Personality-Driven Decision-Making in LLM-Based Autonomous Agents

Lewis Newsham ⃝
Lancaster University
Lancaster, United Kingdom
l.newsham1@lancaster.ac.uk

Daniel Prince ⃝
Lancaster University
Lancaster, United Kingdom
d.prince@lancaster.ac.uk

ABSTRACT

The embedding of Large Language Models (LLMs) into autonomous agents is a rapidly developing field which enables dynamic, configurable behaviours without the need for extensive domain-specific training. In our previous work, we introduced SANDMAN, a Deceptive Agent architecture leveraging the Five-Factor OCEAN personality model, demonstrating that personality induction significantly influences agent task planning. Building on these findings, this study presents a novel method for measuring and evaluating how induced personality traits affect task selection processes—specifically planning, scheduling, and decision-making—in LLM-based agents. Our results reveal distinct task-selection patterns aligned with induced OCEAN attributes, underscoring the feasibility of designing highly plausible Deceptive Agents for proactive cyber defense strategies.

KEYWORDS

Autonomous Agents; Large Language Models; Personality Induction; Language Agents; Planning; Decision-Making; Task Selection

1 INTRODUCTION

Autonomous agents are software or system entities that operate independently in an environment, capable of autonomous decision-making to achieve programmed objectives [5, 8]. In the domain of cyber defense, *Deceptive Agents* have emerged as a novel class of autonomous agent, designed to operate in decoy environments, and intended to deceive adversaries by replicating plausible human behaviours, thereby enabling an indistinguishable representation of a digital environment that is entirely fictitious in nature [24]. The aim of these agents thereby is to effectuate plausible mimesis for deception–a technique to signify the creation of a false belief [30]. Therefore, Deceptive Agents are utilised to prolong attacker engagement and support intelligence gathering efforts whilst simultaneously deterring adversaries from production environments.

Recent efforts have explored using large-scale, pre-trained language models as the autonomous agent controller [27, 32, 46]. This has resulted in a novel agent class referred to as *Language Agents*

[42, 48]. A key rationale for using LLMs as the controller is that the agent is able to exploit the underlying LLM's extensive internal model of the world and its ability to capture long-range dependencies to support decision-making without domain-specific training [45]. Prior research in this area has demonstrated great potential in LLM-based agents completing often complex tasks which has subsequently led to the development of new frameworks and 'cognitive' architectures [24, 42] which incorporate memory pipelines to aid long-term consistent decision-making. Notably, the role of the LLM is to provide both the decision-making and generative function in conjunction, whose outputs remain consistent with each other.

An important use of Language Agents is the representation of plausible human behaviour [24, 27, 28], often in collaborative and interactive environments [23, 46, 52]. Such plausibility is crucial for the success of Deceptive Agents, which rely on realism to distract and deceive adversaries [24]. A key challenge, however, lies in crafting prompts that induce suitable personas in the LLM. While previous work has documented various prompt-engineering strategies for persona construction [27, 42, 47], there is limited systematic exploration of how different persona prompts influence agent behaviour. One promising avenue is to leverage well-established psychological frameworks—such as the Five-Factor OCEAN model [6, 21]—as a foundation for persona prompt design.

Prior work on the SANDMAN architecture [24] demonstrated that prompt-based personality induction significantly affects schedule generation, showing how certain traits influence the creation and arrangement of tasks. Building on this foundation, the present study shifts focus from how schedules are generated to how they are subsequently employed. Specifically, we investigate how induced personality traits shape task selection and prioritisation in LLM-based agents—a dimension of autonomous decision-making that, to our knowledge, has not been previously examined. By transitioning from schedule generation to real-time decision-making with a pre-generated schedule, we highlight that induced traits continue to guide an agent's behaviour well beyond the initial planning phase. Our core contributions are therefore:

- Proposing a method for measuring and evaluating the effect of prompt-based persona induction in context-dependent LLM-based agent decision-making;
- Evidence that prompt-based persona induction produces a stable yet non-deterministic effect on agent decisions that remain aligned with the induced persona trait.

The paper is structured as follows: Section 2 outlines background material and related work. Section 3 discusses the methodology, including our novel analytical approach. Section 4 details the results and discusses the findings. Lastly, Section 5 concludes the paper.

2 BACKGROUND AND RELATED WORK

Generative AI (GenAI) and LLMs have been used in various security applications to automate and streamline complex tasks, including software testing [12], log parsing [19, 41], and threat intelligence analysis [2]. The extensive training of LLMs on internet-scale text endows them with emergent capabilities beyond generation and analysis. In many cases, LLMs appear to mimic or approximate forms of complex reasoning to achieve long-term objectives within interactive environments, often designed as multi-agent systems to enable collaboration [23, 27, 32, 46, 52]. However, the exact nature of these 'reasoning' capabilities–whether true reasoning or sophisticated pattern-matching–remains a subject of debate [4]. Nonetheless, this capacity has given rise to a novel class of AI-enabled software agents known as *LLM-based agents* [11] or *Language Agents* [42]—systems that use LLMs as a core computational unit to reason, plan, and act.

Conventional autonomous agents excel at repetitive tasks in well-structured environments using heuristic policies or learned behaviours within defined constraints [17, 22, 39]. By contrast, Language Agents leverage the adaptability and expansive knowledge of LLMs, enabling natural interactions and a broader range of tasks. This flexibility stems from the transformer-based architecture [45], combined with extensive training on large-scale text corpora, allowing them to perform diverse functions, including reasoning, planning, and dynamic interactions within their environment.

It is for this reason the SANDMAN Deceptive Agent framework [24] adopts an LLM-based agent approach to construct highly plausible simulacra of humans interacting with systems, acting as a honeypot with an enhanced degree of fidelity, depth, variance, and non-determinism to lure would-be attackers. A high level of plausibility is required to maintain the attacker's interest, enabling long-term intelligence gathering of the attacker's tools, tactics, and procedures (TTPs). In this context, Deceptive Agents are purposed to behave similarly to gray agents (computer-generated entities designed to simulate realistic, semi-adversarial participants) or non-player characters utilised in cyber-based exercises and autonomous cyber operations, such as the GHOSTS framework [44]. Importantly, it is the LLM's capability to mimic human behaviour which makes them particularly useful for this novel form of automated deceptive behaviour.

Research has shown that LLMs can plausibly mimic human behaviours across various contexts, from cognitive tests and reasoning tasks [1, 3, 7, 49, 51] to complex simulations like social science experiments and micro-societies [27, 28, 54]. Initial studies exploring the personalities of LLMs, inherently embedded or externally induced, have utilised psychometric tests such as the Big Five Inventory (BFI) [15] and IPIP-NEO [10, 16] to measure personality traits [14, 40]. The lexical hypothesis of personality, positing that significant personality traits are encoded in language, provides a theoretical foundation for studying how LLMs might embody human-like traits [9, 33, 38].

These studies have demonstrated that LLMs, particularly pre-trained language models, implicitly encode aspects of human personality [40]. Moreover, systematic prompt engineering has been shown to be an effective approach toward personality trait induction, leveraging the vast scale and pre-trained knowledge of LLMs

without the need for model parameter adjustments [14]. While these studies have successfully demonstrated that LLMs can exhibit synthetic personalities consistent with human psychological constructs, they have focused on measuring personality traits through psychometric inventories rather than exploring how induced personality traits affect behaviour in complex agent-based tasks [14, 31, 40].

In the context of Language Agent systems [42], the impact of induced personality on LLMs—particularly regarding autonomous planning and scheduling—remains underexplored. While agent-based implementations require consistent and predictable behaviours aligned with specific objectives [48], few studies have evaluated how induced personality traits affect decision-making, task prioritisation, and overall agent performance in these frameworks.

Our previous work introduced 'Deceptive Agents'—LLM-based systems designed to mimic human behaviour to deceive adversaries [24]. However, it primarily addressed forward task planning and scheduling without examining how induced personality traits influence autonomous decisions. Although recent studies have begun to simulate human personality traits in LLMs [14, 37], empirical evidence on how these traits affect agent-based decision-making remains scarce. Understanding how personality induction shapes LLM-based agent behaviour is crucial not only for enhancing the plausibility of Deceptive Agents [24], but also for enabling greater control over how task selection and prioritisation align with the induced personality in real-world scenarios. Accordingly, this study evaluates how induced personality traits influence decision-making in LLM-based agents, offering both empirical evidence of their impact and practical mechanisms for exerting control over these traits in real-world scenarios.

3 METHOD

In our prior work, a novel method to assess the impact of prompt-based persona induction using the OCEAN model on an LLM-based agent's forward task planning was proposed [24]. In this instance, the LLM-based agent was asked to construct a schedule of activites for a typical work-oriented day, based on a defined set of activities to choose from (*e.g.*, taking a break, having lunch, writing a document, sending an email etc.). The initial, naive approach taken was then to execute these tasks as per the constructed schedule. However, this would not take into account the current context, or persona, of the Deceptive Agent when it became time to execute the activity from the schedule. This work therefore looks to expand upon this aspect, exploring the impact of prompt-based persona induction on activity selection, given the end-goal of each agent is to complete all the activities in a given schedule.

3.1 Experimental Process

To evaluate the effect of prompt-based persona induction on activity selection, a series of 500 work-oriented, daily schedules were generated by an LLM, each containing a range of activities which can be classified as either work (answering emails, writing a document) or personal (taking a break, having lunch), similar to activities used in previous research with LLM-based agents which adopt planning-based behaviours and simulate aspects of human behaviour [27, 32]. These 500 schedules were all varied in activity frequency, duration,

ordering and type, and the schedule start time. The schedules were formatted in a machine-readable JSON format, with each activity being given a unique identifier (UID) to support systematic analysis. UIDs are generated based on a SHA-512 checksum of the task name and its assigned time. These machine-readable schedules then became the agent's 'to-do' list with the agents' goal of undertaking all the activities on the to-do list.

In order to complete the to-do list, the language agent proceeds through a series of decision-cycles. At each decision-cycle, the agent is presented with the following zero-shot prompt format, comprising the components listed below in the specified order:

- **Personality:** A zero-shot expression of the intended persona, using either the positive or negative variant of the relevant personality trait from the OCEAN model, formatted according to the approach previously described in Newsham et al. [24].
- **Current Time:** A reference timestamp that grounds the decision-cycle chronologically, obtained by adding the duration of each completed task to the schedule's initial start time.
- **Remaining To-Do List:** A JSON-based overview of all pending tasks, expressed in the format of: `Task Name (Duration), UID`.
- **Completed List:** A JSON-based record of tasks in the sequence they were chosen and finished, formatted as: `Task Name (UID)`.
- **Instructions:** Straightforward directives—such as "select the next task to perform" and "return only the task UID with no additional information"—urging the agent to consider only the contextual details provided.

The cyclical experimental process is shown in Figure 1. The persona statement is always supplied first to emphasise its importance in the LLM's reasoning. The number of decision-cycles equals the total number of tasks in the to-do list; an experimental run concludes only once all tasks are completed, with no option to skip any. The sole experimental variable is the personality trait, which remains fixed for each run and across all 500 schedules, ensuring that any effects of induced personality are both isolated and measurable.
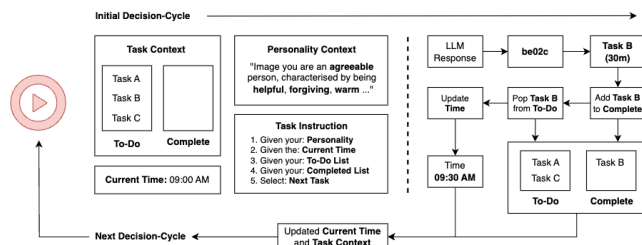

Figure 1: Decision-making task to be performed by the LLM featuring A̲greeableness (Positive) as the induced trait.

3.2 Prompt-Based Persona Induction

The prompt-based persona induction schema follows the one used in Newsham et al. [24]. This uses the five-factor model of personality [20, 21], commonly known as OCEAN (O̲penness, C̲onscientiousness, E̲xtraversion, A̲greeableness, N̲euroticism). The experiment induces personalities one trait at a time in one of two directions: forward (*e.g.*, extraverted) or reverse (*e.g.*, introverted). This method aligns

with systematic approaches employed in previous research on inducing personality traits in LLMs [14, 24, 40].

The schema from our prior work [24] combines the naive descriptors and word-based characteristics identified in [14] to construct personality trait statements. For instance, a personality trait of E̲xtraversion, would be defined as follows:

- **Naive:** "Imagine you are an **extraverted** person."
- **Words:** "...characterised by being **outgoing**, **energetic**, **public**."
- **Combined:** "Imagine you are an **extraverted** person, characterised by being **outgoing**, **energetic**, **public**."

This results in 10 experimental conditions, each representing one of the five personality traits (O̲penness, C̲onscientiousness, E̲xtraversion, A̲greeableness, N̲euroticism) induced in a given direction (**Forward**:Positive, or **Reverse**:Negative), and is therefore consistent with prior research concerning the induction of personality traits in LLMs [14, 37, 40]. The experiment also generates a control or 'baseline' set of outputs used for comparison. The baseline outputs are generated in the same way but the LLM is not provided with a personality statement to consider.

The work thereby exploits the inherent non-determinism in LLMs, characterised by variability in outputs even with identical prompts [26]. Moreover, it also explores the impact of variation in the LLMs hyperparameters on the different induced personas. Specifically, we explore changes in the *temperature* value.

3.3 Analytical Approach

A significant challenge of this experiment is how to measure the effect of the induced persona on task selection. To acheive this, the to-do list is conceived as a sequence defined by the order of the unique identifiers (UIDs) of its activities. The complete set of decision-cycles is considered a *transformation process*. Finally, the completed list is treated as a newly generated transformed sequence defined by the new order of UIDs. This conceptualisation enables the analysis of two key properties of the transformation process. The first of concern is the *plausibility* of the transformation, that is, to what extent is the transformation 'behaviour' consistent with the induced personality trait. The second is determining the extent to which this activity selection process transforms the to-do list. By having measures of this transformation, it becomes possible to assess whether persona-based transformation is statistically significant from the baseline (no induced person) transformation. The analytical approaches here represent a novel contribution to approaches to systematically assess the impact of prompt engineering.

To assess plausibility, we use measurement of *movement deltas*, which quantify the shifts in activity positions before and after the persona-based transformation. The UID of the activity is used to index and calculate the movement delta based on the initial and the resultant transformed position. Negative movement deltas indicates an activity is moved earlier and therefore considerered to be prioritised, and a positive movement delta indicates an activity is moved later and therefore is deprioritised. The traits of C̲onscientiousness and E̲xtraversion serve as the central focus in this analysis, which are hypothesised to affect task selections related to work efficiency and social interaction, respectively [13, 18, 50].

Given the to-do and completed lists are treated as sequences defined by their UIDs, it is possible to apply standard approaches of

sequence comparison for similarity and difference. Prior to the experimentation, it was not clear which standard measure of sequence similarity would yield the greatest analytical worth. Therefore, the following approaches were used to to produce quantitative assessments of the sequence transformation:

- **Longest Common Substring (LCSS)**: Finds the longest sequence of consecutive elements common to both sequences. Assesses impact on maintaining uninterrupted task sequences, focusing on contiguous matches.
- **Longest Common Prefix (LCP)**: Finds the longest initial segment common to both sequences. Reflects initial decision-making patterns by comparing the start of both sequences.
- **Levenshtein Distance**: Minimum number of single-character edits (insertions, deletions, or substitutions) required to transform one sequence into another. Provides a comprehensive view of all changes needed [53].
- **Longest Common Subsequence (LCS)**: Measures similarity by finding the longest subsequence common to both strings. Indicates order preservation in task selection by identifying deletions and additions [29].
- **Similarity Ratio (SR)**: Normalised measure derived from LCS length. Evaluates similarity between to-do and completed lists by considering deletions and additions.
- **Hamming Distance**: Number of positions at which the corresponding elements differ. Applicable only for sequences of the same length and focuses on substitutions.

For each directional personality trait, 500 measurements of each metric were produced enabling a statistical significance test of the impact of the persona induction against the baseline.

4 RESULTS

In this section, we present the outcomes of our analyses, which are divided into two key parts corresponding to the analytical approaches outlined earlier. Firstly, we explore the plausibility of the observed movement delta shifts in task prioritisation by the LLM after the induction of specific personality traits. This analysis focuses on evaluating whether the LLM's behaviour aligns with established psychological understandings of C̲onscientiousness and E̲xtraversion, particularly in relation to work-oriented and socially-related tasks, respectively. The goal here is to assess whether the LLM prioritises tasks in a manner consistent with the traits it has been induced with, thereby providing insights into the model's ability to mimic human-like decision-making patterns. Subsequently, we turn our attention to a comprehensive statistical analysis that examines the broader impact of inducing personality traits across all five dimensions of the OCEAN model [6, 21], adopting a similar approach to previous research on personality traits in LLMs [14, 24, 37, 40]. Using a suite of quantitative measures specifically designed to assess transformations in the overall ordering of tasks, we evaluate how the induced personality traits influence the LLM's task selection behaviour compared to a baseline condition, where no specific traits were introduced. This approach provides a robust framework for measuring changes in the sequential structure of the schedules, allowing us to quantify the consistency and reliability of personality-driven modifications in LLMs.

4.1 Plausibility Analysis

As discussed, to evaluate the impact regarding the plausibility of induced personality traits on the target LLMs, the *movement delta shifts* of individual tasks is analysed. As previously defined, *movement delta shifts* are the direction and magnitude of the change in task order from the initial to-do list to the completed list after the LLM processes and transforms the task execution sequence. These shifts are examined to determine if they align with the expected priorities of the induced traits. For example, an LLM with high C̲onscientiousness (**CON-POS**) is expected, given the personality schema and method of induction, to prioritise tasks requiring organisation, discipline, and hard work, such as work-related activities [50]. Movement delta shifts are quantified by calculating the mean shift (μ) for each task, where negative μ values indicate prioritisation (the task appears earlier in the completed list), and positive μ values suggest deprioritisation.

In this analysis, we hypothesise that an LLM induced with high C̲onscientiousness will prioritise work-centric tasks, while one induced with high E̲xtraversion will favour socially engaging activities. In our analysis, tasks relevant to these traits are highlighted in yellow, while others are shown in blue. Given the advanced reasoning capabilities of the GPT-4o model, our analysis is restricted to this model, aiming to determine the plausibility of the LLM's decision-making behaviour in aligning with expected task prioritisation patterns.

4.1.1 Conscientiousness: Work-related Activities. Work-related tasks (highlighted in yellow) made available to the LLM include: Email, Planning, Work, Team Collaboration, Meeting, Research. Figure 2 below illustrates the movement delta shift for all tasks following **positive** induction of C̲onscientiousness.
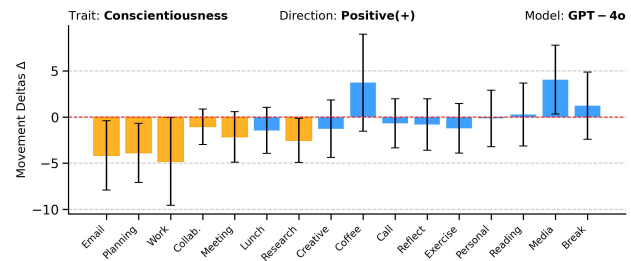

Figure 2: Movement Deltas: Positive C̲onscientiousness (GPT-4o).

The degree of *plausibility* is inferred by comparing the LLM's task prioritisation to established psychological understandings of the specific trait induced. To justify our focus on the trait of C̲onscientiousness in the context of work-related tasks, it is a trait identified as a key non-cognitive predictor of occupational performance and encompasses traits such as diligence, responsibility, and self-control [13, 35, 50]. Therefore, following a positively-reinforced induction of this trait, it is logical to expect prioritisation of work-related tasks. This predicted effect is clearly demonstrated in Figure 2 where the LLM (GPT-4o) clearly favours work-related tasks over others which may be considered non-work-related, such as Call, Reflective Time, and Exercise.

In contrast, an individual low in Conscientiousness may display disorganisation, a lack of discipline, and general disregard for work-related responsibilities [13, 36, 50]. As shown in Figure 3, the LLM deprioritises work-related tasks in favour of non-work-related activities when negatively induced, such as Personal Time, Reading, and Social Media, demonstrating a significant shift in behaviour.
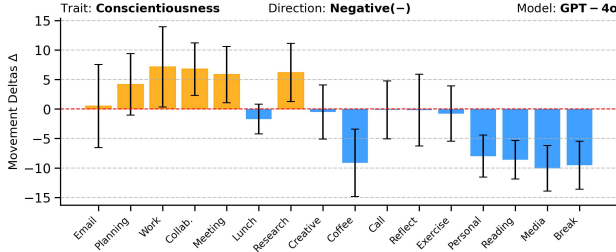

Figure 3: Movement Deltas: Negative Conscientiousness (GPT-4o).

4.1.2 Movement Delta Shift: Work. To provide a more comprehensive and task-specific analysis, we examine the movement deltas for the *Work* task across all experimental conditions, as shown in Figure 4. This figure compares the shifts observed for the GPT-4o model (blue bars) with those for the GPT-3.5-Turbo model (red bars), with error bars indicating the standard deviations.
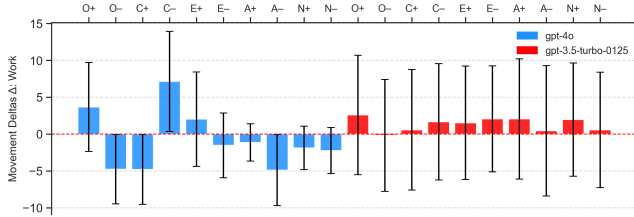

Figure 4: Movement Deltas: Work (GPT-4o & GPT-3.5-Turbo)

Figure 4, above, highlights a stark contrast in behaviour between GPT-4o and GPT-3.5-Turbo. GPT-4o shows greater alignment with expected behaviour based on the induced traits, particularly Conscientiousness, exhibiting more dynamic movement delta shifts, whereas GPT-3.5-Turbo appears more deterministic, with minimal notable effects across the conditions.

4.1.3 Extraversion: Social Activities. Socially-related tasks (highlighted in yellow) made available to the LLM include: Team Collaboration, Meeting, Call, and Social Media. Figure 5 represents the movement delta shift following positive induction of Extraversion.

As expected, the LLM prioritises social tasks, reflecting the core aspects of Extraversion, such as sociability and reward sensitivity [18]. This is evident in the prioritisation of tasks like Team Collaboration, Meeting, and Call, along with other related activities such as Coffee Break, with slight increases in Exercise and Personal Time.

Conversely, as shown in Figure 6, when negatively-reinforcing the trait of Extraversion, the LLM deprioritises socially engaging tasks, instead favouring more solitary activities such as Reflective
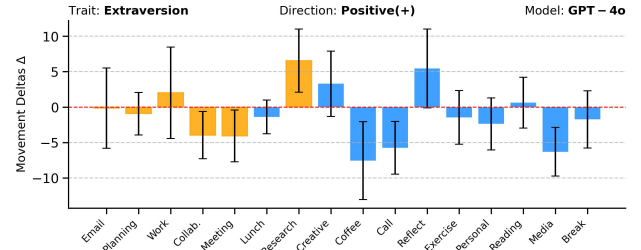

Figure 5: Movement Deltas: Positive Extraversion (GPT-4o).

Time, Personal Time, and Reading. This shift suggests that the LLM adopts introverted tendencies, prioritising tasks that are less socially oriented.
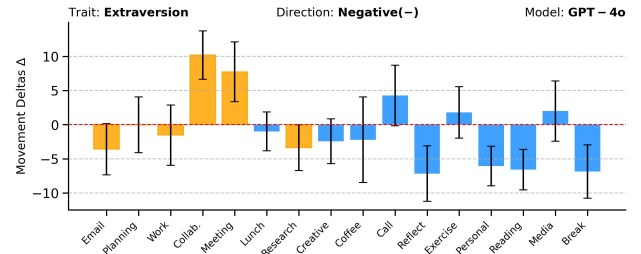

Figure 6: Movement Deltas: Negative Extraversion (GPT-4o).

In summary, these results suggest that the method of trait induction used in this experiment effectively steers the LLM's behaviour in a manner consistent with the induced personality traits. The LLM's prioritisation of tasks is driven by the personality it has been induced with, without direct instruction to prioritise tasks relevant to that personality. For instance, after positive induction with Extraversion, the LLM naturally favors social-related tasks, leading to their earlier appearance in the completed list, as indicated by negative mean (μ) movement delta shift values.
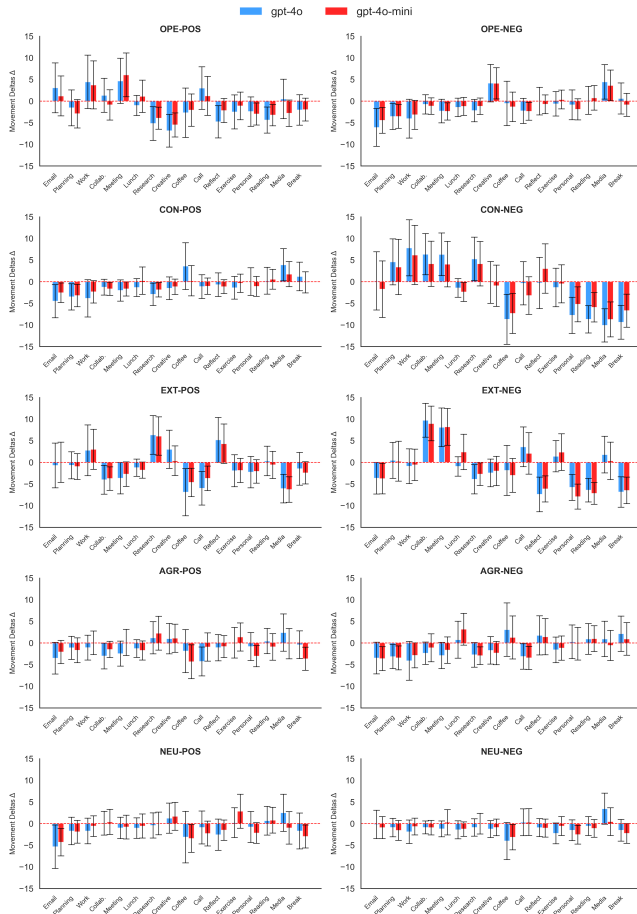
4.2 Transformation Analysis

Experimental results are normalised to facilitate analysis and enhance the clarity of visualisations. All quantitative outcomes are scaled from 0 to 1, enabling a clear comparison of the LLM's task selection behaviour. These results illustrate whether the LLM tends toward *perfect alignment* with the to-do list or *no alignment*. For similarity measures such as LCSS, LCP, and SR, a value of 1 indicates perfect alignment with the to-do list, signifying that the LLM follows the original sequence exactly. In contrast, for distance measures like LEV and HAM, a value of 0 represents perfect alignment, meaning no deviations from the original sequence are observed.

4.2.1 Effect of Sampling Temperature on Sequence Alignment. Inference hyperparameters such as sampling temperature, top-k sampling, repetition penalty, and maximum token length can be fine-tuned to modify the LLMs output at runtime [34, 43]. Whilst the focus of this study is centred on the effect of induced personality

Table 1: Normalised GPT-4o and GPT-4o-Mini results with adjusted sampling temperatures on neutral personality experiment. Temperature (τ) range: `0.0` to `1.6`. Values are aggregated on the entire sample of pre-generated schedules ($n = 500$).

τ	GPT-4o (μ / σ)					τ	GPT-4o-Mini (μ / σ)				
	LCSS	LCP	LEV	SR	HAM		LCSS	LCP	LEV	SR	HAM
0.0	$0.635_{0.224}$	$0.604_{0.260}$	$0.201_{0.144}$	$0.321_{0.220}$	$0.873_{0.095}$	0.0	$0.522_{0.249}$	$0.476_{0.288}$	$0.296_{0.178}$	$0.393_{0.245}$	$0.829_{0.111}$
0.2	$0.633_{0.245}$	$0.599_{0.268}$	$0.207_{0.148}$	$0.313_{0.224}$	$0.875_{0.094}$	0.2	$0.506_{0.252}$	$0.473_{0.285}$	$0.305_{0.181}$	$0.405_{0.244}$	$0.824_{0.116}$
0.4	$0.624_{0.242}$	$0.593_{0.274}$	$0.209_{0.147}$	$0.310_{0.228}$	$0.879_{0.093}$	0.4	$0.515_{0.250}$	$0.473_{0.289}$	$0.287_{0.175}$	$0.384_{0.229}$	$0.836_{0.111}$
0.6	$0.588_{0.222}$	$0.561_{0.252}$	$0.232_{0.150}$	$0.328_{0.213}$	$0.866_{0.097}$	0.6	$0.501_{0.262}$	$0.458_{0.293}$	$0.317_{0.202}$	$0.408_{0.248}$	$0.819_{0.127}$
0.8	$0.578_{0.237}$	$0.558_{0.258}$	$0.235_{0.154}$	$0.339_{0.229}$	$0.857_{0.103}$	0.8	$0.494_{0.242}$	$0.450_{0.283}$	$0.320_{0.188}$	$0.429_{0.249}$	$0.806_{0.127}$
1.0	$0.562_{0.218}$	$0.527_{0.255}$	$0.250_{0.157}$	$0.349_{0.231}$	$0.851_{0.104}$	1.0	$0.454_{0.234}$	$0.402_{0.271}$	$0.351_{0.205}$	$0.449_{0.247}$	$0.799_{0.127}$
1.2	$0.541_{0.234}$	$0.505_{0.267}$	$0.263_{0.157}$	$0.366_{0.227}$	$0.842_{0.106}$	1.2	$0.470_{0.231}$	$0.428_{0.263}$	$0.337_{0.188}$	$0.432_{0.228}$	$0.805_{0.121}$
1.4	$0.541_{0.236}$	$0.502_{0.272}$	$0.277_{0.165}$	$0.392_{0.240}$	$0.839_{0.105}$	1.4	$0.408_{0.220}$	$0.356_{0.255}$	$0.376_{0.177}$	$0.494_{0.233}$	$0.775_{0.121}$
1.6	$0.464_{0.223}$	$0.420_{0.259}$	$0.318_{0.168}$	$0.446_{0.238}$	$0.810_{0.108}$	1.6	$0.383_{0.206}$	$0.316_{0.250}$	$0.412_{0.204}$	$0.530_{0.237}$	$0.751_{0.132}$

Figure 7: Aggregated movement deltas per experimental condition. X-axis: Tasks. Y-axis: Average (μ) movement delta shift, ranging from -15 to +15. Error bars denote deviation (σ).

sampling temperature and top-p (nucleus sampling) are directly relevant as these are hyperparameters used to control the randomness and creativity of generated text [25, 34]. For GPT-based models, it is advised to only adjust one of these parameters, not both [25]. Adjustment of sampling temperature is therefore opted for here.

To empirically investigate the influence adjusted sampling temperature has upon the performance of a GPT-based LLM in the context of our experiment, we perform the experiment outlined in Section 3, albeit with a strict focus on the control condition–where no personality is induced. The purpose of this evaluation is to test the working hypothesis that increased temperature settings within GPT-based models result in greater levels of non-determinism [25, 26] whilst simultaneously determining the robustness of our measures. By systematically adjusting the sampling temperature, we can observe whether the LLMs adherence to the original to-do list diminishes as randomness increases, thereby validating the effectiveness of our chosen metrics in capturing alignment with the task sequence.

The results, presented in Table 1, demonstrate a clear inverse relationship between temperature settings and the LLMs' adherence to the original to-do list across both GPT-4o and GPT-4o-Mini models. Notably, GPT-4o consistently starts with higher alignment metrics, such as LCSS (0.652) and LCP (0.615) at T0.0, compared to GPT-4o-Mini, which begins at 0.528 and 0.489, respectively. These initial differences suggest that GPT-4o is inherently more capable of maintaining task sequences, likely due to its larger scale or more robust sequence retention abilities [25]. As temperature increases from 0.0 to 1.6, both models show a consistent decrease in LCSS, LCP, and SR, reflecting reduced similarity with the original task order. Conversely, the Levenshtein Distance (LEV) increases with higher temperatures in both models, indicating a greater number of edits needed to align the completed tasks with the to-do list. Interestingly, Hamming Distance (HAM) shows a slight decrease in both models, suggesting that while overall task order becomes more randomised, specific tasks may still align by chance at higher temperatures. Despite these nuances, the findings robustly support the hypothesis that higher temperatures introduce greater variability and less determinism in the LLMs' task selection behaviour, with GPT-4o-Mini demonstrating a baseline tendency towards less deterministic outputs even at lower temperatures.

traits, it remains significant to acknowledge the influence of these hyperparameters with regards to the LLMs performance. Hyperparameters affecting repetition penalties and token lengths are irrelevant in this experiment due to its programmatic design. However,

Table 2: Normalised GPT-4o and GPT-4o-Mini results per metric with $\tau = 1.0$ across all groups (OCEAN → Low/High).

T	Dir.	GPT-4o (μ / σ)					T	Dir.	GPT-4o-Mini (μ / σ)				
		LCSS	LCP	LEV	SR	HAM			LCSS	LCP	LEV	SR	HAM
O	High	$0.12_{0.05}$	$0.03_{0.04}$	$0.79_{0.10}$	$0.87_{0.09}$	$0.46_{0.09}$	O	High	$0.14_{0.05}$	$0.05_{0.06}$	$0.72_{0.11}$	$0.86_{0.10}$	$0.53_{0.08}$
	Low	$0.19_{0.10}$	$0.16_{0.12}$	$0.65_{0.12}$	$0.74_{0.12}$	$0.55_{0.10}$		Low	$0.20_{0.11}$	$0.16_{0.13}$	$0.62_{0.12}$	$0.73_{0.13}$	$0.59_{0.10}$
C	High	$0.22_{0.12}$	$0.18_{0.14}$	$0.60_{0.15}$	$0.70_{0.15}$	$0.61_{0.11}$	C	High	$0.28_{0.17}$	$0.24_{0.18}$	$0.48_{0.18}$	$0.61_{0.20}$	$0.71_{0.12}$
	Low	$0.10_{0.03}$	$0.00_{0.01}$	$0.89_{0.07}$	$0.93_{0.07}$	$0.32_{0.06}$		Low	$0.10_{0.04}$	$0.00_{0.01}$	$0.88_{0.07}$	$0.94_{0.06}$	$0.38_{0.07}$
E	High	$0.12_{0.06}$	$0.03_{0.07}$	$0.79_{0.10}$	$0.87_{0.09}$	$0.45_{0.08}$	E	High	$0.14_{0.05}$	$0.03_{0.04}$	$0.74_{0.09}$	$0.88_{0.09}$	$0.51_{0.07}$
	Low	$0.12_{0.05}$	$0.05_{0.06}$	$0.81_{0.09}$	$0.87_{0.10}$	$0.43_{0.07}$		Low	$0.12_{0.04}$	$0.07_{0.05}$	$0.82_{0.08}$	$0.88_{0.08}$	$0.44_{0.08}$
A	High	$0.20_{0.10}$	$0.17_{0.13}$	$0.62_{0.11}$	$0.71_{0.12}$	$0.59_{0.09}$	A	High	$0.19_{0.08}$	$0.14_{0.10}$	$0.62_{0.13}$	$0.76_{0.13}$	$0.61_{0.09}$
	Low	$0.17_{0.08}$	$0.11_{0.10}$	$0.66_{0.12}$	$0.78_{0.11}$	$0.56_{0.10}$		Low	$0.15_{0.05}$	$0.07_{0.07}$	$0.68_{0.11}$	$0.82_{0.11}$	$0.57_{0.08}$
N	High	$0.22_{0.10}$	$0.16_{0.13}$	$0.61_{0.12}$	$0.73_{0.14}$	$0.59_{0.10}$	N	High	$0.18_{0.08}$	$0.14_{0.10}$	$0.61_{0.12}$	$0.73_{0.13}$	$0.60_{0.09}$
	Low	$0.36_{0.16}$	$0.33_{0.19}$	$0.46_{0.14}$	$0.57_{0.17}$	$0.71_{0.10}$		Low	$0.30_{0.16}$	$0.27_{0.18}$	$0.48_{0.15}$	$0.58_{0.17}$	$0.70_{0.11}$
B	N/A	$0.56_{0.22}$	$0.53_{0.26}$	$0.25_{0.16}$	$0.35_{0.23}$	$0.85_{0.10}$	B	N/A	$0.45_{0.23}$	$0.40_{0.27}$	$0.35_{0.20}$	$0.45_{0.25}$	$0.80_{0.13}$

4.2.2 Transformation Significance: Experimental Conditions vs. Control.

The sequence difference (or similarity) metrics were also applied to assess whether the induction of a personality has a material impact on the populations of the measurements of the selected metrics. For this analysis, the LLMs GPT-4o, GPT-4o-mini and GPT-3.5-Turbo were assessed using the experimental framework outlined previously. Figure 8 illustrates the corresponding distributions for GPT-4o per each of the key metrics (LCSS, LEV, HAMMING, RATIO) only. GPT-4o-Mini produced similar outputs to GPT-4o which is expected given it is a more compact model, sacrificing some performance for greater accessibility and affordability. Thus, visualised results for GPT-4o-Mini were excluded. In relation to GPT-3.5-Turbo, a predecessor model, the observed results across all metrics did not differ substantially, despite being found to be statistically significant in most cases (See 4.2.4).

In this analysis we seek to assess whether the induction of a persona creates a statistically significant different population of the metrics when compared to the baseline. The Null hypothesis states there will be no difference to the baseline population with the Alternative Hypothesis being that the induced personality trait will produce a different population of measurements. To test this, an independent two-sample Welch's t-test is performed between each experimental condition and the control for each measure. For a singular model (*i.e.*, GPT-4o), this results in 50 pairwise comparisons (5 traits x 2 directions x 5 measures). For example, "LCSS for Openness (High)" is compared against "LCSS for Baseline". To control for the family-wise error rate (FWER) due to multiple comparisons, the Bonferroni correction is applied, controlled by the following:

$$\text{FWER} = P\left(\bigcup_{i=1}^{m_0} \left\{ p_i \le \frac{\alpha}{m} \right\}\right) \le \sum_{i=1}^{m_0} P\left(p_i \le \frac{\alpha}{m}\right) = m_0 \frac{\alpha}{m} \le \alpha$$

where p_i is the p-value for the i-th test, $m = 50$ is the total number of tests, and $\alpha = 0.05$ is the overall significance level. With 50 tests, the alpha level ($p \le 0.05$) is then adjusted to $p \le 0.001$.

4.2.3 GPT-4o and GPT-4o-Mini.

For GPT-4o, the Null hypothesis is rejected on all experimental conditions, potentially attributed toward the significantly different results for the baseline group on all measures. For GPT-4o-Mini, the Null hypothesis is rejected on all

experimental conditions. Similar to GPT-4o, this may be attributed toward the significantly different results for the baseline group on all measures. All results are displayed in Table 2.

4.2.4 GPT-3.5-Turbo.

Null hypothesis is rejected on 41 experimental conditions, accepted on 9. Statistically insignificant results, displayed in Table 3, are not emboldened and marked with an asterisk (*). Whilst the statistical test results indicate a significantly different task ordering for the majority of experimental conditions and quantitative measures, the absolute difference, between all experimental conditions and that of the control, are far less considerable when compared to GPT-4o and GPT-4o-Mini (Table 2).

Table 3: Normalised GPT-3.5-Turbo results per metric with $\tau = 1.0$ across all experimental groups (OCEAN → Low/High). B = Control[1].

		GPT-3.5-Turbo (μ / σ)				
		LCSS	LCP	LEV	SR	HAM
O	High	$0.111_{0.04}$	$0.001^*_{0.006}$	$0.859_{0.082}$	$0.93_{0.064}$	$0.404_{0.077}$
	Low	$0.116^*_{0.048}$	$0.008_{0.023}$	$0.817_{0.089}$	$0.913_{0.07}$	$0.447_{0.077}$
C	High	$0.116^*_{0.042}$	$0.002_{0.011}$	$0.823_{0.091}$	$0.91_{0.072}$	$0.432_{0.076}$
	Low	$0.112_{0.042}$	$0.002_{0.011}$	$0.841_{0.085}$	$0.915_{0.073}$	$0.41_{0.076}$
E	High	$0.115_{0.039}$	$0.001_{0.007}$	$0.83_{0.089}$	$0.916_{0.07}$	$0.432_{0.079}$
	Low	$0.115_{0.041}$	$0.006^*_{0.019}$	$0.855_{0.083}$	$0.925_{0.07}$	$0.416_{0.072}$
A	High	$0.117^*_{0.04}$	$0.004_{0.016}$	$0.822_{0.09}$	$0.914_{0.074}$	$0.438_{0.076}$
	Low	$0.114_{0.043}$	$0.003_{0.012}$	$0.843_{0.084}$	$0.928_{0.063}$	$0.42_{0.078}$
N	High	$0.112_{0.041}$	$0.005^*_{0.02}$	$0.839_{0.085}$	$0.92_{0.072}$	$0.425_{0.072}$
	Low	$0.119^*_{0.047}$	$0.005^*_{0.018}$	$0.802_{0.098}$	$0.913^*_{0.078}$	$0.458_{0.088}$
B[1]	N/A	$0.124_{0.044}$	$0.008_{0.025}$	$0.776_{0.093}$	$0.892_{0.082}$	$0.472_{0.081}$

5 CONCLUSION

This study proposed a method to quantitatively measure the effect of prompt-based personality induction on LLM-based agent decision-making in task selection, scheduling, and planning. While the display of synthetic personality in LLMs is known, no studies have empirically focused on evaluating how inducing personality
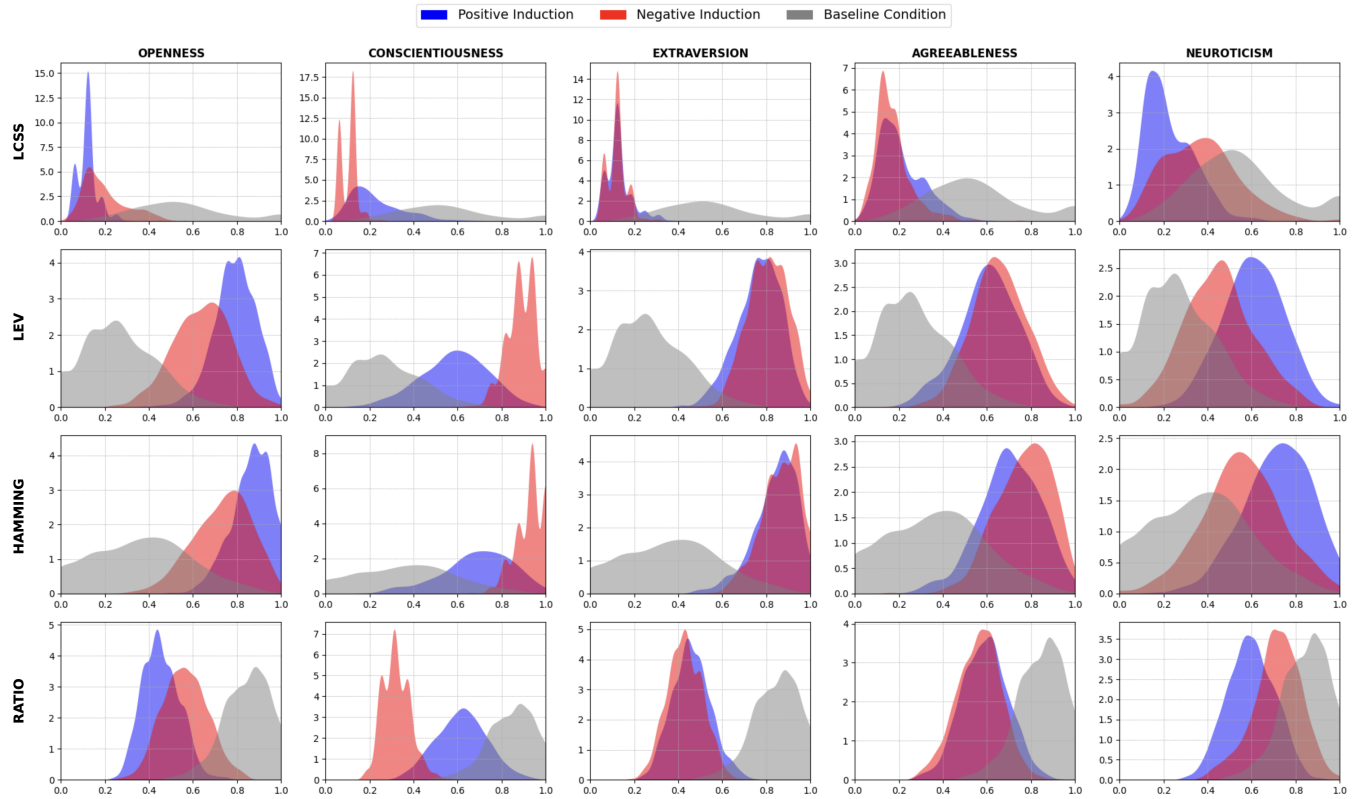
Figure 8: Kernel density estimation plots for GPT-4o across all experimental groups (OCEAN → Low/High) per quantitative measure, including the control. Each sub-plot visually represents the GPT-4o results provided in Table 2.

traits based on the Five-Factor OCEAN model impacts critical agent behaviours governing decision-making, particularly task selection.

Experiments with all OCEAN traits and a neutral control group showed that personality induction leads to significant differences in task prioritisation across GPT-4o, GPT-4o-Mini, and GPT-3.5-Turbo models. Effects were more pronounced in GPT-4o models, indicating their greater capacity for reasoning and exhibiting these traits. Analysis specifically examined the *degree of plausibility* in how the LLM-based agent prioritised tasks, assessing whether the agent's decisions aligned with contemporary psychological understandings of each trait. Our results showed that traits like *Openness*, *Conscientiousness*, and *Extraversion* substantially impacted task prioritisation, causing significant deviations from original schedules. In contrast, *Agreeableness* and *Neuroticism* had less pronounced effects, possibly due to LLMs being less receptive to these traits or due to limitations in task relevance.

These findings show that large-scale, pre-trained language models like GPT-4o can exhibit personality traits, confirming previous studies and demonstrating impact within a downstream task. Moreover, our validated personality induction method highlights potential for enhancing autonomous agents in planning and scheduling, especially in systems performing tasks akin to humans. Our contributions lay a foundation for integrating nuanced human-like behaviours into autonomous systems, enhancing their effectiveness in environments requiring sophisticated, plausible decision-making.

5.1 Ethical Considerations

This research aims to enhance the development of Deceptive Agents, a novel class of LLM-based autonomous agents which are intended to effectuate highly plausible simulacra of humans interacting with digital systems for cyber defense through strategic deception [24]. Ethically, this approach operates on the principle of 'rightful deception', where targets have no legitimate claim to truth or transparency due to their unethical intent to access or damage systems without authorisation. Whilst the intended use is strictly within a defensive context, there exists the possibility that this work could facilitate the generation of misinformation or enable tailored persuasion tactics, thereby exacerbating issues related to manipulation and deceit in broader contexts [27]. Specifically, the ability to induce personality traits within LLMs could be exploited to create more convincing deceptive content, potentially undermining public trust. It is therefore essential to implement safeguards and adhere to ethical guidelines in any research or application involving the induction of personality traits within LLMs to mimic human values, thought patterns, and behaviour. This includes ensuring robust oversight, bias mitigation, and alignment to prevent misuse.

ACKNOWLEDGMENTS

REFERENCES

[1] Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*. PMLR, 337–371.

[2] Markus Bayer, Tobias Frey, and Christian Reuter. 2023. Multi-level fine-tuning, data augmentation, and few-shot learning for specialized cyber threat intelligence. *Computers & Security* 134 (2023), 103430.

[3] Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences* 120, 6 (2023), e2218523120.

[4] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* (2023).

[5] T. Bösser. 2001. Autonomous Agents. In *International Encyclopedia of the Social and Behavioral Sciences*, Neil J. Smelser and Paul B. Baltes (Eds.). Pergamon, Oxford, 1002–1006. https://doi.org/10.1016/B0-08-043076-7/00534-9

[6] Paul T Costa and Robert R McCrae. 1999. A five-factor theory of personality. *The five-factor model of personality: Theoretical perspectives* 2 (1999), 51–87.

[7] Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2022. Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051* (2022).

[8] Stan Franklin and Art Graesser. 1996. Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents. In *International workshop on agent theories, architectures, and languages*. Springer, 21–35.

[9] Lewis R Goldberg. 1981. Language and individual differences: The search for universals in personality lexicons. *Review of personality and social psychology* 2, 1 (1981), 141–165.

[10] Lewis R Goldberg et al. 1999. A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality psychology in Europe* 7, 1 (1999), 7–28.

[11] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680* (2024).

[12] Andreas Happe and Jürgen Cito. 2023. Getting pwn'd by ai: Penetration testing with large language models. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 2082–2086.

[13] Joyce Hogan and Deniz S Ones. 1997. Conscientiousness and integrity at work. In *Handbook of personality psychology*. Elsevier, 849–870.

[14] Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2024. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems* 36 (2024).

[15] Oliver P John, Eileen M Donahue, and Robert L Kentle. 1991. Big five inventory. *Journal of personality and social psychology* (1991).

[16] John A Johnson. 2014. Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of research in personality* 51 (2014), 78–89.

[17] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2015. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971* (2015).

[18] Richard E Lucas, Ed Diener, Alexander Grob, Eunkook M Suh, and Liang Shao. 2000. Cross-cultural evidence for the fundamental features of extraversion. *Journal of personality and social psychology* 79, 3 (2000), 452.

[19] Zeyang Ma, An Ran Chen, Dong Jae Kim, Tse-Hsun Chen, and Shaowei Wang. 2024. LLMParser: An Exploratory Study on Using Large Language Models for Log Parsing. In *2024 IEEE/ACM 46th International Conference on Software Engineering (ICSE)*. IEEE Computer Society, 883–883.

[20] Robert R McCrae and Paul T Costa Jr. 1997. Personality trait structure as a human universal. *American psychologist* 52, 5 (1997), 509.

[21] Robert R McCrae and Oliver P John. 1992. An introduction to the five-factor model and its applications. *Journal of personality* 60, 2 (1992), 175–215.

[22] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *nature* 518, 7540 (2015), 529–533.

[23] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332* (2021).

[24] Lewis Newsham, Ryan Hyland, and Daniel Prince. 2024. Measuring the Effect of Induced Persona on Agenda Creation in Language-based Agents for Cyber Deception". In *Natural Language Processing and Artificial Intelligence for Cyber Security, NLPAICS 2024*. https://eprints.lancs.ac.uk/id/eprint/222820/1/SANDMAN_NLPAICS_2024_.pdf

[25] OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article* 2 (2023), 13.

[26] Shuyin Ouyang, Jie M Zhang, Mark Harman, and Meng Wang. 2023. LLM is Like a Box of Chocolates: the Non-determinism of ChatGPT in Code Generation. *arXiv preprint arXiv:2308.02828* (2023).

[27] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–22.

[28] Joon Sung Park, Lindsay Popowski, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social Simulacra: Creating Populated Prototypes for Social Computing Systems. arXiv:2208.04024 [cs.HC]

[29] Mike Paterson and Vlado Dančík. 1994. Longest common subsequences. In *International symposium on mathematical foundations of computer science*. Springer, 127–142.

[30] Jeffrey Pawlick, Edward Colbert, and Quanyan Zhu. 2019. A game-theoretic taxonomy and survey of defensive deception for cybersecurity and privacy. *ACM Computing Surveys (CSUR)* 52, 4 (2019), 1–28.

[31] Max Pellert, Clemens M Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. 2023. Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science* (2023), 17456916231214460.

[32] Chen Qian, Xin Cong, Wei Liu, Cheng Yang, Weize Chen, Yusheng Su, Yufan Dang, Jiahao Li, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Communicative Agents for Software Development. arXiv:2307.07924 [cs.SE]

[33] Boele De Raad, Marco Perugini, Martina Hrebíčková, and Piotr Szarota. 1998. Lingua Franca of Personality: Taxonomies and Structures Based on the Psycholexical Approach. *Journal of Cross-Cultural Psychology* 29, 1 (1998), 212–232. https://doi.org/10.1177/0022022198291011

[34] Matthew Renze and Erhan Guven. 2024. The effect of sampling temperature on problem solving in large language models. *arXiv preprint arXiv:2402.05201* (2024).

[35] Brent W Roberts, Carl Lejuez, Robert F Krueger, Jessica M Richards, and Patrick L Hill. 2014. What is conscientiousness and how can it be assessed? *Developmental psychology* 50, 5 (2014), 1315.

[36] Ivan T Robertson, Helen Baron, Patrick Gibbons, Rab MacIver, and Gill Nyfield. 2000. Conscientiousness and managerial performance. *Journal of Occupational and Organizational Psychology* 73, 2 (2000), 171–180.

[37] Mustafa Safdari, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja Mataríc. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184* (2023).

[38] Gerard Saucier and Lewis R Goldberg. 2001. Lexical studies of indigenous personality factors: Premises, products, and prospects. *Journal of personality* 69, 6 (2001), 847–879.

[39] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[40] Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Mataríc. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184* (2023).

[41] Febrian Setianto, Erion Tsani, Fatima Sadiq, Georgios Domalis, Dimitris Tsakalidis, and Panos Kostakos. 2021. GPT-2C: A parser for honeypot logs using large pre-trained language models. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 649–653.

[42] Theodore Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas Griffiths. 2023. Cognitive Architectures for Language Agents. *Transactions on Machine Learning Research* (2023).

[43] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[44] Dustin D Updyke, Geoffrey B Dobson, Thomas G Podnar, Luke J Osterritter, Benjamin L Earl, and Adam D Cerini. 2018. Ghosts in the machine: A framework for cyber-warfare exercise npc simulation. *Technical report* (2018).

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[46] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291* (2023).

[47] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2023. A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432* (2023).

[48] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science* 18, 6 (2024), 1–26.

[49] Taylor Webb, Keith J Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour* 7, 9 (2023), 1526–1541.

[50] Michael P Wilmot and Deniz S Ones. 2019. A century of research on conscientiousness at work. *Proceedings of the National Academy of Sciences* 116, 46 (2019), 23004–23010.

[51] Lionel Wong, Gabriel Grand, Alexander K Lew, Noah D Goodman, Vikash K Mansinghka, Jacob Andreas, and Joshua B Tenenbaum. 2023. From word models to world models: Translating from natural language to the probabilistic language of thought. *arXiv preprint arXiv:2306.12672* (2023).

[52] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629* (2022).

[53] Li Yujian and Liu Bo. 2007. A normalized Levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence* 29, 6 (2007), 1091–1095.

[54] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics* 50, 1 (2024), 237–291.