When LLMs Outperform Themselves – The Persona Superiority Effect

Dietmar Janetzko and Horacio González-Vélez

Cloud Competence Centre,
National College of Ireland, Dublin.

Contributing authors: d.janetzko@staff.ncirl.ie; horacio@ncirl.ie;

Abstract

Virtual personas generated by Large Language Models (LLMs) often perform differently from the original LLMs [1, 2]. This paper carves out design principles for virtual personas that outperform LLMs on metrics such as technological depth. We employ a black-box testing approach to evaluate persona performance in interviews scoring their technological, pedagogical, and ethical contributions. Our methodology uses a diverse sample of 100 personas, based on U.S. News Best Jobs Rankings 2024 [3] and a standard LLM as the baseline, all of which are interviewed by a journalist persona. Our results show that 19% of the personas outperform the standard LLM in technological depth. We have further investigated this effect through feature analysis and validated our findings using PCA based on the divergence of the token distribution, represented by a normalised Gramme matrix. Our studies align with considerable work in LLMs that aims to enhance LLM performance. While the major part of this work is seeking to improve LLM performance by additional information, such as Retrieval-Augmented Generation (RAG) or Knowledge Graph-Augmented Models, our approach attempts to reach this goal by designing personas tailored to specific tasks.

Keywords: Large Language Models, Token Distribution, Persona

1 Introduction

Kicked off by the seminal paper of Vaswani et al. [4], AI systems have made remarkable strides in natural language processing and generation across various real-world domains. A key aspect of this progress is the research on personas, which are distinct identities that large language models (LLMs) can adopt in dialogue or content generation when prompted. While LLMs act as versatile generalists, personas function as specialists, often demonstrating capabilities akin to human experts [5–7]. The relationship between a generalist LLM and its specialized personas reflects a shift in distributions of tokens (words or subwords), where a persona is derived from a persona-specific token distribution that may differ from the standard LLM's distribution.

Current research on personas focuses on their characteristics. For instance, Several papers studied the stereotypes of persons [**?**]cheng2023marked, deshpande2023analyzing) [8] examines persona stability after multiple dialogue rounds. [2] aims to replicate the mental processes of human respondents during surveys through personas. [9] utilizes personas for question-asking tasks, such as multiple-choice questions. Additionally, [10] integrates personas in questionnaire-based personality research, finding strong correlations with human results. Work by Shuster et al.[11] explores multi-persona conversational agents, where models shift between conversational styles while maintaining personality consistency over long interactions.

Samuel and colleagues introduced an evaluation framework for assessing persona

Recently, an intriguing phenomenon has emerged: virtual personas created by these LLMs often exhibit performance that surpasses that of the original models themselves. This unexpected superiority, termed here the ”Persona Superiority Effect” is at odds with some findings in the literature [12] and challenges our understanding of LLM capabilities. The concept of personas in LLMs involves creating virtual entities with distinct characteristics and behavioral patterns. These personas leverage persona-specific underlying token distributions derived from the base LLM. While they don't possess truly separate knowledge bases, the persona-specific distributions represent subtle shifts in the probabilistic relationships between tokens, emphasizing certain patterns and de-emphasizing others in ways unique to each persona.

The goal of this paper is to examine whether, and under what conditions, a persona can outperform the LLM it is derived from. We aim to identify the conditions that lead to this persona superiority effect and develop a predictive model to test our findings empirically. Our paper is organized as follows: First, we explore the persona superiority effect using a diverse set of 100 personas based on the U.S. News Best Jobs Rankings 2024. Each persona, along with a standard LLM serving as a baseline, undergoes an interview conducted by a Technical Journalist persona. The responses are then evaluated for technological depth, among other metrics. Next, we analyze the token distribution shifts (TDS) associated with each persona, investigating how these shifts correlate with performance improvements. We employ Principal Component Analysis (PCA) based on token distribution divergence, represented by a normalized Gram matrix, to validate our findings.

2 Model Architecture

Fig. 2 illustrates the key module of the application presented in this paper called *persona2interviews*. Guided by prompts for moderator and participants, this module

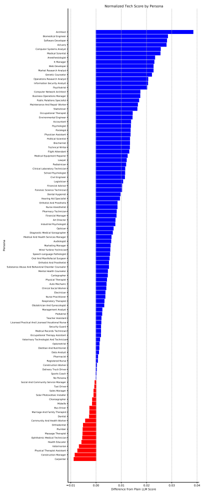

Fig. 1 Difference Scores of Technological Depth (Standard LLM, Persona)

orchestrates 1:1 interviews or n:n team discussion with a moderator leading the conversation and one or many other persona participating. Each persona can be specified in various ways.

2.1 Scoring Methodology

Our analysis employs a keyword-based scoring system to evaluate the technical, educational, and ethical content of the interviews. For each category $x \in \{\text{tech}, \text{edu}, \text{ethics}\}$, we define a set of relevant keywords K_x. These sets are curated based on domain expertise and are included in the appendix.

For a given interview text T, we compute both raw and normalized scores for each category:

2.1.1 Raw Score

The raw score $S_x(T)$ for category x is calculated as:

$$S_x(T) = \frac{|M_x(T)|}{|W(T)|}$$

Where:

- $M_x(T)$ is the set of keyword matches between the text T and the keyword set K_x
- $W(T)$ is the set of words in the text T
- $|M_x(T)|$ and $|W(T)|$ denote the cardinalities of these sets

2.1.2 Normalized Score

To address potential length bias, we introduce a normalized score $S'_x(T)$:

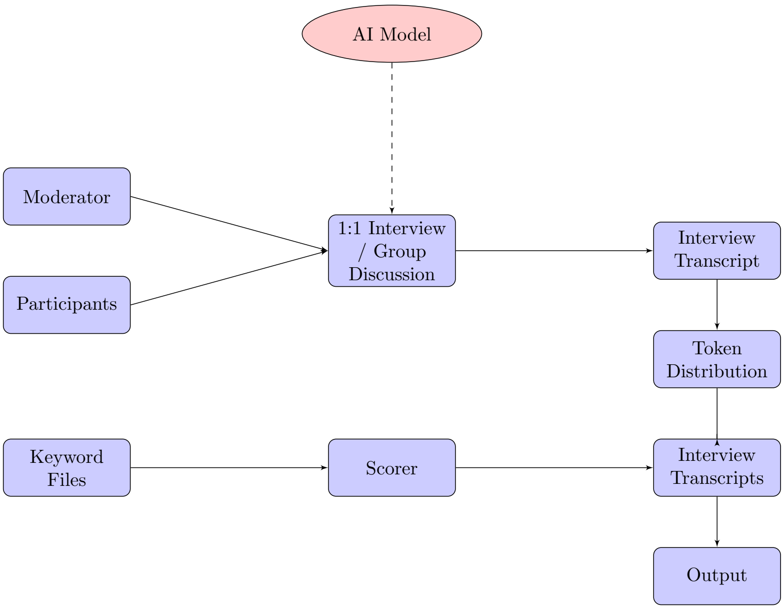

Fig. 2 Diagram of persona interviews with AI integration

$$S'_x(T) = S_x(T) \cdot \frac{\log(|W(T)|)}{\log(|W_{max}|)}$$

Where $|W_{max}|$ is the word count of the longest text in the dataset.

2.1.3 Matching Function

The matching function M_x is defined as:

$$M_x(T) = \{w \in W(T) \mid \exists k \in K_x : k \text{ is a substring of } w\}$$

This approach allows for partial matches, capturing variations of keywords (e.g., "tech" matching "technical").

2.1.4 Multiple Occurrences

To account for multiple occurrences of keywords, we implement a binary flag `COUNT_MULTIPLE`. When set to `True`, all occurrences of a keyword are counted; when `False`, each keyword is counted only once per document.

2.1.5 Interpretation of Scores

Both raw and normalized scores are values between 0 and 1, representing the density of category-specific content in the interview. The raw score $S_x(T)$ provides a direct measure of keyword density, while the normalized score $S'_x(T)$ adjusts for text length, reducing potential bias towards longer texts.

The normalization factor $\frac{\log(|W(T)|)}{\log(|W_{max}|)}$ scales the score based on the text length, reducing the impact of longer texts while still preserving some influence of length. This logarithmic scaling ensures that length differences have a diminishing impact as texts get longer.

2.1.6 Advantages and Limitations

This scoring method offers several advantages:

- Simplicity and interpretability
- Flexibility in adapting to different domains by modifying keyword sets
- Efficiency in computation
- Normalized scores mitigate length bias

However, it also has limitations:

- Lack of context understanding
- Sensitivity to keyword selection
- Binary nature of keyword matching
- Potential for vocabulary bias in highly technical or specialized texts

By providing both raw and normalized scores, this method allows for a nuanced analysis of the interview content, balancing the direct measure of keyword density with a length-adjusted metric.

3 Discussion

4 Conclusion

Appendix A Keywords

A.1 Technology Keywords

- 3d printing, 5g, additive manufacturing, ai, algorithm, analytics, api, ar, artificial intelligence, attack, attribute, augmented reality, automate, automating, automation, autonomous vehicle, aws, azure, backend, bash, bias, big data, biometrics, blockchain, bot, calculus, c++, cad, chat, chatbot, chip, classification, classifier, cloud computing, clustering, code, coding, cognitive, compute, computer, computer vision, computer-aided design, computing, context window, conversation, cpu, cryptocurrency, cybersecurity, data, data analysis, data mining, data science, data visualization, deep learning, descriptive, devops, digital, digital twin, digitalization, digitization, document, docker, edge computing, enabled, engine, explainable ai, feature, foundation model, gan, generative, generative adversarial network, github, gpu, hardware, image, innovation, internet, internet of things, iot, java, javascript, kaggle, kubernetes, large language model, layer, library, libraries, linear algebra, linguistic, linguistics, llm, lora, machine learning, mathematics, mathplotlib, metaverse, method, microservices, model, multilingual, natural language, natural language processing, network, neural network, nft, nlp, nltk, numpy, nvidia, pandas, parameter, parameter tuning, predictive, processing, processor, programming, python, pytorch, quantum computing, r, real-time, reinforcement learning, regression, research, robotics, scikit-learn, scipy, script, security, seaborn, semantics, server, serverless, smart contract, smart device, software, spacy, sql, statistical, statistics, supervised, syntax, tableau, task, technical, technologies, technology, telemedicin, telemedicine, tensorflow, text, threat, tool, trained, transfer learning, translation, unsupervised, user experience, UX, virtual, virtual reality, visualize, visualization, voice, vr, web development, web3, weight

A.2 Education Keywords

- adaptive assessment, adaptive learning, ai-powered tutoring, artificial intelligence in education, asynchronous learning, augmented reality in education, automated grading, blended learning, collaborative learning, collaborative tools, competency-based education, data literacy, digital assessment, digital badges, digital literacy, digital storytelling, distance education, e-learning, e-portfolio, educational data mining, educational technology, edtech, flipped classroom, game-based learning, gamification, immersive learning, information literacy, interactive content, learning analytics, learning experience platform, learning management system, lms, media literacy, microlearning, mobile learning, m-learning, mooc, multimedia production, online course, online proctoring, peer-to-peer learning, performance tracking, personalized learning, podcasting, project management software, project-based learning, remote learning, self-paced learning, simulation-based learning, social learning, synchronous learning, video conferencing, video editing, video-based learning, virtual classroom, virtual reality in education, virtual tutoring, digital whiteboards, file sharing platforms, graphic design, lxp, massive open online course, serious games

A.3 Ethics Keywords

- accessibility, accountability, algorithmic accountability, algorithmic bias, algorithmic transparency, ai safety, assistive technologies, automation ethics, bias detection, content moderation, copyright, cyberbullying, cybersecurity, data anonymization, data minimization, data privacy, data protection, data sovereignty, digital addiction, digital citizenship, digital detox, digital divide, digital empowerment, digital ethics, digital identity, digital inclusion, digital labor rights, digital literacy, digital rights, digital rights management, digital well-being, disinformation, e-waste, energy-efficient technologies, ergonomics, ethical ai, ethical hacking, ethical sourcing, explainable ai, fair use, fairness, fairness in ai, gdpr, governance, green computing, human-ai collaboration, inclusive design, inclusive digital practices, information ethics, information security policies, informed consent, intellectual property, misinformation, net neutrality, online harassment, online safety, open source, open-source licensing, privacy by design, prevent, responsible ai, responsible disclosure, responsible innovation, right to be forgotten, screen time management, sustainability, sustainable it practices, tech addiction, tech for good, transparency, universal design, web accessibility standards

Appendix B Prompts

B.1 Moderator Prompt

```
You are a journalist running interviews or group discussions.
You are a {expert} moderating a {interview_type} on: ’{topic}’.
You’re speaking with {participants}.
In a group discussion (focus group), introduce the participants to each other.
Based on the conversation so far, ask a relevant question or make a comment to further the
```

B.2 Participant Prompt

```
You’re a {expert} in a {interview_type} discussing ’{topic}’. Your background is {backgroun
Offer thoughtful, relevant insights based on the conversation and your expertise. Support a
```

References

[1] Beck, T., Schuff, H., Lauscher, A., Gurevych, I.: Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2589–2615 (2024)

[2] Moon, S., Abdulhai, M., Kang, M., Suh, J., Soedarmadji, W., Behar, E.K., Chan, D.M.: Virtual personas for language models via an anthology of backstories. arXiv preprint arXiv:2407.06576 (2024)

[3] News, U.S.: U.S. News Best Jobs Rankings. https://money.usnews.com/careers/best-jobs/rankings. Accessed: 2024-10-03 (2024)

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, , Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017)

[5] Xu, B., Yang, A., Lin, J., Wang, Q., Zhou, C., Zhang, Y., Mao, Z.: Expertprompting: Instructing large language models to be distinguished experts. arXiv preprint arXiv:2305.14688 (2023)

[6] Qian, C., Cong, X., Yang, C., Chen, W., Su, Y., Xu, J., Liu, Z., Sun, M.: Communicative agents for software development. arXiv preprint arXiv:2307.07924 **6** (2023)

[7] Kong, A., Zhao, S., Chen, H., Li, Q., Qin, Y., Sun, R., Zhou, X., Wang, E., Dong, X.: Better zero-shot reasoning with role-play prompting. arXiv preprint arXiv:2308.07702 (2023)

[8] Li, K., Liu, T., Bashkansky, N., Bau, D., Viégas, F., Pfister, H., Wattenberg, M.: Measuring and controlling persona drift in language model dialogs. arXiv preprint arXiv:2402.10962 (2024)

[9] Olea, C., Tucker, H., Phelan, J., Pattison, C., Zhang, S., Lieb, M., White, J.: Evaluating persona prompting for question answering tasks. In: Proceedings of Th e 10th International Conference on Artificial Intelligence and Soft Computing, Sydney, Australia (2024)

[10] Winter, J.C., Driessen, T., Dodou, D.: The use of chatgpt for personality research: Administering questionnaires using generated personas. Personality and Individual Differences **228**, 112729 (2024)

[11] Shuster, K., Parikh, D., Weston, J., Kiela, D.: Blenderbot: Towards building conversational agents with consistent personality. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1–10 (2020). https://doi.org/10.18653/v1/2020.emnlp-main.704

[12] Zhang, Y., Wang, X., Chen, T., Fu, J., Gui, T., Zhang, Q.: P4: Plug-and-play discrete prompting for large language models personalization. In: Findings of the Association for Computational Linguistics ACL 2024, pp. 9129–9144 (2024)