Play **Y**our **P**art

Improving LLMs Script Adherence and Consistency in Long-Form Interactions

Large Language Models (LLMs)—neural networks trained as auto-regressive generative models on web-scale text datasets—can be prompted to perform various tasks [6], including dialogue, enabling natural, human-like interaction. This has led to their widespread use in chatbots like ChatGPT. **These systems prompt an LLM to role-play an agent** by describing its persona and following a dialogue template, e.g. "`You are a helpful and smart assistant at the service of User <You>Hello, I'm here to assist you<\You><User>...<\User>`".

To facilitate interaction with LLMs and prevent harmful behavior, complex prompts are crafted to shape the persona of the simulated character. For instance, the initial prompt for Snapchat's My AI chatbot included instructions like "`Never have negative opinions or make adversarial judgments on sensitive topics such as politics, religion, (...)`" [1]. Additionally, most LLMs undergo *human preference alignment* (HPA), where they are fine-tuned to increase helpful, harmless outputs and reduce harmful or non-helpful content, as defined by human evaluators [16]. However, due to their inherent nature, **LLMs are difficult to control**, which reduces trust in their use —particularly in sensitive or high-risk scenarios— since **they can unpredictably deviate from the intended "script"**. Such deviations may occur due to hallucinations or shifts in their behavior caused by altered instructions. For example, LLMs can be prompted—intentionally or unintentionally—to bypass initial instructions and exhibit unwanted behaviors, a process called *jailbreaking*. This issue is accentuated in long-form interaction, with empirical evidence and theoretical arguments showing that long contexts result in reduced controllability through initial instructions [22].

This project aims to address the issue of consistency and controllability in LLM agents within the challenging context of long-form interactions. We propose a dual-pronged approach. Firstly, we will explore metrics to identify and quantify deviations from desired behavior, along with the necessary evaluation sets to measure these metrics effectively. Secondly, we will delve into mitigating such deviations through the development of improved control techniques. Our methods will be based on gaining a deeper understanding of the mechanisms underlying role-playing and jailbreaking through modern mechanistic interpretability techniques, and the analysis of interaction dynamics using a model-based approach. Two applications involving long-form interaction and of significant practical relevance—multi-turn task-oriented dialogues and the simulation of doctor-patient interactions with diverse personas—will inform the design of our methods and serve as testbeds for their evaluation.

Research axes

Agent interpretability and control

The control methods described above treat LLMs as black boxes, aiming to manage them by conditioning through input (*prompting*) or fine-tuning to achieve desired outputs (*HPA*), without addressing the mechanisms by which they process inputs and generate outputs. Various works suggest that **understanding the underlying mechanisms of LLM behavior can lead to more robust control over the agent's actions**. Specifically, recent findings in *mechanistic interpretability*—a field focused on breaking down neural networks into comprehensible components—indicate that **LLMs encode semantic features with causal effects on behavior as linear directions in their representation space** [5]. This idea, known as the *linear representation hypothesis* (LRH) [18], has informed recent work identifying features in LLMs that govern behaviors such as truthfulness [13] and request refusal [1]. Building on these findings, we aim to address the issue of role consistency by leveraging this understanding. Our approach will involve two key steps: **1) uncovering representation-level mechanisms that underpin role-playing in LLMs**; and **2) using this knowledge to develop methods that reinforce role consistency**.

In the first part, we will operate within the LRH framework, using techniques like dictionary learning [5] and difference-in-means analysis [3] to isolate directions in the embedding space for persona traits (e.g., helpfulness, politeness) and entities (e.g., self, user) defined by role-playing prompts. To understand how roles are implemented and lose consistency, we will draw on insights from the binding problem—how LLMs associate attributes with entities—an area where recent mechanistic interpretability research has advanced notably [8]. In the second part, we will focus on utilizing the mechanisms discovered in the first part to develop methods for controlling LLM role-playing behavior. Specifically, we plan to employ representation-based intervention techniques, such as representation steering [21, 24], to allow for context-independent behavior control.

[1] Example taken from `https://github.com/jujumilk3/leaked-system-prompts`

Modeling, Adherence & Evaluation

Another approach to assess and enforce adherence to a persona or script in LLM multi-turn interactions is through a model of the dialogue. The **modeling of task-oriented dialogues** has been the object of a previous JSALT project in 2023. In particular, one of the concrete results were the Dialog2Flow (D2F) embeddings [7]. These embeddings represent points in a latent space that encodes both utterance semantics and the speaker's communicative intention. By using D2F embeddings, conversations can be modeled as trajectories in a conversational latent space that can be merged and pruned to extract the underlying dialogue flow. We aim to explore how to take advantage of dialogue flows to ground LLMs-based dialogue systems to improve their controllability and interpretability [19]: i) **script adherence scores**: as of today, there is no standard metric(s) to quantify hallucinations and misaligned behavior of conversational dialogues beyond simple word-based metrics such as perplexity, ROUGE or BLEU. We aim to explore how to take advantage of dialogue flows to design evaluation metrics that go beyond word-based towards flow-based metrics that also consider the expected conversational steps within dialogues. This stage will also involve exploring metrics beyond flow-based ones to compare against, including more subjective metrics like human or LLM-based ones. ii) **dialogue flow-grounded LLMs**: recently, a significant effort has been devoted to help LLMs better support the contextualization and abstraction within a conversation, resulting in a richer language model encoding in the embedding space [23, 9]. As a possible extension to this work we consider the use of the audio or multimodal D2F approach, i.e., to operate directly on speech or multimodal (audio-textual) inputs.

Synthetic Data Generation

As a flagship use-case of this work we propose **generating faithful synthetic data by creating realistic doctor-patient conversations using fictional, plausible personas, electronic medical records (EMRs), or clinical notes as sources**. Our approach covers the entire pipeline, from creating a dialogue [12, 2] based on personas e.g. from EMRs, to generating faithful multichannel audio that mimics the examination room acoustics, including non-verbal noises like typing on a keyboard or the patient coughing/panting (based on the EMR indicating these) [14]. We aim to develop a modular framework for generating multi-channel audio that can be customized for various conditions. We will also create a set of faithfulness metrics to evaluate the quality of the generated dialogues across multiple dimensions, including naturalness, room acoustics, and non-verbal cues. Depending on progress made during the workshop, we plan to expand support for a wider range of acoustic variability, such as differences in age or accent. Research questions that we want to address are: i) how to generate doctor-patient conversations effectively, while **including the influence of medical history, personality traits, and underlying health conditions**?, ii) what are effective techniques for generating **realistic and diverse synthetic doctor-patient conversations**?

Resources

Datasets. Interpretability methods require input samples that clearly exhibit presence (or absence) of the features that we attempt to discover. Since most datasets will not contain the required annotations for our features, most of our data for interpretability will be synthesized using LLMs. Prior work on the task-oriented dialogues produced a unified and standardized dataset introduced in Burdisso, Madikeri, and Motlicek [7] and the flowchart-based dataset [19] expanded by using instruction-tuned LLMs. These will serve as the foundations for the dialogue modeling approach. For the definition of doctor-patient personas and synthesis, the conversational datasets we plan to rely on are PriMock57 [17], MediQA-Chat 2023 [4], and MIMIC-III [10].

Models. For data generation and as object of study for interpretability we plan to use state-of-the-art open-source LLMs like Llama 3 [20], for which dictionary models and datasets with interpreted features are also available[2]. The speech synthesis of doctor-patient conversations will leverage tried-and-true toolkits such as ESPnet and K2.

Benchmarks. Task-oriented dialogues and persona consistency in patient-doctor interaction simulation will be our main practical benchmarks. Initially, we also intend to use open jailbreak benchmarks (e.h. HarmBench [15]) in the evaluation of the *Agent interpretability and control* axis. We also **propose Sequential Social Dilemma (SSD) games [11] as a scalable artificial task to systematically elicit and evaluate role misalignment in long-form interaction**. Due to their game-theoretical rational incentives for misbehavior, we believe SSDs present a challenging setup to evaluate agents' consistency and alignment.

Organization

Team. The members of the team are organized in three sub-teams corresponding to our main research axes, and are presented in Table 1. Undergraduate students will be added to the team upon application.

Work-plan. A preliminary work timeline is presented on the next page on a monthly basis for pre-workshop tasks and on a weekly basis for the duration of the workshop. Several of these tasks are conditional to the developments seen throughout the project and are therefore likely to change.

[2] `https://huggingface.co/[EleutherAI/sae-llama-3.1-8b-64x,datasets/EleutherAI/auto_interp_explanations]`

Full-Time Members		Part-Time/Remote Members	
Santiago Cuervo*	Antonio Almúdevar*	Milos Cernak	Esau Villatoro
Adel Moumen*	Ricard Marxer	Markus Müller	Detlef Koll
Petr Motlicek	Sergio Burdisso	Michael White	Reed van Deusen MD
Srikanth Madikeri	Thomas Schaaf	Adam Rothschild MD	Alfonso Ortega
Amy Chun*	Andrew Perrault		
Tomiris Kaumenova*	3 grad student TBD		

Table 1: Project Team Structure. * indicates graduate students, not present indicates senior members.

Timeline of pre-workshop and during workshop actions

M denotes months and **W** denotes weeks.

Pre workshop

M1 Open data collection and preparation (e.g., HarmBench, PriMock57, MIMIC-III).

M2 Setting up repository for LLM inference and set of prompts for data generation. Prepare SSDs environment for interaction with LLMs. Setup repository for linear feature extraction and automated interpretability methods.

M3 Synthesis of initial test datasets for interpretability and patient-doctor persona simulations.

M4 Initial meetings. Team-wide reading group. Initial experiments to validate the binding vectors finding from Feng and Steinhardt [8] with our synthesized data.

M5 Final pre-workshop meetings. Team-wide reading group. Initial exploratory experiments to familiarize new members with the codebase.

During the workshop

W1 – W2 Search for interpretable agent features. Representation steering evaluation on synthetic data. Task-oriented dialogues adherence scoring using Dialog2Flow. Refinement of prompts for doctor-patient conversations.

W3 Validation of found features in single-turn benchmarks (e.g. persona attacks in HarmBench). Possible search for features specific to the benchmarks. Flow-based grounding of task-oriented dialogue LLMs. Study of factors of variation in personas for doctor-patient conversation synthesis.

W4 – W5 Validation of found features in task-oriented dialogues, persona-based simulation of doctor-patient dialogues, and SSDs. Possible search for features specific to these tasks. Assessment of multi-agent setups in doctor-patient conversation synthesis. Development of evaluations for synthetic conversational data.

W6 Final evaluations. Prepare results and conclusions for final presentation.

Expected contributions

Scientific contributions:

- The groundwork towards a theoretical framework for representation-level mechanisms controlling persona simulation and jailbreaking in LLM agents.
- A study on the elicitation of environment-driven temporally extended jailbreaks in LLM agents through game-theoretical rational incentives for misbehavior with SSDs.

Practical deliverables:

- A benchmark for evaluating role consistency in long-form interactions, including the tasks of task-oriented dialogues with dialog flow references, simulation of personas in doctor-patient interactions, and SSDs.
- Development of tools and metrics for evaluating script adherence and consistency in LLM-based dialogues (e.g., flow-based adherence metrics).
- Extendable open source framework for generating synthetic doctor-patient conversations with multi-channel audio, including non-verbal cues like keyboard typing or coughing.

References

[1] Andy Arditi et al. *Refusal in Language Models Is Mediated by a Single Direction*. 2024. arXiv: 2406.11717 [cs.LG]. URL: https://arxiv.org/abs/2406.11717.

[2] Pulkit Arya et al. "Bootstrapping a Conversational Guide for Colonoscopy Prep". In: *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Ed. by Svetlana Stoyanchev et al. Prague, Czechia: Association for Computational Linguistics, Sept. 2023, pp. 413–420. DOI: 10.18653/v1/2023.sigdial-1.38. URL: https://aclanthology.org/2023.sigdial-1.38.

[3] Nora Belrose. *Diff-in-means concept editing is worst-case optimal: Explaining a result by Sam Marks and Max Tegmark*. Accessed on: October 12, 2024. 2023. URL: https://blog.eleuther.ai/diff-in-means/.

[4] Asma Ben Abacha et al. "Overview of the MEDIQA-Chat 2023 Shared Tasks on the Summarization & Generation of Doctor-Patient Conversations". In: *Proceedings of the 5th Clinical Natural Language Processing Workshop*. Ed. by Tristan Naumann et al. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 503–513. DOI: 10.18653/v1/2023.clinicalnlp-1.52. URL: https://aclanthology.org/2023.clinicalnlp-1.52.

[5] Trenton Bricken et al. "Towards Monosemanticity: Decomposing Language Models With Dictionary Learning". In: *Transformer Circuits Thread* (2023). https://transformer-circuits.pub/2023/monosemantic-features/index.html.

[6] Tom Brown et al. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[7] Sergio Burdisso, Srikanth Madikeri, and Petr Motlicek. "Dialog2Flow: Pre-training Soft-Contrastive Action-Driven Sentence Embeddings for Automatic Dialog Flow Extraction". In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 5421–5440. URL: https://aclanthology.org/2024.emnlp-main.310.

[8] Jiahai Feng and Jacob Steinhardt. "How do Language Models Bind Entities in Context?" In: *International Conference on Learning Representations (ICLR)*. 2024. URL: https://iclr.cc/virtual/2024/poster/17378.

[9] Tingchen Fu et al. "Learning towards conversational AI: A survey". In: *AI Open* 3 (2022), pp. 14–28. ISSN: 2666-6510. DOI: https://doi.org/10.1016/j.aiopen.2022.02.001. URL: https://www.sciencedirect.com/science/article/pii/S2666651022000079.

[10] Alistair EW Johnson et al. "MIMIC-III, a freely accessible critical care database". In: *Scientific data* 3.1 (2016), pp. 1–9.

[11] Joel Z. Leibo et al. "Multi-agent Reinforcement Learning in Sequential Social Dilemmas". In: *Proceedings of the 16th International Conference on Autonomous Agents and MultiAgent Systems*. AAMAS '17. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2017, pp. 464–473. ISBN: 9781450349460. URL: https://dl.acm.org/doi/10.5555/3091125.3091194.

[12] Ashley Lewis and Michael White. "Mitigating Harms of LLMs via Knowledge Distillation for a Virtual Museum Tour Guide". In: *Proceedings of the 1st Workshop on Taming Large Language Models: Controllability in the era of Interactive Assistants!* Ed. by Devamanyu Hazarika, Xiangru Robert Tang, and Di Jin. Prague, Czech Republic: Association for Computational Linguistics, Sept. 2023, pp. 31–45. URL: https://aclanthology.org/2023.tllm-1.4.

[13] Kenneth Li et al. "Inference-Time Intervention: Eliciting Truthful Answers from a Language Model". In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023. URL: https://openreview.net/forum?id=aLLuYpn83y.

[14] Haohe Liu et al. "AudioLDM 2: Learning Holistic Audio Generation With Self-Supervised Pretraining". In: *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* 32 (May 2024), 2871–2883. ISSN: 2329-9290. DOI: 10.1109/TASLP.2024.3399607. URL: https://doi.org/10.1109/TASLP.2024.3399607.

[15] Mantas Mazeika et al. "HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal". In: *International Conference on Learning Representations (ICLR)*. 2024. URL: https://openreview.net/forum?id=f3TUipYU3U.

[16] Long Ouyang et al. "Training language models to follow instructions with human feedback". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 27730–27744. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

[17] Alex Papadopoulos Korfiatis et al. "PriMock57: A Dataset Of Primary Care Mock Consultations". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 588–598. DOI: 10.18653/v1/2022.acl-short.65. URL: https://aclanthology.org/2022.acl-short.65.

[18] Kiho Park, Yo Joong Choe, and Victor Veitch. "The Linear Representation Hypothesis and the Geometry of Large Language Models". In: *Forty-first International Conference on Machine Learning*. 2024. URL: https://openreview.net/forum?id=UGpGkLzwpP.

[19] Dinesh Raghu et al. "End-to-End Learning of Flowchart Grounded Task-Oriented Dialogs". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by Marie-Francine Moens et al. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 4348–4366. DOI: 10.18653/v1/2021.emnlp-main.357. URL: https://aclanthology.org/2021.emnlp-main.357.

[20] Hugo Touvron, Priya Goyal, Piotr Bojanowski, et al. "The Llama 3 Herd of Models". In: *arXiv preprint arXiv:2407.21783* (2024). URL: https://arxiv.org/abs/2407.21783.

[21] Alexander Matt Turner et al. "Activation Addition: Steering Language Models Without Optimization". In: *CoRR* abs/2308.10248 (2023). URL: https://doi.org/10.48550/arXiv.2308.10248.

[22] Yotam Wolf et al. "Fundamental Limitations of Alignment in Large Language Models". In: *Proceedings of the 41st International Conference on Machine Learning*. Ed. by Ruslan Salakhutdinov et al. Vol. 235. Proceedings of Machine Learning Research. PMLR, July 2024, pp. 53079–53112. URL: https://proceedings.mlr.press/v235/wolf24a.html.

[23] Ruijian Xu et al. "Learning an Effective Context-Response Matching Model with Self-Supervised Tasks for Retrieval-based Dialogues". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.16 (May 2021), pp. 14158–14166. DOI: 10.1609/aaai.v35i16.17666. URL: https://ojs.aaai.org/index.php/AAAI/article/view/17666.

[24] Andy Zou et al. *Improving Alignment and Robustness with Circuit Breakers*. 2024. arXiv: 2406.04313 [cs.LG]. URL: https://arxiv.org/abs/2406.04313.