Linguistic Accommodation in Corpora and Large Language Models

Antonio Innocenti, Agnieszka Gwizdek and Susanne Schmalwieser

universität wien

Project Aim

To investigate Linguistic Accommodation in Large Language Models:

- Find the best corpora and compare them
- Find the most suitable models & prompts
- Investigate text similarity via authorship attribution and singular stylistic features

Stylistic Features

Looking at words' *functions* instead of *meanings*

Key features analysed:

Length of utterances (in characters), Proper noun mentions (=names/key person or place mentions), Novelty tokens, Use of hedge words, Part of speech usage

Data Understanding & Preparation

Initially, 16 different text corpora were converted for the analysis. Based on a set of predefined criteria, we selected three corpora to perform further analysis on linguistic accommodation: the ***DailyDialog*** dataset, the ***Cornell Movie-Dialogs Corpus***, and the ***NPR Interview 2P Dataset Corpus***.

Selection Criteria for Corpora:

- corpus size,
- text quality,
- conversation coherence,
- and publication credibility.

Key Preprocessing Decisions:

- Conversations that involve only two speakers.
- Conversations with at least five turns.
- Speaker roles were assigned such that user X corresponds to odd-numbered turns and user Y to even-numbered turns.
- Starting from the 5th turn, user X is replaced by the LLM.

Modeling

Models used:

Gemma family:

- Gemma 2-2b-it
- Gemma 2-9b-it

Meta Llama family:

- Llama 3.2-3b-Instruct
- Llama 3.1-8b-Instruct

Selection criteria for LLMs:

- Task: text generation.
- Popularity on Hugging Face.
- Size of the models: "tiny" vs larger models for comparison.
- Instruction tuned models.

Framework:

Environment: Google Colab
Language: Python
Initial Model: Gemma 2-2b-it,
Initial Corpus: Daily Dialog (train/dev/test sets

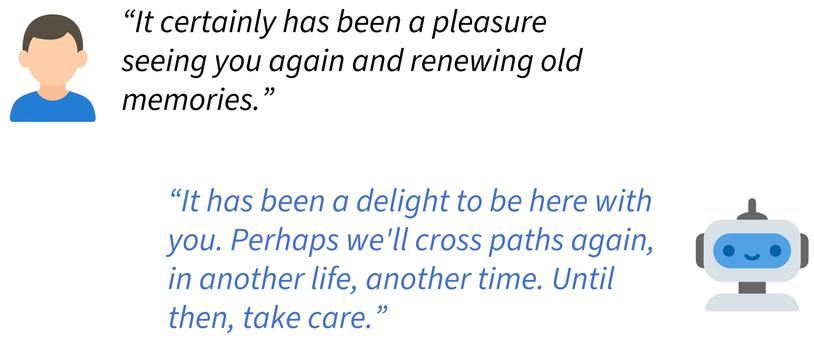
Steps performed:

1. Train set - **Prompting**: Choose a way to provide conversation history (context) to the LLM.
2. Dev set - **Hyperparameter tuning:** Choose the parameters for generation of output (temperature, top-k, top-p)
3. Test set - **Text Generation**: Apply the best strategy to the rest of models and corpora.

Evaluation process:

Comparing regenerated utterances from the LLM with original conversations:

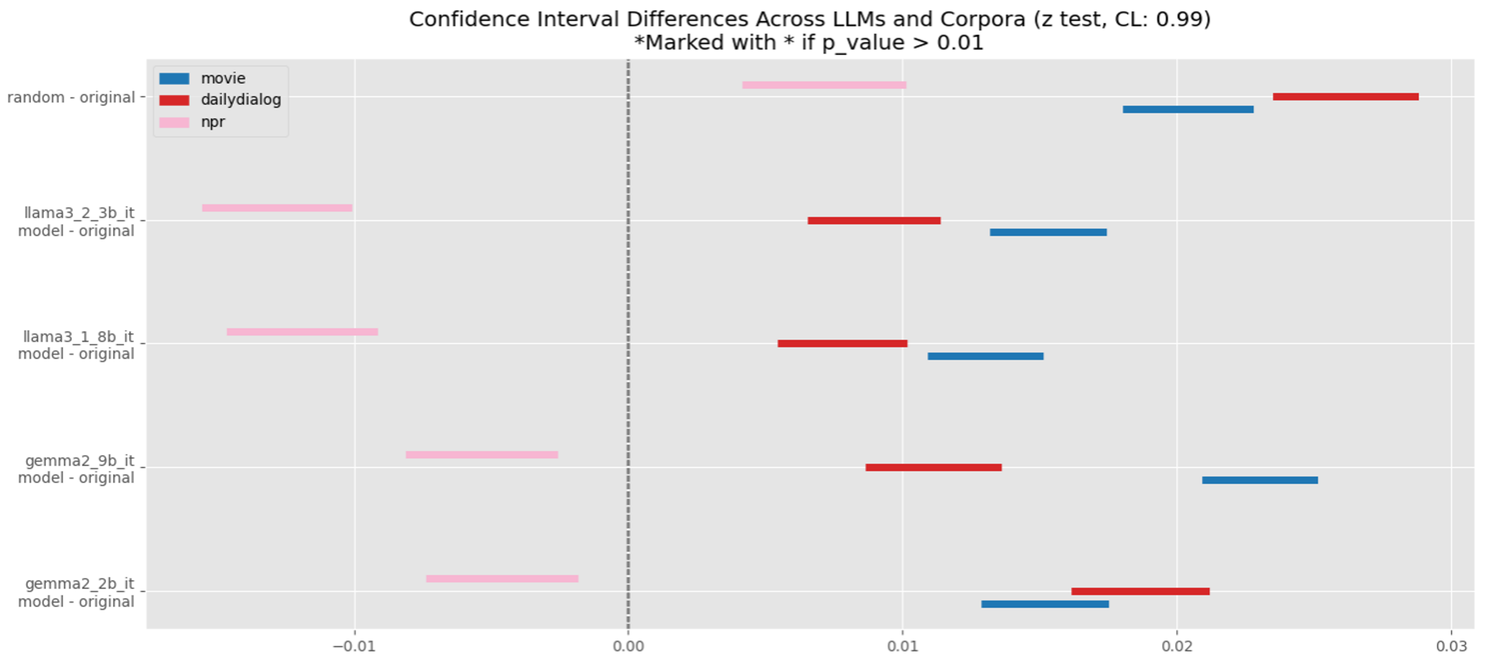
- quantitatively: GLEU, STS (Semantic Textual Similarity), ROUGE, BERTScore F1.
- qualitatively: manual inspection of the model's outputs.



Results I

Results from the AdHominem authorship model:

1. The degree of accommodation varies depends on both the language model and the dataset.
2. Some LLMs, especially those in the LLaMA family, mimic user input more closely than others.
3. Model size did not consistently predict accommodation behavior.
4. Within the Gemma models, performance varied with size, but the LLaMA 3-2.3B and LLaMA 3-1.8B models showed very similar behavior.
5. Factors beyond scale may significantly influence linguistic alignment.



Results II

Main findings for Stylistics:

1. For various features, accommodation can be observed.
2. The models exhibit less linguistic „creativity" than natural users.
3. The models adapt regarding the specific tokens used (links to „content words" used by user Y).
4. We can observe certain „AI model habits" regarding speech production.
5. Bigger accommodation might not necessarily mean more natural sounding speech.