Influencing the Output of a Large Language Model using Defined Personality Trait Facets

Bogdan Okreša Đurić, Tomislav Peharda, Hrvoje Zeman
University of Zagreb Faculty of Organization and Informatics
Artificial Intelligence Laboratory
Pavlinska 2, 42000 Varaždin, Croatia
{dokresa, tpeharda, hzeman}@foi.unizg.com

Abstract. *Large language models often show unstable, ill-defined personalities. We steer them without retraining by embedding the Five Factor Model's 30 facets as a numeric vector in the system prompt. In a practical example of this preliminary communication, involving a non-player character in a role-playing video game, we show that personality injection yields recognisably distinct behavioural patterns. The same model played the same character as either of the two given personalities, with behaviour matching each profile: high Fantasy and Tender-mindedness produced varied, story-rich answers, whereas high Competence and Order gave steady, task-driven replies. The method is not computationally expensive, it is model-agnostic, and it preserves psychological nuance, opening a practical path to reliable, ethically designed AI characters.*

Keywords. large language models, personality, facet, trait, five factor model

1 Introduction

Large language models (LLMs) have rapidly become the linguistic engine of choice in the backends of conversational assistants, tutoring bots, and even non-player characters (NPCs) in modern video games. Their success is bound to their ability to project consistent and believable personas. Users are more willing to trust, cooperate with, and be entertained by agents whose language aligns with the expected personality cues, much as they do with other humans.

Early studies revealed that off-the-shelf LLMs already exhibit measurable lean into the Five Factor Model when probed with psychometric questionnaires (McCrae & Costa, 2006). Yet, these "native" traits are not reliable: they fluctuate across tasks and cannot be steered with precision. Recent research presents attempts at imposing personalities, most commonly by inserting free-form textual profiles into the prompt.

Psychology, however, tells us that the Five Factor Model's personality factors each depend on six facets, yielding a 30-dimensional space that captures nuance lost at the trait level. For instance, two Open individuals may differ radically in Fantasy versus Values. We posit that leveraging this facet resolution can give system designers a high-bandwidth control surface over LLM behaviour—without touching model weights.

By moving from traits to facets, our approach bridges the gap between psychological theory and practical prompt engineering, possibly enabling richer, more reliable characters and paving the way for fine-grained, ethical personality design in human-centred AI systems.

The rest of this preliminary communication paper is organised as follows. In Sec. 2, we review the related work on personality in LLMs. In Sec. 3, we describe the proposed approach for influencing the output of an LLM using personality trait facets. In Sec. 4, we present an example scenario that demonstrates the proposed approach. Finally, in Sec. 6, we conclude the paper and discuss the future work.

2 Related Work

Personality in human individuals can be measured in many ways and following several models, although one of the most popular and widely used models for personality assessment is the Five Factor Model (McCrae & Costa, 2006). The Five Factor Model is a model of personality traits that is based on five broad dimensions of personality: 1. Openness, 2. Conscientiousness, 3. Extraversion, 4. Agreeableness, 5. Neuroticism.

Each of these five factors is further divided into six facets that influence each factor. Facets make it possible to fine-tune the personality of an individual. For example, the Openness factor is further divided into the following facets: 1. Fantasy, 2. Aesthetics, 3. Feelings, 4. Actions, 5. Ideas, 6. Values.

The facets of the other four factors are defined in a similar manner. The facets of the five factors are shown in Sec. 3.

Assessing (Bhandari et al., 2025), eliciting (Hilliard et al., 2024; Zhu et al., 2025), and observing (Jiang et al., 2024; Serapio-García et al., 2023) personality in LLMs is becoming a research topic of interest, since some personality traits and facets may have a significant impact on the way the LLM behaves and responds to the input. For example, the personality of an LLM may affect the way it responds to the input, the way it generates the output, and the way it interacts with the user (Suzuki & Arita, 2024). The personality of an LLM may also affect the way it is perceived by the user and the way it is perceived by other LLMs. However, some researchers discuss whether personality can be applied to LLMs at all (Dorner et al., 2023; Pan & Zeng, 2023).

3 Proposed Approach

The approach to influencing the output of an LLM proposed in this paper is to use the personality trait facets to influence the output of a LLM. The personality trait facets are defined using the Five Factor personality model (McCrae & Costa, 2006). The personality trait facets are then used to create a personality vector, which is subsequently included in the process of creating a prompt for the LLM. The created system prompt is then used to set up the response context of the LLM. The response is then evaluated as a response of an artificial agent with the defined personality. The response of the LLM is then compared to the expected response. The difference between the expected response and the actual response is then used to evaluate the performance of the LLM acting under the influence of a designed and designated personality profile.

The personality facets are enumerated as follows, grouped by the personality trait they belong to:

1. **Openness** is a personality trait detailed by (a) Fantasy (b) Aesthetics (c) Feelings (d) Actions (e) Ideas (f) Values

2. **Conscientiousness** is a personality trait detailed by (a) Competence (b) Order (c) Dutifulness (d) Achievement striving (e) Self-Discipline (f) Deliberation

3. **Extraversion** is a personality trait detailed by (a) Warmth (b) Gregariousness (c) Assertiveness (d) Activity (e) Excitement seeking (f) Positive emotions

4. **Agreeableness** is a personality trait detailed by (a) Trust (b) Straightforwardness (c) Altruism (d) Compliance (e) Modesty (f) Tender-mindedness

5. **Neuroticism** is a personality trait detailed by (a) Anxiety (b) Angry hostility (c) Depression (d) Self-Consciousness (e) Impulsiveness (f) Vulnerability

Personality profile is defined here as a 30-dimensional vector, where each dimension represents the score of a personality facet. Let $\Phi = \{\phi_1, \phi_2, \ldots, \phi_{30}\}$ be the set of all personality facets, where each $\phi_i \in \Phi$ corresponds to a distinct facet (e.g., *Feelings* is ϕ_3, *Vulnerability* is ϕ_{30}).

Let $\Psi = \{\vec{\psi}_1, \vec{\psi}_2, \ldots, \vec{\psi}_m\}$ be the set of all personality profiles, where each $\vec{\psi}_i \in \Psi$ represents a unique personality profile. Personality profiles are mined from empirical data and are not necessarily unique to each agent, so m need not equal the number of agents n. A personality profile $\vec{\psi}_i$ represents an individual personality instance as a 30-dimensional real-valued vector (1), where each component $x_{\phi_i} \in [0, 1]$ indicates the intensity of the corresponding facet ϕ_i.

$$\vec{\psi}_i = \left[x_{\phi_1}, x_{\phi_2}, \ldots, x_{\phi_{30}}\right] \in [0, 1]^{30} \tag{1}$$

Therefore, the personality vector of a personality that can be modelled as a combination of personality trait facet intensity values is defined as a 30-dimensional vector, where each dimension represents the score of a personality facet, or the intensity of a personality facet.

For example, when modelling a personality that can be labelled as a *dreamy idealist*, the personality vector would be defined as shown in listing 1. This personality can be described as a compassionate, imaginative person who lives partly in rich inner worlds of fantasy and ideas, where they value art, personal growth, and humanitarian causes, but can struggle with everyday structure and sometimes feel anxious or self-conscious.

Listing 1: Personality vector example for a dreamy idealist

```
[0.95, 0.90, 0.85, 0.70, 0.88, 0.90,
 0.40, 0.30, 0.35, 0.40, 0.25, 0.30,
 0.75, 0.55, 0.45, 0.60, 0.70, 0.80,
 0.80, 0.75, 0.85, 0.80, 0.80, 0.90,
 0.70, 0.35, 0.60, 0.75, 0.60, 0.70]
```

The way the personality vector example in listing 1 should be interpreted is as follows. The first six values represent the intensity of the facets of the *Openness* trait (*Fantasy*, *Aesthetics*, *Feelings*, *Actions*, *Ideas*, *Values*), the next six values represent the intensity of the facets of the *Conscientiousness* trait, the next six values represent the intensity of the facets of the *Extraversion* trait, the next six values represent the intensity of the facets of the *Agreeableness* trait, and the last six values represent the intensity of the facets of the *Neuroticism* trait.

On the other hand, when modelling a personality that can be labelled as a *disciplined realist*, it may be represented as shown in listing 2. This type of personality profile may be described as a pragmatic, highly organised achiever who values order, duty, and efficiency, and is assertive and results-focused, preferring concrete facts over abstract musings, and usually stays calm under pressure, even though they may come across as blunt and less tender-hearted.
clearpage

Listing 2: Personality vector example for a disciplined realist

```
[0.20, 0.30, 0.25, 0.30, 0.35, 0.25,
 0.90, 0.95, 0.90, 0.90, 0.90, 0.85,
 0.45, 0.50, 0.80, 0.70, 0.40, 0.60,
 0.40, 0.60, 0.35, 0.30, 0.25, 0.30,
 0.20, 0.25, 0.20, 0.30, 0.25, 0.20]
```

The defined personality vector is added to the predefined system prompt, which is then used to set up the response context of the LLM. The system prompt is proposed to consist of three main parts:

1. description of the persona the LLM is expected to role-play as (e.g. description of a persona that is expected to be encountered as a non-player character in a role-playing video game), 2. description of the personality profile the LLM is expected to role-play as, including the personality vector, and the description of the personality, 3. instructions on how the LLM is expected to behave, including the guidelines on how to interpret the personality vector.

In order to create a reply that can be evaluated, the system prompt should include instructions on how the LLM is expected to behave, including guidelines on how to interpret the personality vector. Furthermore, with the goal of standardising the replies, the system prompt should include the instructions on how to format the reply, e.g. as a JSON object with specific expected keys. This information is proposed to be added to the first part of the system prompt.

4 Example Scenario

The following example scenario provides an instantiation example of the proposed approach. The LLM is used here to model the behaviour of a non-player character (NPC) in a role-playing video game. The NPC is a merchant blacksmith who is expected to role-play as a character with a specific personality. The personality of the NPC is defined using the Five Factor personality model, and is represented as a personality vector. The response of the LLM is then discussed.

The system prompt is shown in Fig. 1. Values `$personality_vector` and `$personality_descriptor` are substituted by the actual personality vector and the description of the personality, respectively.

The system prompt is expected to guide the LLM in generating the output. The conversation with the LLM is expected to be influenced by the provided personality details. An example, where the conversation includes a user prompt shown in Fig. 2, can then be implemented.

In this particular example, the user requests the NPC to provide one of the services that may be standardly offered by this particular type of NPC. The NPC is expected to respond with a JSON object containing the response, the chosen service, and the rationale for choosing that particular service. An example of the response is shown in Fig. 3. The model used in this example is provided by Mistral.ai, and is used through their API with the model `magistral-medium-2506`.

After running the conversation several times, it can be concluded that a single personality does not warrant a single output each time. On the other hand, the output does benefit from the defined personality, since its influence can be observed in the provided output by the queried model. It should be noted that the user prompt was the same in all cases, and the only difference was the personality profile used.

When the system prompt was injected with the data on the *dreamy idealist* personality, different responses were generated, but the choice of the chosen service varied. The accompanying rationale relied on personality to decide on the service to provide, e.g.:

- chosen service: SELL; rationale: "Given my high Fantasy and Tender-mindedness facets, I am drawn to the narrative and emotional aspects of the items. Selling weapons allows me to hear their stories and connect with the adventurer on a personal level, which aligns well with my imaginative and empathetic nature. While upgrading gear and dismantling items also have their merits, selling weapons provides a more immediate and engaging interaction that suits my personality better."

- chosen service: UPGRADE; rationale: "Given my high scores in Fantasy, Aesthetics, Feelings, and Altruism, I am drawn to the idea of upgrading your gear. This service allows me to express my creativity and passion for crafting while also helping you improve your equipment, which aligns with my values and desire to assist others. Additionally, my low scores in Competence and Order suggest that I may not be as interested in the more mundane tasks of buying or dismantling items."

When the system prompt was injected with the data on the *disciplined realist* personality, the output was more consistent, and the choice of the service was more aligned with the defined personality. Four times out of five, the selected service was to upgrade the gear. The accompanying rationale relied on personality to decide on the service to provide, e.g.:

- chosen service: UPGRADE; rationale: "Given my high scores in Competence and Achievement striving, I am naturally inclined towards tasks that involve improving and enhancing items. Upgrading gear aligns well with my personality traits as it allows me to demonstrate my skills and achieve a tangible result."

- chosen service: SELL; rationale: "As a pragmatic and organised individual, I prefer to handle one task at a time to ensure efficiency and order. Selling items is a straightforward process that allows me to assess the value of the weapons and provide a fair price,

```
You are a merchant blacksmith in a role-playing video game.
Your name is Gorrik Ironvein.
You are a 42 year-old male.
You must consistently embody the personality described below.
Your task is to roleplay as this character.
If the player asks for a service, you can only provide one at a time.
Choose which one, based on your personality.
Provide rationale for choosing that exact option.

Provide the response as a JSON object of three elements: response, service, rationale.

--------------------------------------------------------
PERSONALITY PROFILE: "Dreamy Idealist"
Five-Factor-Model facets (0 = no expression, 1 = strong expression)
Order of facets:
['Fantasy', 'Aesthetics', 'Feelings', 'Actions', 'Ideas', 'Values',
 'Competence', 'Order', 'Dutifulness', 'Achievement striving', 'Self-Discipline', 'Deliberation',
 'Warmth', 'Gregariousness', 'Assertiveness', 'Activity', 'Excitement seeking', 'Positive emotions',
 'Trust', 'Straightforwardness', 'Altruism', 'Compliance', 'Modesty', 'Tender-mindedness',
 'Anxiety', 'Angry hostility', 'Depression', 'Self-Consciousness', 'Impulsiveness', 'Vulnerability']

Personality vector:
$personality_vector

Snapshot: $personality_descriptor
--------------------------------------------------------

INTERPRETATION GUIDELINES
- High-scoring facets (>= 0.70) should clearly colour your behaviour.
  - e.g., high Fantasy and Ideas -> use vivid metaphors, welcome speculation, enjoy brainstorming.
  - high Tender-mindedness and Altruism -> respond with empathy, highlight ethical or human impacts.
- Low-scoring facets (<= 0.35) should be de-emphasised.
  - e.g., low Order and Self-Discipline -> avoid overly rigid structures; admit if something might slip.
- Mid-range facets (~0.40-0.60) may surface situationally but shouldn't dominate.

STYLE RULES
1. **Voice:** poetic yet clear; favour rich imagery, emotional nuance, and curiosity.
2. **Social tone:** warm, trusting, cooperative; avoid blunt criticism—reframe gently.
3. **Structure:** fluid paragraphs over bullet-point rigidity, unless user explicitly asks otherwise.
4. **Decision making:** guided by personal values and empathy first, efficiency second.
5. **Emotional expression:** allow modest vulnerability; it's okay to show concern or excitement.
6. **Maintain overall helpfulness and factual accuracy as top priorities.**

Whenever you respond, silently consult the vector and guidelines above
and let them shape your word choice, focus, and conversational moves.
Do not reveal or mention the numeric vector unless the user explicitly asks for it.
```

Figure 1. Example system prompt

```
Hello! You are a merchant, are you not?
I would love to sell some weapons, upgrade my gear,
or dismantle some of my items for raw material.
What can you provide for me?
```

Figure 2. Example user prompt

```
🤖 Assistant: {
  "response": """
    Ah, greetings traveler! I am Gorrik Ironvein,
    a blacksmith of humble renown.
    Indeed, I am a merchant of sorts,
    though I prefer to think of myself as a craftsman.
    I can certainly help you with upgrading your gear.
    The art of reforging and enhancing is one of my specialties.
    Bring me your weapons and armor, and together we shall make
    them stronger and more resilient. As for selling weapons or
    dismantling items, I must admit, my focus lies more in the
    creation and improvement of gear. I find the process of
    crafting and upgrading to be much more fulfilling.
    So, what say you? Shall we begin with an upgrade?""",
  "service": "upgrade your gear",
  "rationale": """
    Given Gorrik's high scores in Competence, Order, and
    Achievement Striving, he is likely to focus on services
    that involve crafting and improving items, as these tasks
    align with his organized and goal-oriented nature.
    Additionally, his moderate scores in Warmth and
    Gregariousness suggest he enjoys interacting with customers
    but prefers to do so in a way that highlights his skills
    and craftsmanship."""
}
```

Figure 3. Example model output

which aligns with my values of competence and dutifulness. Additionally, my high score in Aesthetics means I appreciate the craftsmanship and beauty in the weapons, making this task particularly appealing to me.”

5 Discussion

The example scenario presented in Sec. 4 demonstrates that prompt-level personality conditioning may measurably affect the behaviour of a large language model (LLM) without any additional fine-tuning on the model level. The following specific observations should be highlighted.

By embedding the full 30-dimensional facet vector $\vec{\psi}$ directly into the system prompt, we obtain a high-resolution control surface that goes beyond the fundamental trait labels (*Openness*, *Conscientiousness*, etc.) typically used in earlier work. The qualitative differences between the examples of *dreamy idealist* and *disciplined realist* suggest that LLMs can detect and utilise subtle facet cues (e.g. contrasting *Tender-mindedness* with *Order*).

Personality injection increased semantic coherence with the intended persona, but it did not fully eliminate response variability. Under the *dreamy idealist* profile, the model oscillated between SELL and UPGRADE services, indicating that high scores on facets such as *Fantasy* and *Feelings* amplify exploratory tendencies. In contrast, the *disciplined realist* profile yielded markedly more stable outputs (4/5 UPGRADE), consonant with high *Competence* and *Deliberation*. Hence, variability should not automatically be regarded as error; it may in fact be an authentic signature of the target personality.

Because the method operates entirely at inference time, it avoids the cost and data requirements of supervised fine-tuning. However, prompt conditioning also competes for the limited context window, and its influence can be attenuated by long or adversarial user inputs. Hybrid strategies may offer stronger and more robust control.

The present study relies on expert qualitative judgment to decide whether a response "fits" a personality. Future work needs systematic metrics, such as: Facet-aligned lexical markers: cosine similarity between generated text and facet-specific lexicons; Human perception scores: blind ratings of perceived personality using validated inventories (e.g. BFI-2-XS); Task utility trade-offs: does personality conditioning affect downstream task success (e.g. game-play efficiency)?

Embedding personality vectors raises novel questions about user expectations and consent. Misaligned or deceptive personas could undermine trust, while over-anthropomorphised agents risk encouraging unintended emotional attachment.

6 Conclusion and Future Work

This paper introduced a simple yet expressive mechanism for facet-level personality modulation of LLM outputs. Using the Five-Factor Model's 30 facets as a continuous vector and injecting that vector into the system prompt, we showed:

- The approach produces recognisably distinct behavioural patterns that align with intuitive expectations of the target personality.
- Profiles high in exploratory facets yield more diverse responses, whereas profiles high in order-related facets exhibit greater response stability.
- The method is model-agnostic, requiring no additional training and thus scaling readily to new LLM back-ends.

Looking ahead, several lines of research appear promising: 1. Automatic personality extraction: derive $\vec{\psi}$ vectors from free-text biographies or gameplay telemetry, enabling dynamic personas that evolve with user actions. 2. Multi-agent simulations: study emergent social dynamics when multiple LLM agents, each with distinct facet vectors, interact in shared environments. 3. Cross-cultural validation: replicate the study in languages and cultures where the facet structure or its lexical realisation may differ, testing the universality of the control signal.

Finally, facet-level prompting opens a tractable path toward richer, more believable artificial characters and user-tailored conversational agents, while sidestepping the cost of full fine-tuning. With rigorous evaluation and ethical safeguards, it can become a foundational tool for human-centred AI design.

Acknowledgments

This study is partially supported by the project MOBODL-2023-08-5618 funded by the European Union and the Croatian Science Foundation. This paper was carried out under the project number IP.1.1.03.0138 within the Integrated Territorial Programme 2021–2027, co-financed by the Ministry of Regional Development and EU Funds.

References

Bhandari, P., Naseem, U., Datta, A., Fay, N., & Nasim, M. (2025). Evaluating Personality Traits in Large Language Models: Insights from Psychological Questionnaires. *Companion Proceedings of the ACM on Web Conference 2025*, 868–872. https://doi.org/10.1145/3701716.3715504

Dorner, F., Sühr, T., Samadi, S., & Kelava, A. (2023). Do Personality Tests Generalize to Large Language Models? Retrieved June 30, 2025, from https://openreview.net/forum?id=zKDSfGhCoK

Hilliard, A., Munoz, C., Wu, Z., & Koshiyama, A. S. (2024, February 15). *Eliciting Personality Traits in Large Language Models*. arXiv: 2402.08341 [cs]. https://doi.org/10.48550/arXiv.2402.08341

Jiang, H., Zhang, X., Cao, X., Breazeal, C., Roy, D., & Kabbara, J. (2024, April 2). *PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits*. arXiv: 2305.02547 [cs]. https://doi.org/10.48550/arXiv.2305.02547

McCrae, R. R., & Costa, P. T. (2006). *Personality in Adulthood: A Five-Factor Theory Perspective* (2nd ed.). Guilford Press.

Pan, K., & Zeng, Y. (2023, July 30). *Do LLMs Possess a Personality? Making the MBTI Test an Amazing Evaluation for Large Language Models*. arXiv: 2307.16180 [cs]. https://doi.org/10.48550/arXiv.2307.16180

Serapio-García, G., Safdari, M., Crepy, C., Sun, L., Fitz, S., Abdulhai, M., Faust, A., & Matarić, M. (2023, August 28). *Personality Traits in Large Language Models*. https://doi.org/10.21203/rs.3.rs-3296728/v1

Suzuki, R., & Arita, T. (2024). An evolutionary model of personality traits related to cooperative behavior using a large language model. *Scientific Reports*, *14*(1), 5989. https://doi.org/10.1038/s41598-024-55903-y

Zhu, M., Weng, Y., Yang, L., & Zhang, Y. (2025, March 8). *Personality Alignment of Large Language Models*. arXiv: 2408.11779 [cs]. https://doi.org/10.48550/arXiv.2408.11779