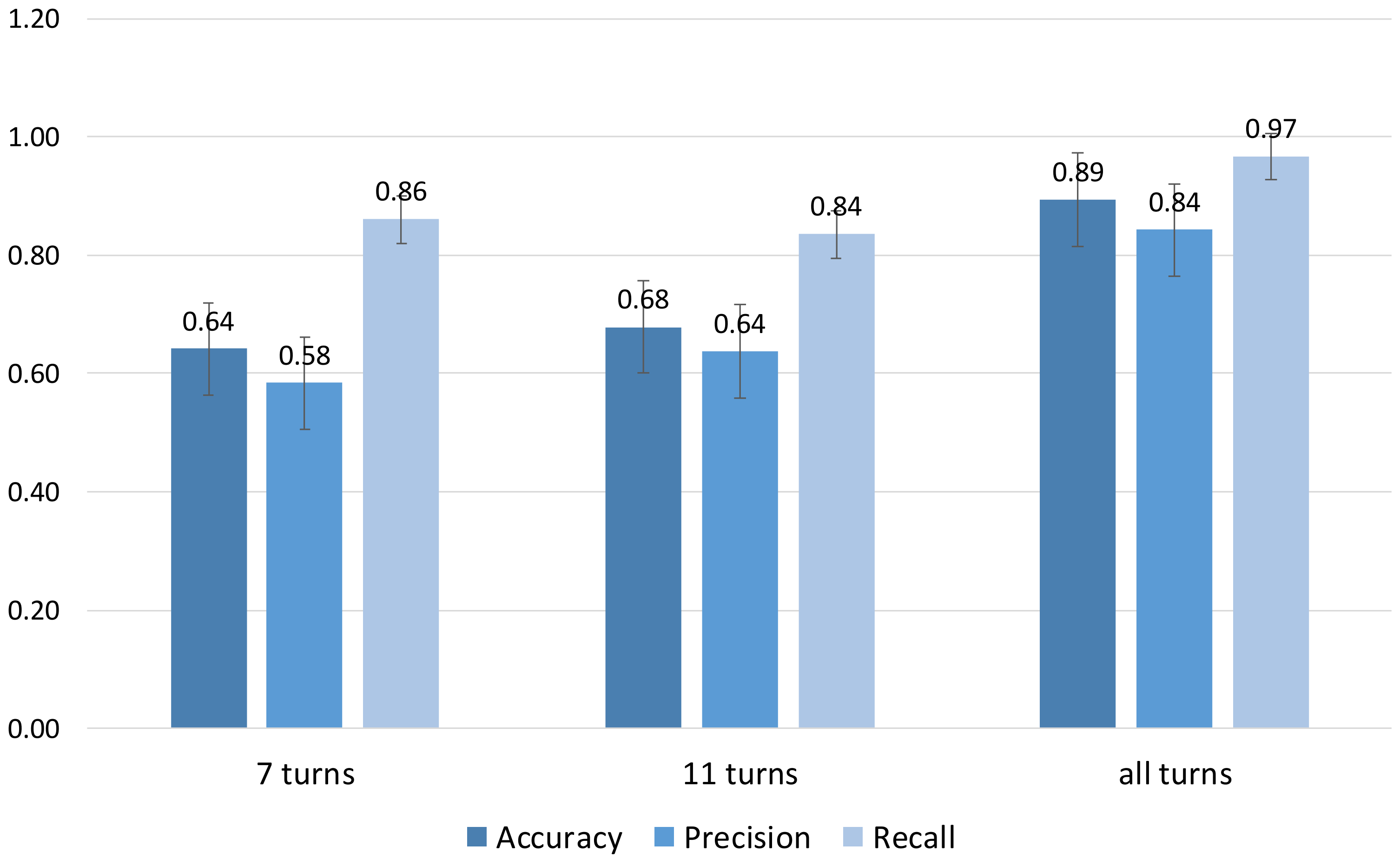
Thank you

Vielen Dank für Ihre Aufmerksamkeit!

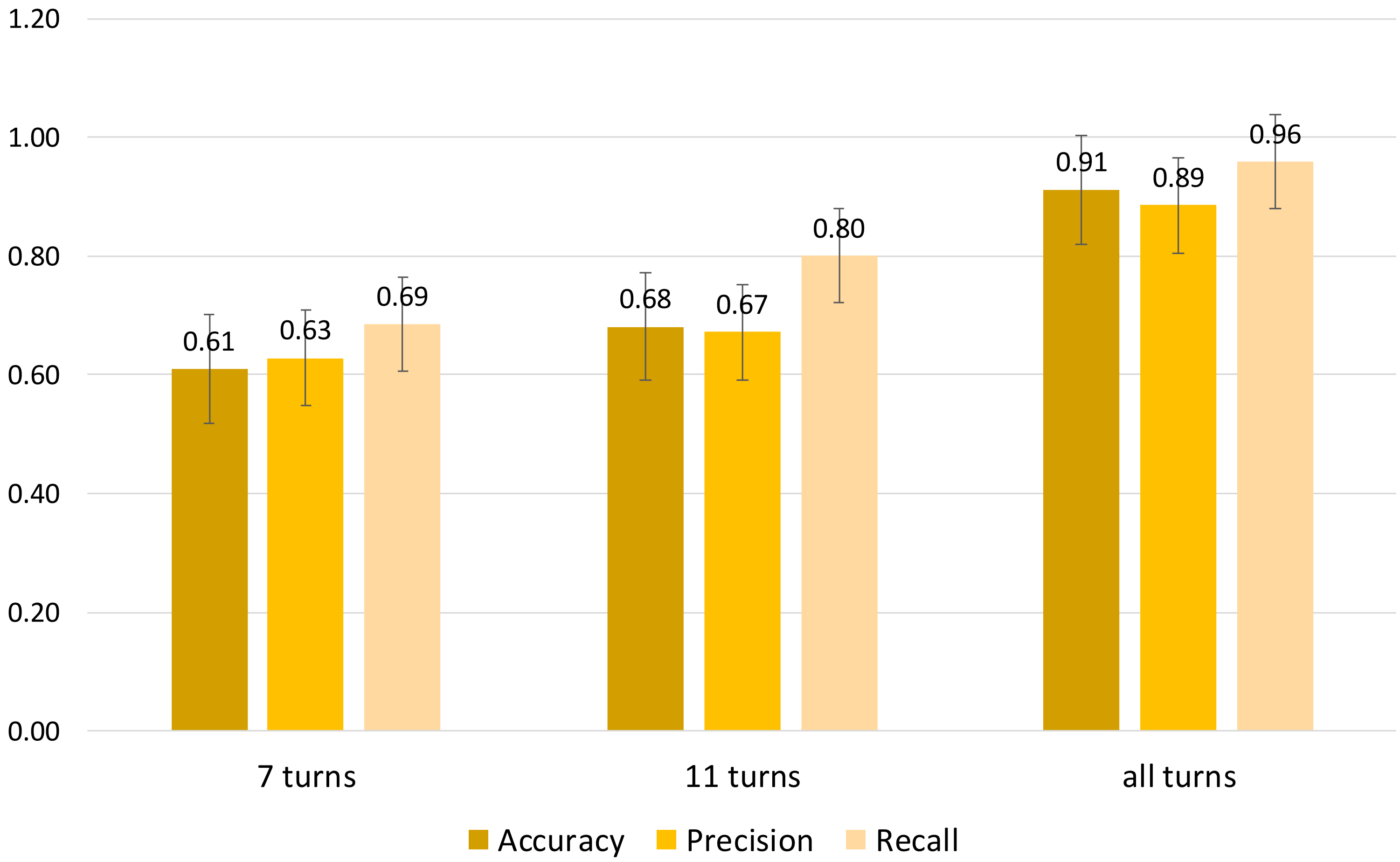
Experiment I, Mistral



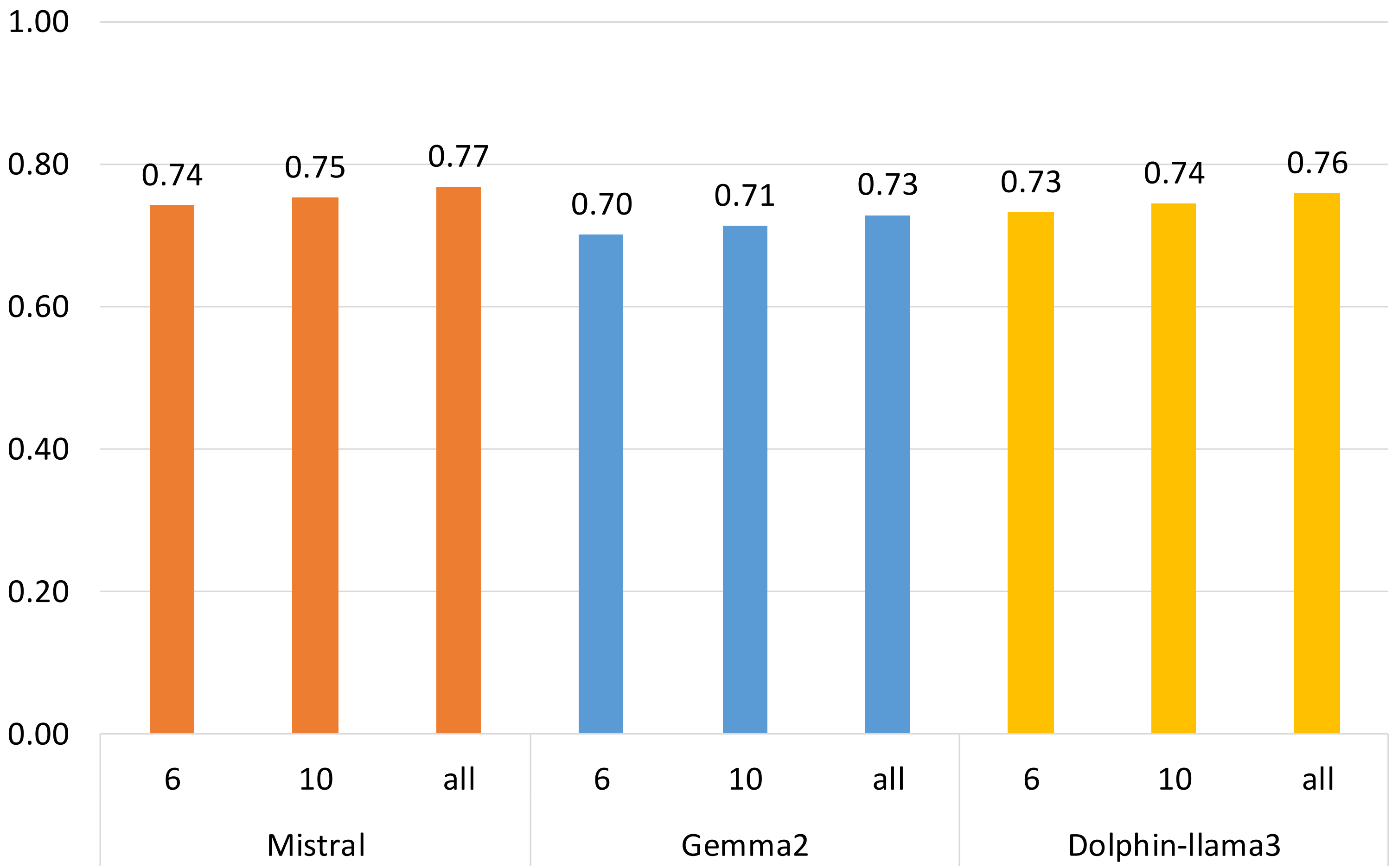
Experiment I, Gemma 2



Experiment I, Dolphin-Llama 3

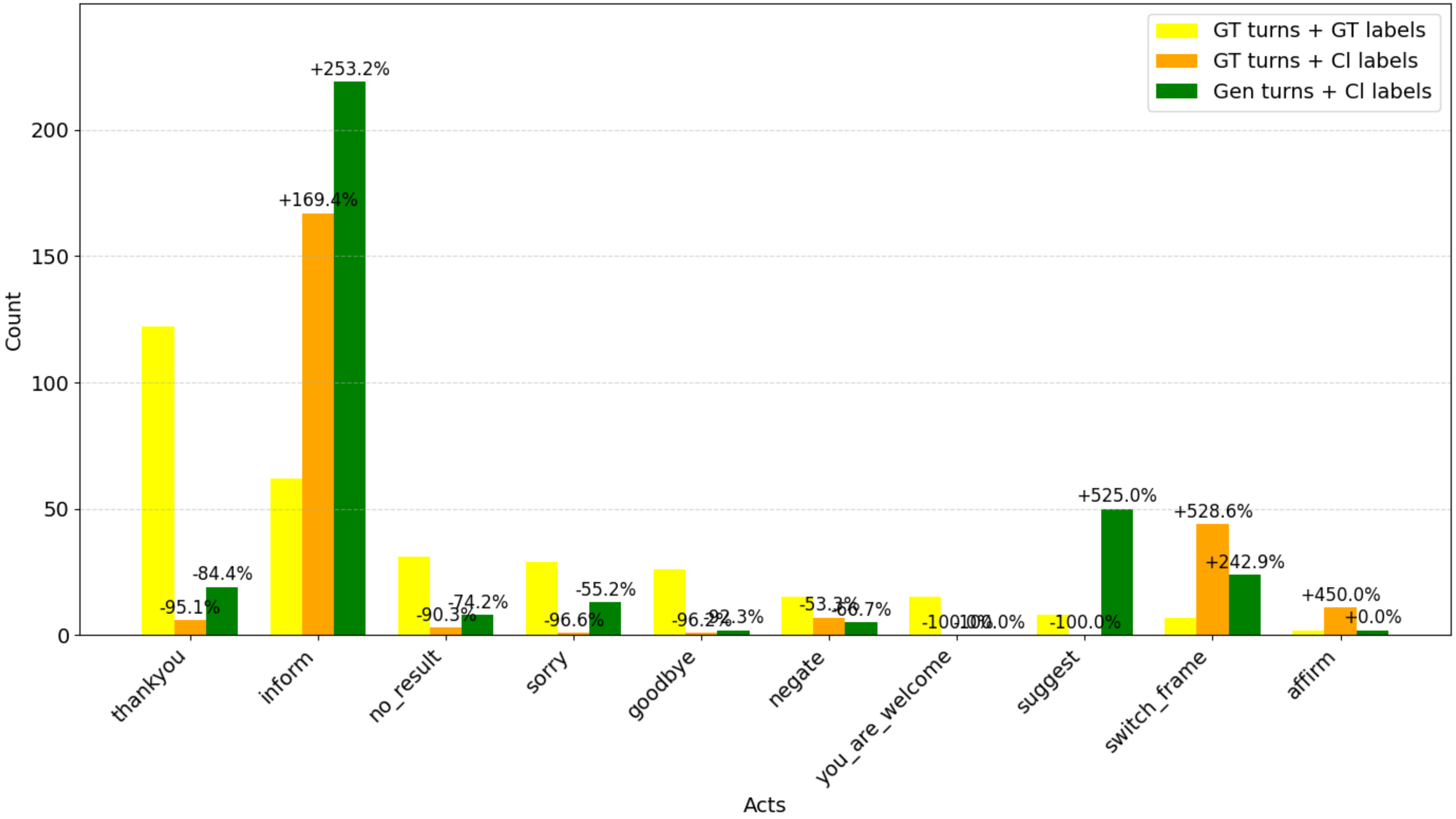


Experiment II, Micro-F1



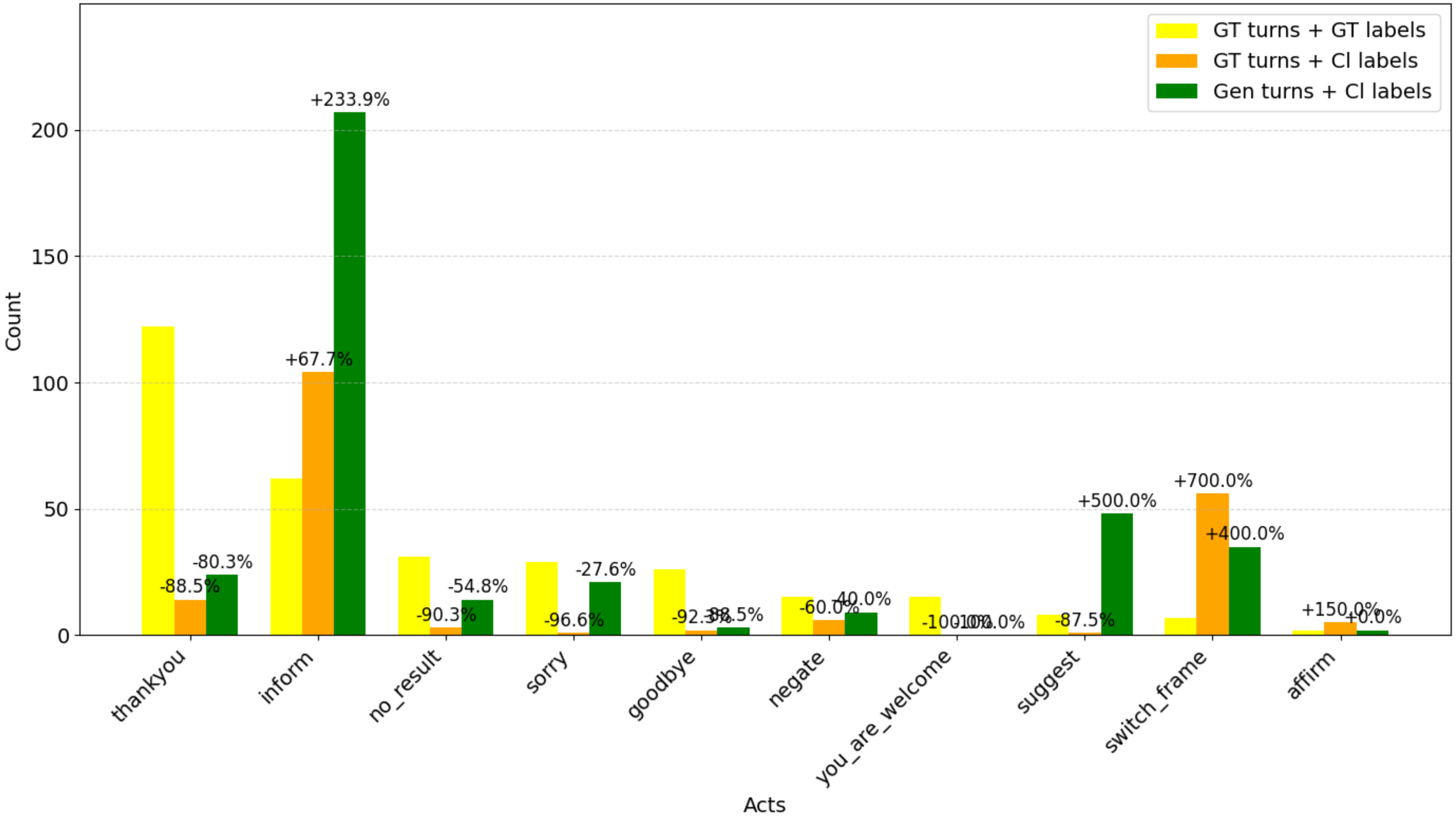
Experiment II Results

Mistral, after 6 turns



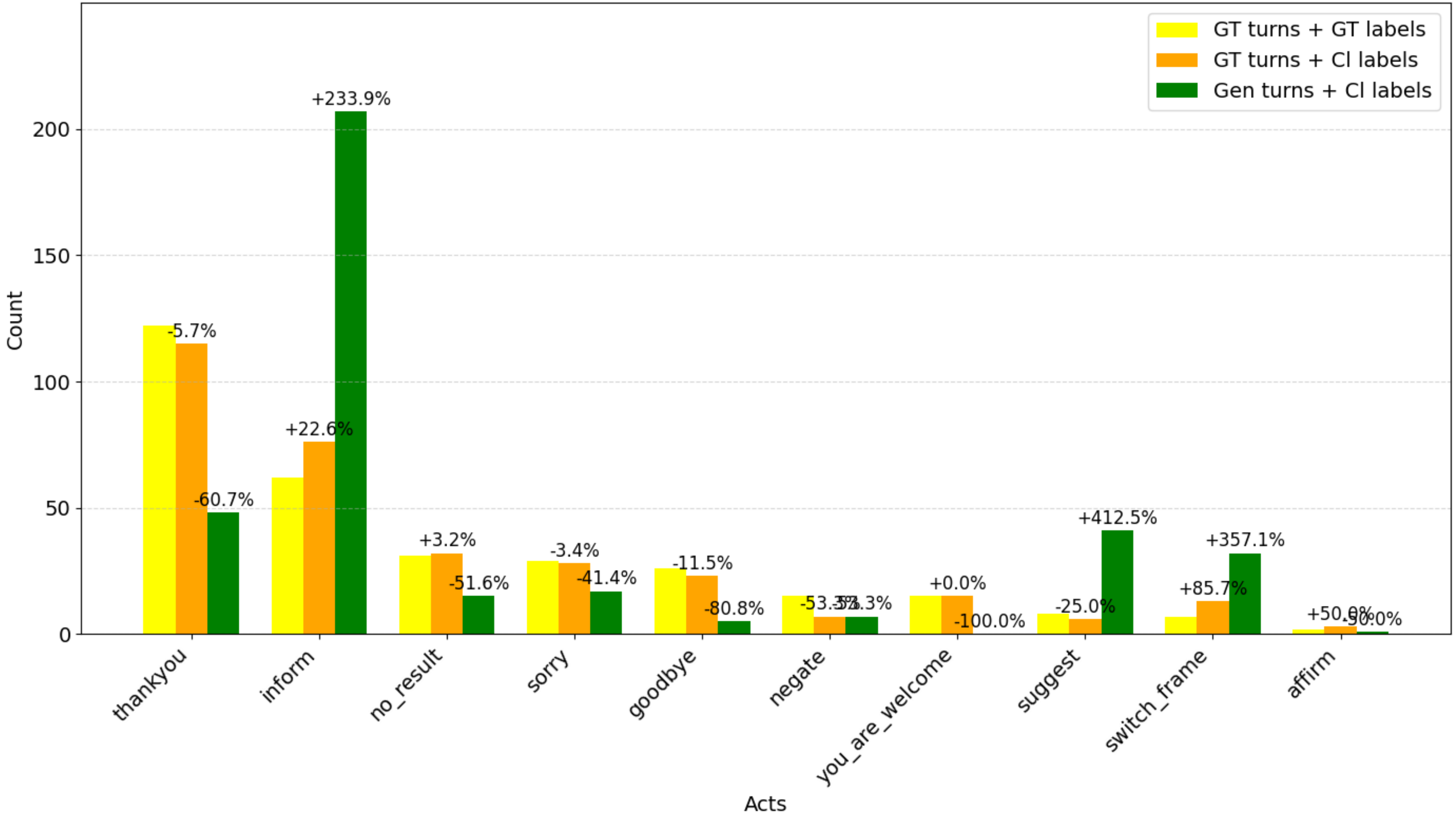
Experiment II Results

Mistral, after 10 turns



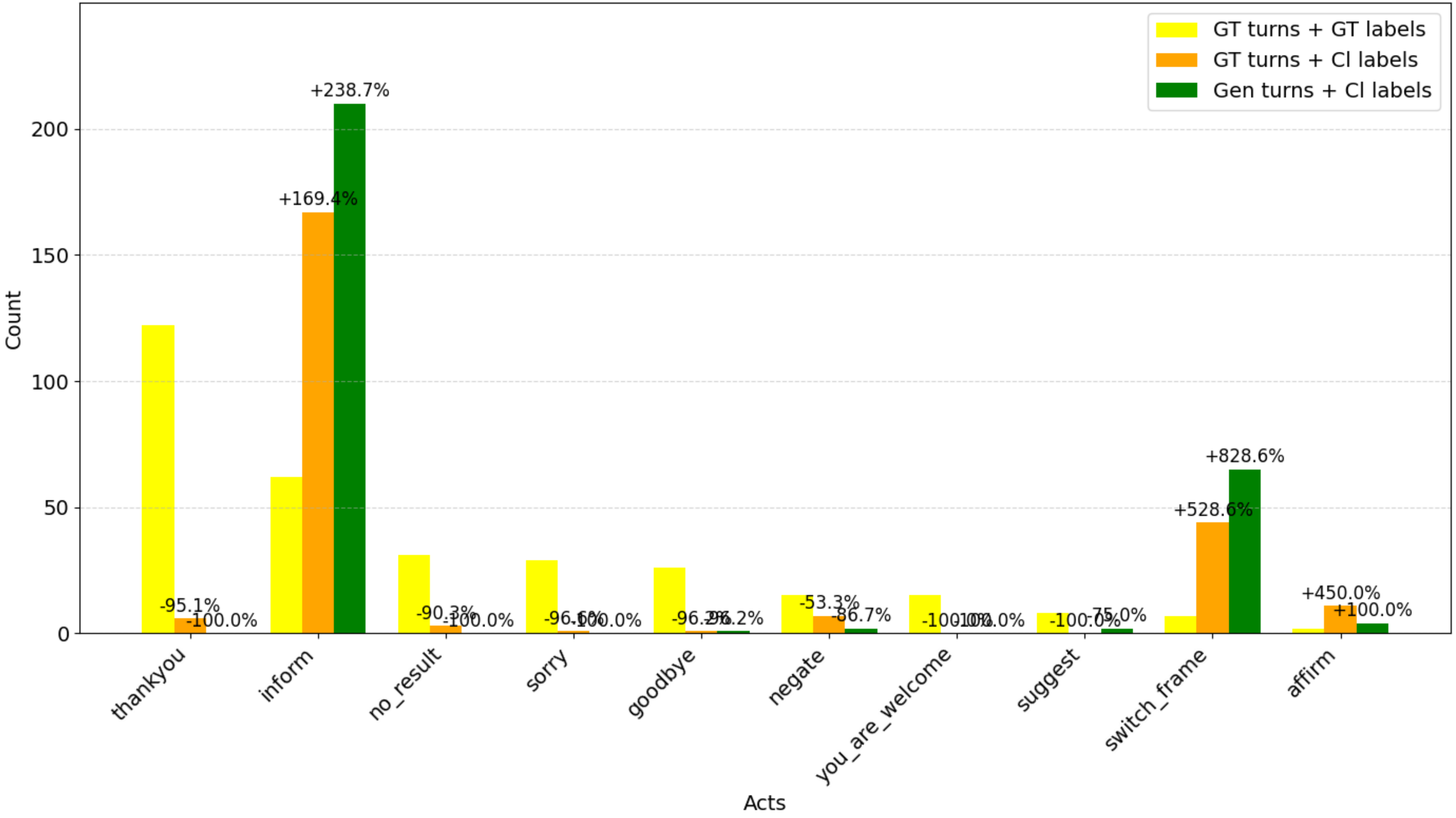
Experiment II Results

Mistral, after all turns



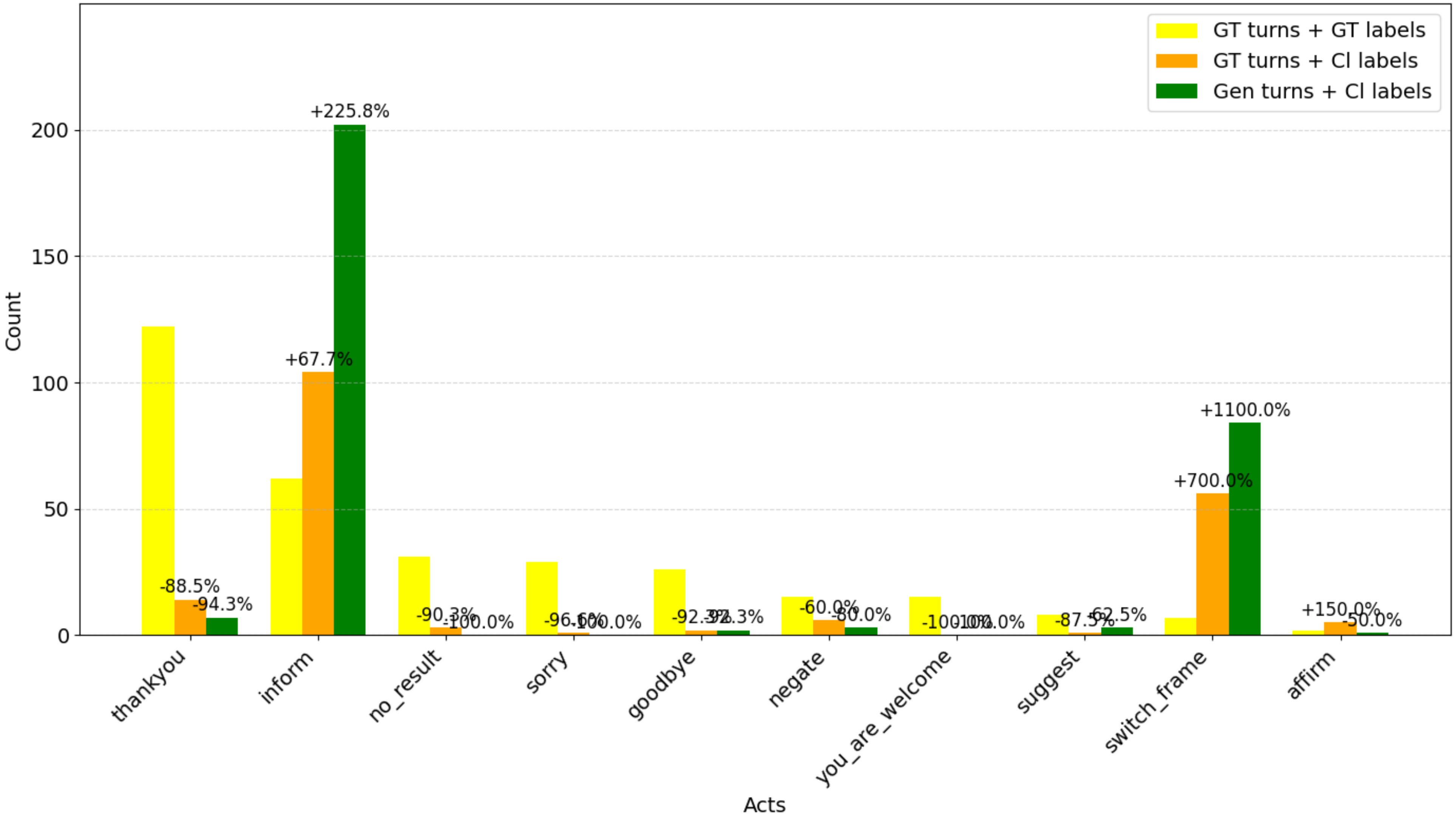
Experiment II Results

Gemma 2, after 6 turns



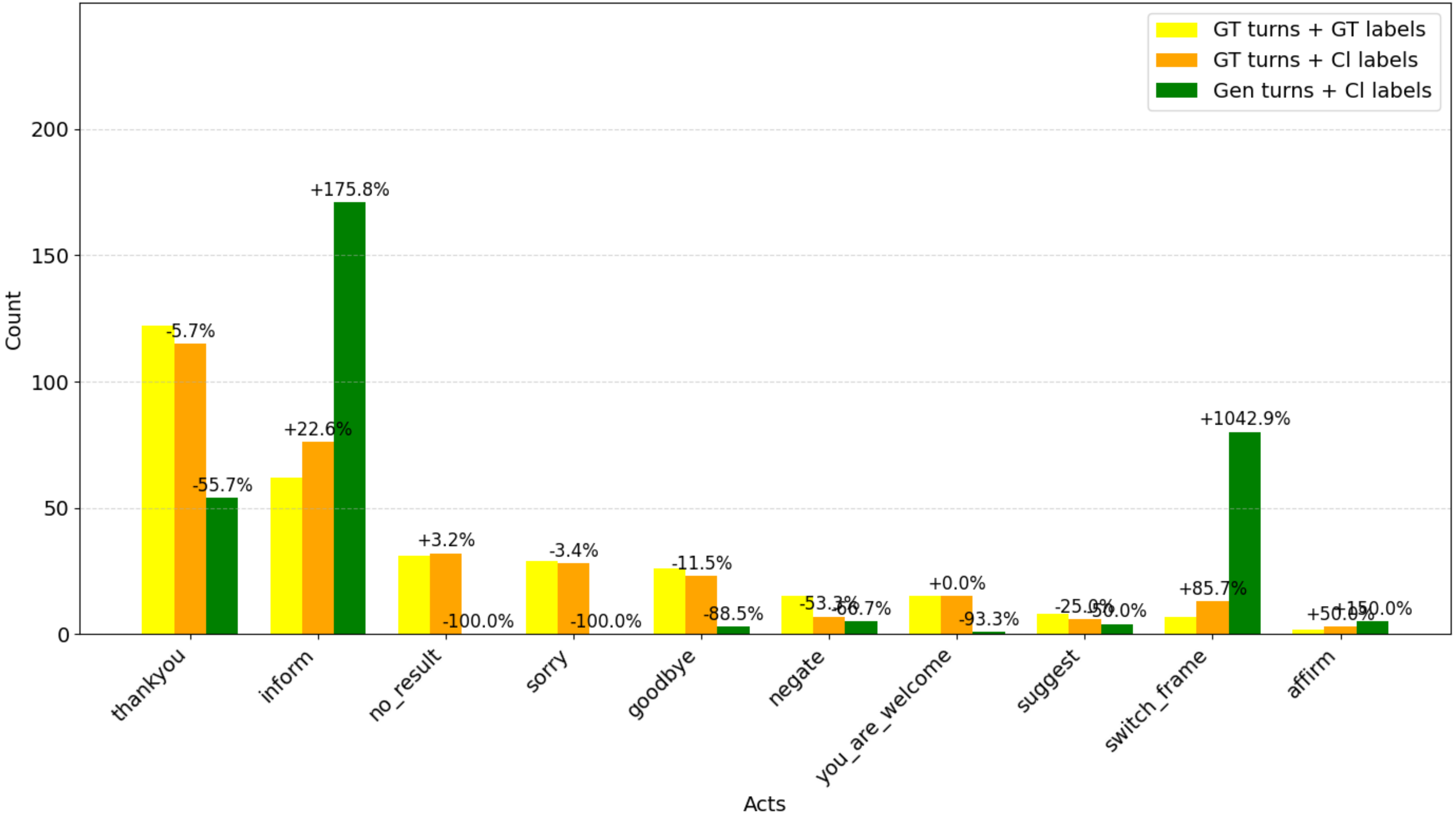
Experiment II Results

Gemma 2, after 10 turns



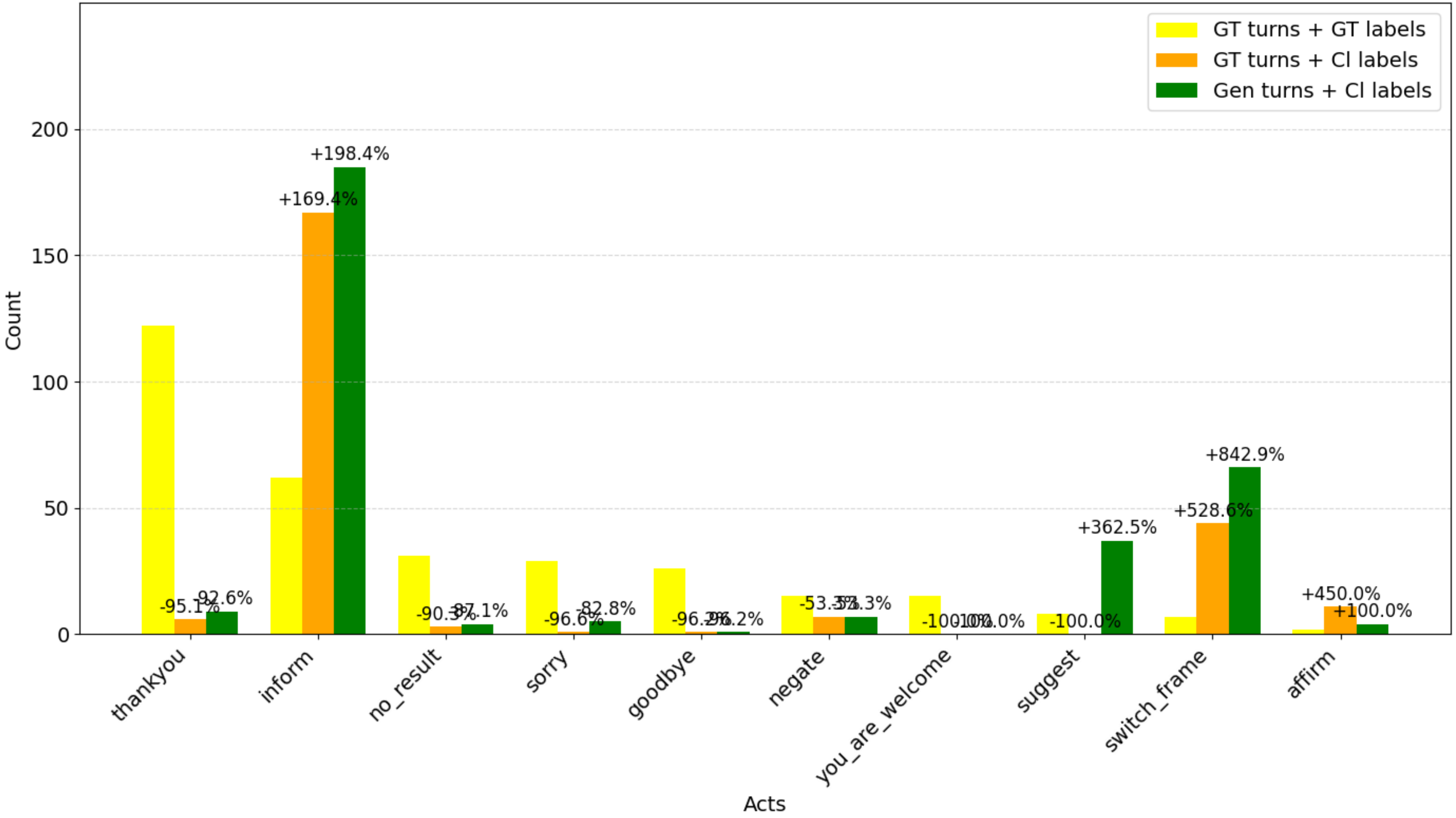
Experiment II Results

Gemma 2, after all turns



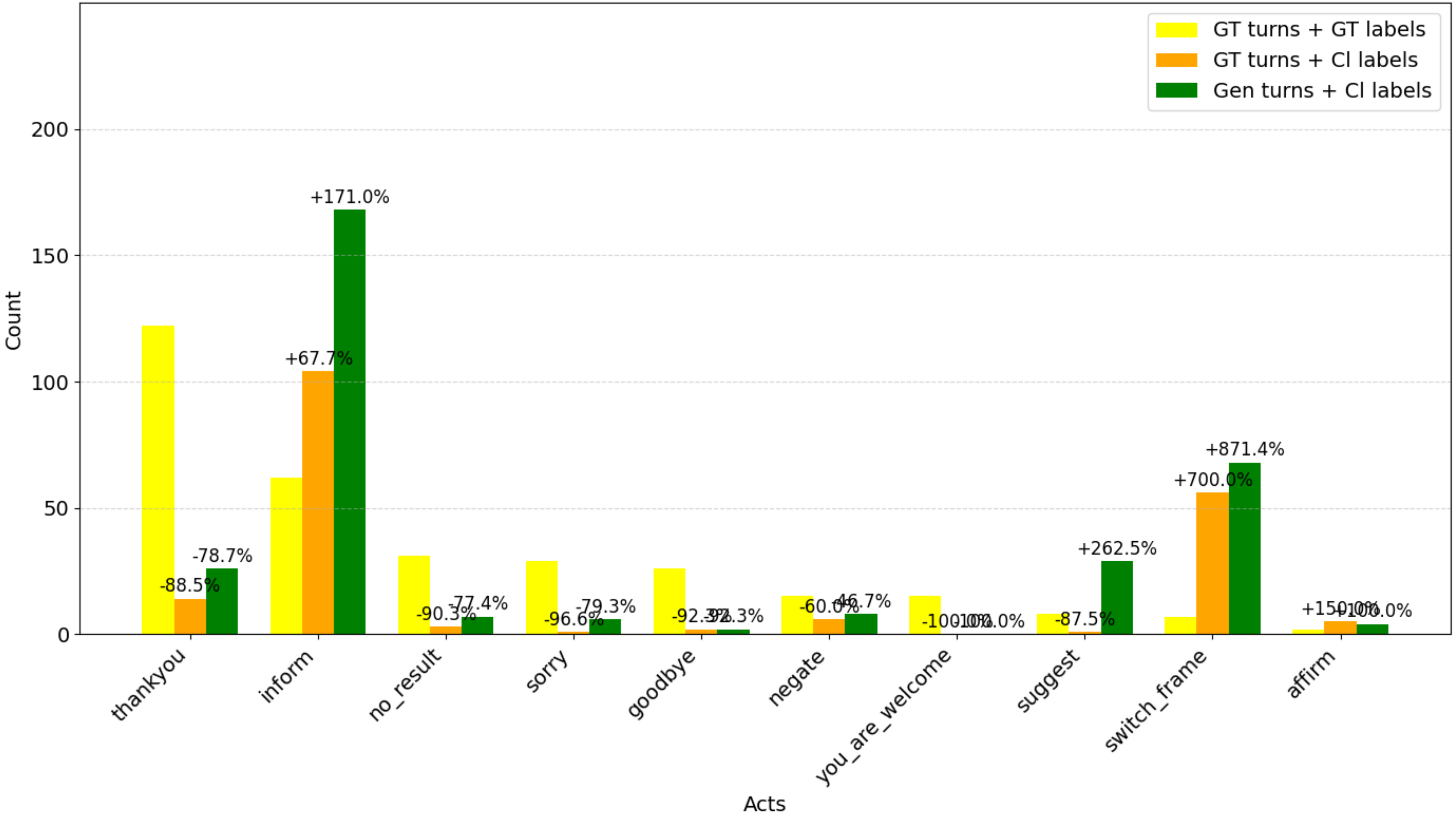
Experiment II Results

Dolphin-Llama 3, after 6 turns



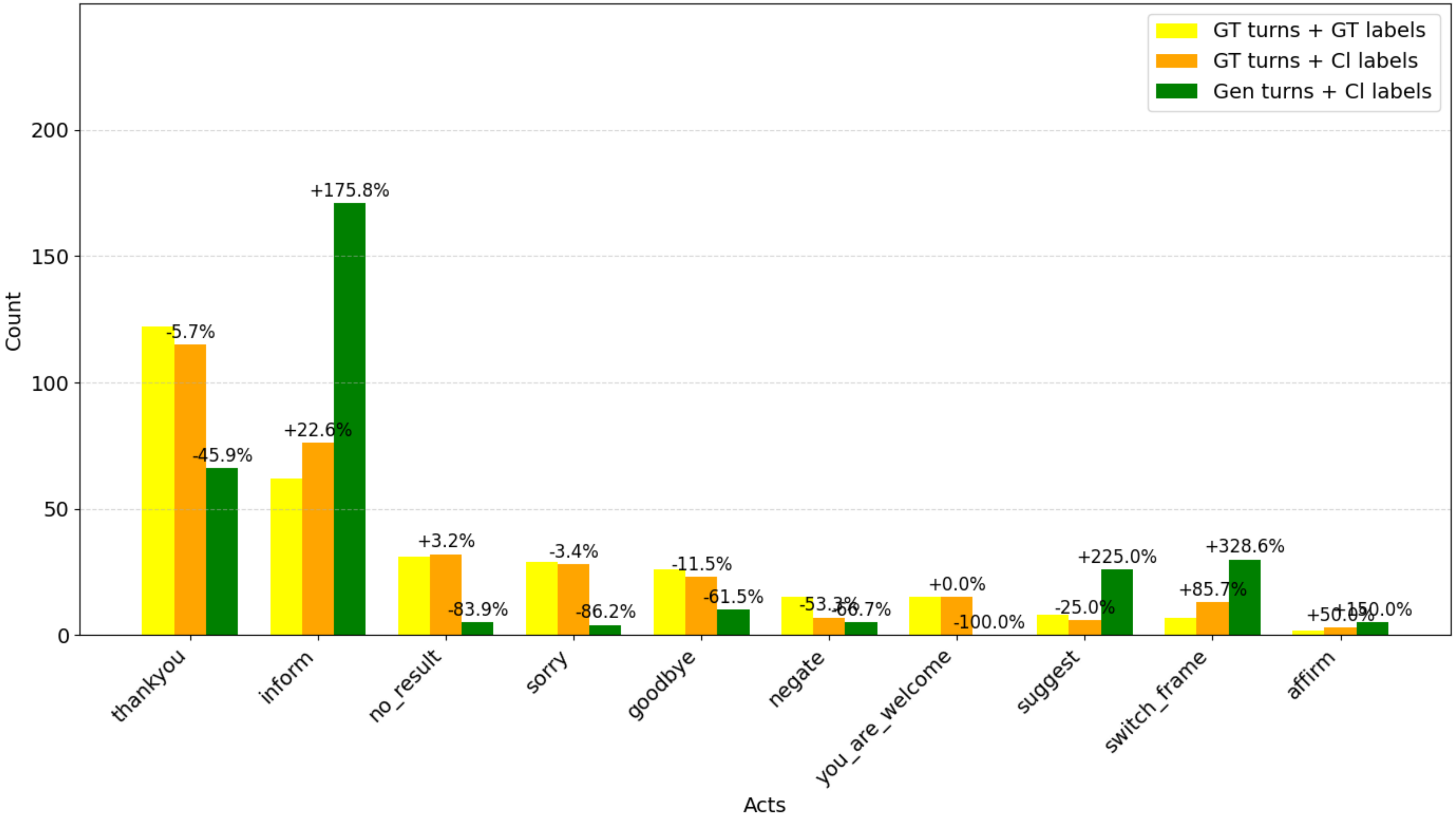
Experiment II Results

Dolphin-Llama 3, after 10 turns



Experiment II Results

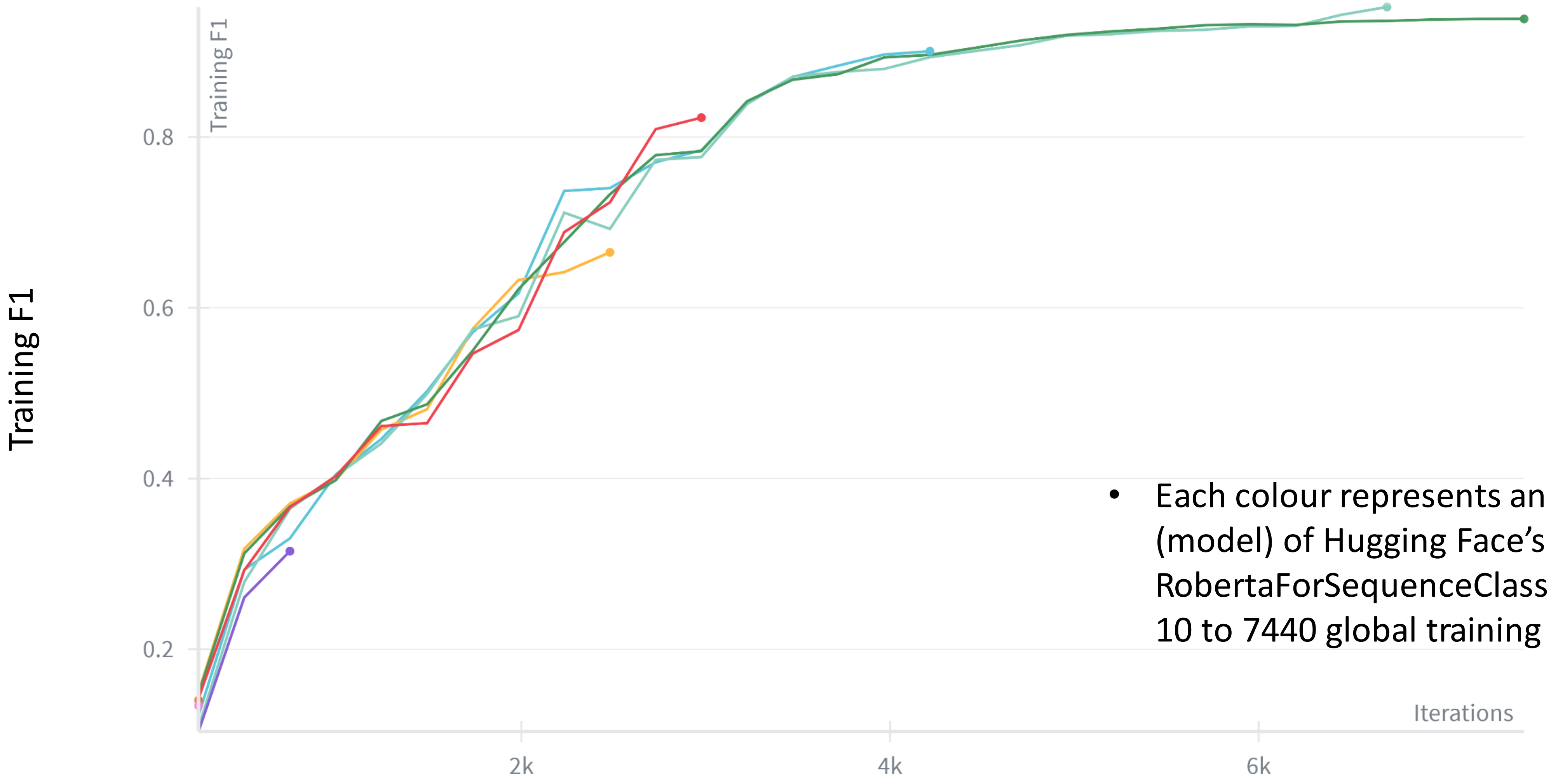
Dolphin-Llama 3, after all turns



Experiment II, Classifier

- Each colour represents an instance (model) of Hugging Face's RobertaForSequenceClassification, from 10 to 7440 global training steps.

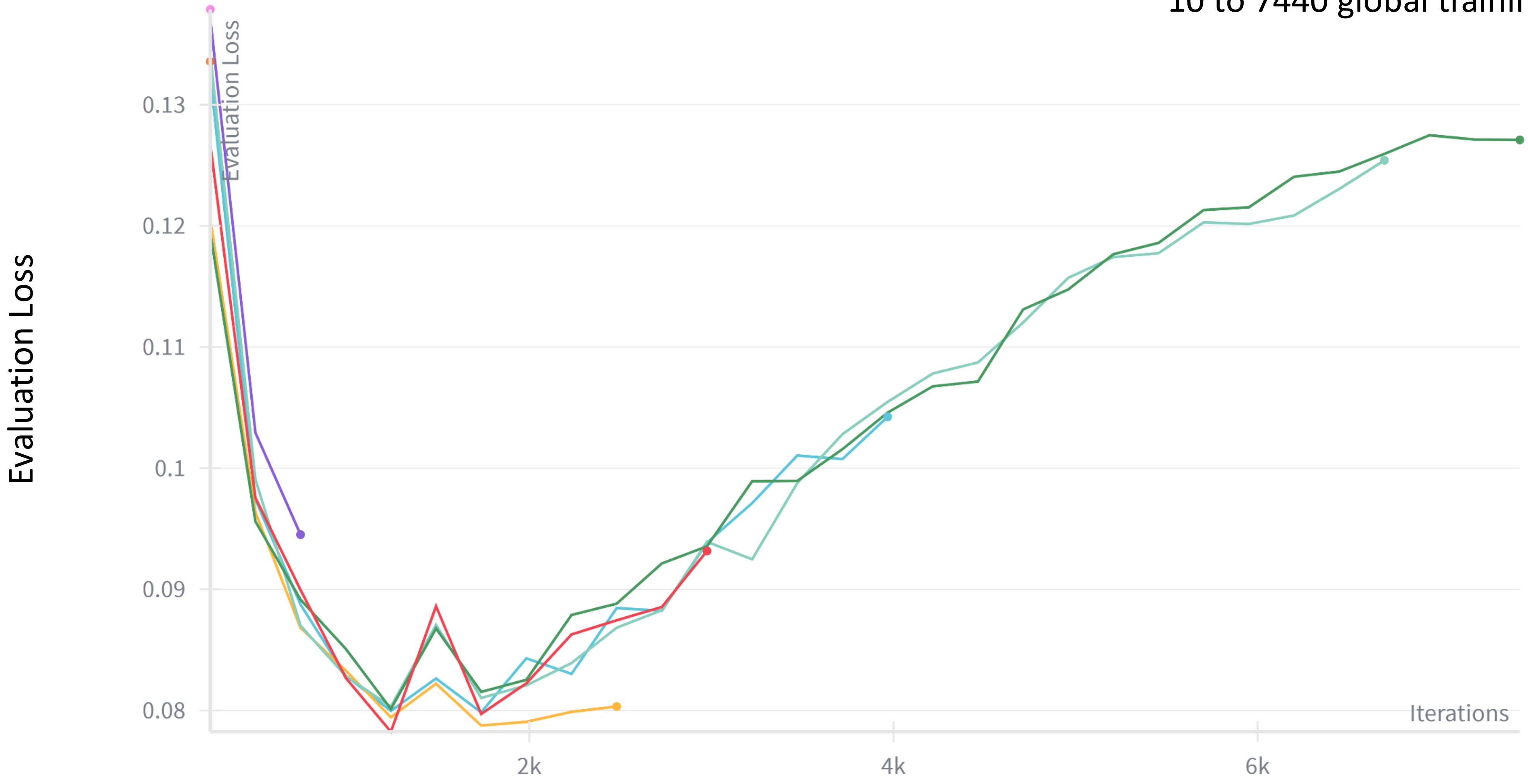
Experiment II, Classifier



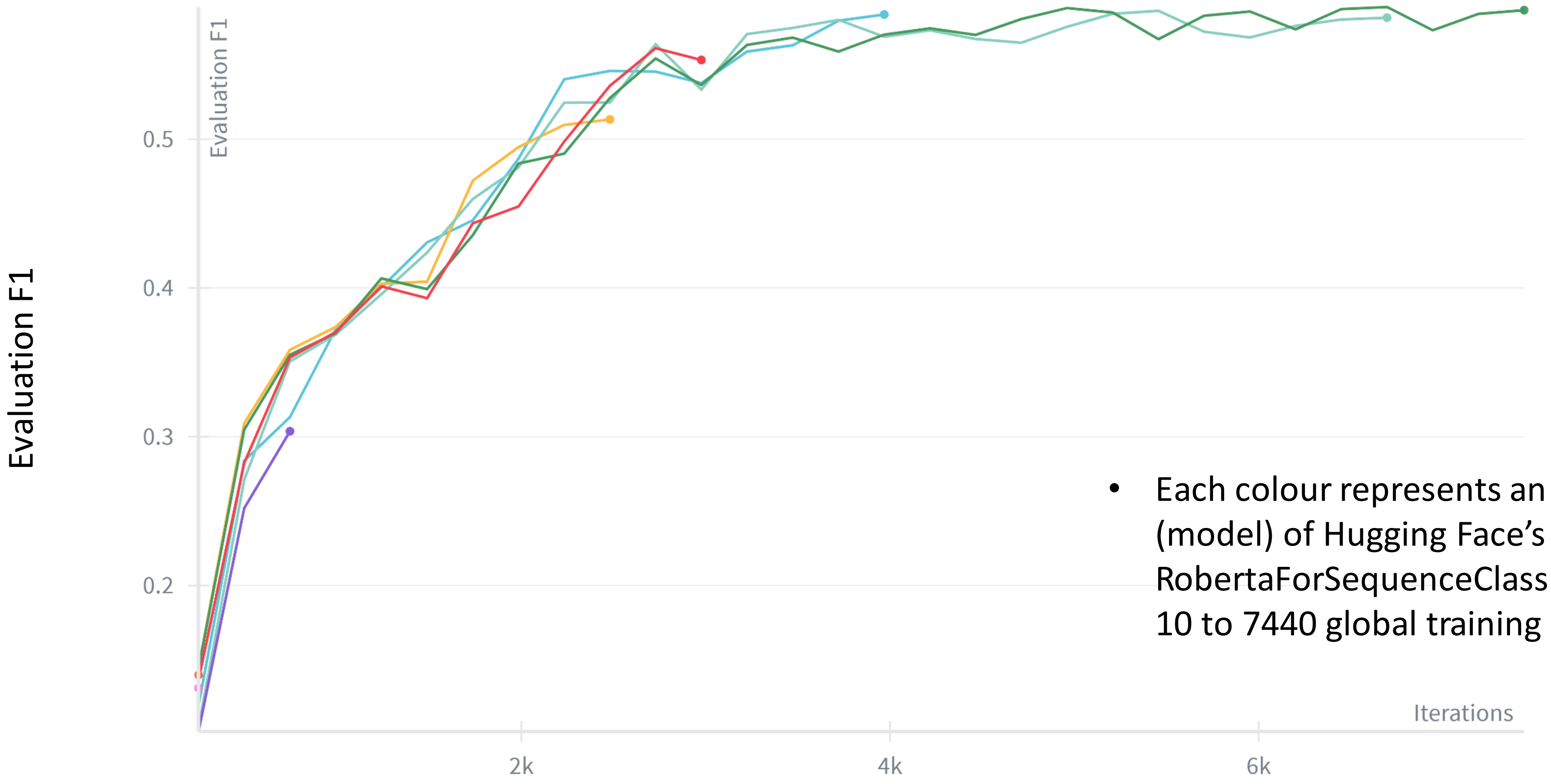
- Each colour represents an instance (model) of Hugging Face's RobertaForSequenceClassification, from 10 to 7440 global training steps.

Experiment II, Classifier

- Each colour represents an instance (model) of Hugging Face's RobertaForSequenceClassification, from 10 to 7440 global training steps.



Experiment II, Classifier



- Each colour represents an instance (model) of Hugging Face's RobertaForSequenceClassification, from 10 to 7440 global training steps.