Measuring LLM Behaviour: A Context-Independent Sampling Framework for Questionnaire-Based Evaluation

Prabhash Rathnayake[1,*], Judd D. Bradbury[2], Rosanna E. Guadagno[1], Rodrigo Ortega-Casellas[1] and Ella Peltonen[1]

[1]SEIS, ITEE, University of Oulu, P.O. Box 4500, FI-90014 University of Oulu, Finland
[2]Department of Business Analytics and AI, , The University of Texas at Dallas, 800 W. Campbell Road, Richardson, TX 75080, USA

Abstract

Large language models (LLMs) are increasingly evaluated using questionnaire-style prompts to probe behavioural tendencies such as decision-making, ethical reasoning, and alignment responses. However, existing approaches often rely on conversational interfaces or multi-turn interactions, introducing sequential dependencies and contextual carryover that compromise measurement validity and reproducibility. We introduce a context-independent sampling framework for measuring LLM behaviour, in which each questionnaire item is executed as an isolated API request under controlled system conditions. This design enforces response independence, eliminates conversational drift, and enables reproducible behavioural measurement. The framework further incorporates role-conditioned prompting, structured response schemas, and systematic logging of refusal behaviour, allowing consistent comparison across models and experimental settings. We evaluate the framework through a large-scale study of thirteen contemporary LLMs using multiple behavioural questionnaires under different role conditions. The results show that behavioural responses vary systematically across both models and roles, and that refusal patterns provide a measurable signal of alignment constraints and safety mechanisms. This work establishes a reproducible methodology for measuring behavioural signatures in LLM outputs and provides a foundation for systematic, non-anthropomorphic analysis of model behaviour across systems and deployment contexts.

Keywords

Large language models, Questionnaire-based probing, Context-independent API sampling, LLM behaviour analysis

CEUR-WS.org/Vol-4309/paper01.pdf

1. Introduction

Large language models (LLMs) have rapidly evolved from research prototypes to widely deployed systems that shape decision-making, knowledge work, communication, and creative processes [1]. As tutors, assistants, analysts, and embedded agents, they influence not only task performance but also how information is generated, framed, and interpreted, including recommendations, uncertainty handling, safety enforcement, and user guidance.

Despite their growing impact, current evaluation frameworks primarily focus on capability (e.g., reasoning, coding, summarisation) and safety compliance (e.g., filtering harmful content). Although essential, these dimensions do not fully capture how models behave across interactions. LLMs exhibit recurring response patterns in how they handle ambiguity, prioritise safety, balance helpfulness and refusal, and adapt to user roles. These behavioural tendencies directly influence user trust, decision-making, and system-level outcomes. In high-stakes situations, this variability creates uncertainty that can undermine reliability and accountability [2].

The behaviour patterns of large language models emerge from large-scale statistical training [3] and alignment processes such as reinforcement learning from human feedback (RLHF) [4] and constitutional approaches [5]. The composition of training data and alignment procedures is often partially undisclosed

TKTP 2026: Annual Doctoral Symposium of Computer Science, 8.—9.6.2026 Turku, Finland
*Corresponding author
*Corresponding author
prabhash.rathnayake@oulu.fi (P. Rathnayake); judd.bradbury@utdallas.edu (J. D. Bradbury); rosanna.guadagno@oulu.fi (R. E. Guadagno); rodrigo.ortegacasellas@oulu.fi (R. Ortega-Casellas); ella.peltonen@oulu.fi (E. Peltonen)
0009-0007-9894-1579 (P. Rathnayake); 0000-0003-3664-2315 (J. D. Bradbury); 0000-0001-8247-5154 (R. E. Guadagno); 0009-0006-2788-1882 (R. Ortega-Casellas); 0000-0002-3374-671X (E. Peltonen)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-02

and heterogeneous [6], making it difficult to inspect or isolate the underlying drivers of behaviour. As a result, existing explainability methods remain limited in their ability to characterise how models behave in practice [7].

Consequently, researchers are increasingly relying on output-based inference to study model behaviour [8]. Empirical studies indicate that LLMs can display consistent patterns of cooperation, assertiveness, risk framing, and alignment responses when placed in controlled conditions. These findings are increasingly important as LLM-based agents move into practical application settings, including mental well-being support, recommender systems, legal assistants, and web-interaction environments [9].

Research on human-machine interaction shows that users often assign personality, intentions, and mental states to computational systems [10]. This tendency is heightened by the fluent, context-sensitive dialogue generated by large language models [11]. This creates a methodological challenge: while users may perceive these models as socially responsive agents, scientific analysis must refrain from anthropomorphic interpretations. Previous studies suggest that LLMs can exhibit consistent *behavioural signatures*, defined as recurring, measurable response patterns observed under controlled conditions, without implying human-like traits or subjective experiences [12].

A common method for investigating such behaviour involves using questionnaire-based prompts, including tools for assessing personality and decision-making. Questionnaires offer a structured and standardised way to gather comparable responses across different models, making them suitable for systematic behavioural analysis. However, many existing studies administer these questionnaires via conversational interfaces or multi-turn interactions.This approach can introduce sequential dependencies and contextual carryover, where information from earlier turns influences responses to later questionnaire items, as well as other uncontrolled system conditions. As a result, individual responses may not be treated as independent observations, which diminishes reproducibility and complicates interpretation [13]. Despite these challenges, there is currently no standardised methodology to ensure response independence in questionnaire-based evaluations of LLMs.

To address this gap, we introduce a Context-Independent Sampling (CIS) framework for measuring LLM behaviour. In this approach, each questionnaire item is executed as an isolated Application Programming Interface (API) request under fixed system conditions. This method eliminates conversational carryover and ensures that responses are independent of one another. Additionally, the framework uses structured response schemas, role-conditioned prompting, and systematic logging of refusal behaviour. This enables reproducible and comparable measurements of behaviour across different models and experimental settings.

Rather than interpreting outputs as psychological traits, we treat responses as observable behavioural signals generated under controlled conditions. This allows behavioural patterns, including refusals, to be analysed as measurable properties of model outputs shaped by training data, alignment mechanisms, and system constraints. The contributions of our work can be summarised as follows:

- We introduce *behavioural signatures* as a measurable construct for analysing recurring response patterns in LLM outputs without anthropomorphic assumptions.
- We propose a CIS protocol that enforces response independence and enables reproducible behavioural measurement.
- We demonstrate that role-conditioned prompting produces systematic variation in behavioural responses across models.
- We show that refusal behaviour serves as a measurable indication of alignment constraints and safety mechanisms.
- We establish a reproducible experimental protocol for questionnaire-based LLM evaluation, including prompt design, structured output schemas, and logging of behavioural signals.

Together, these contributions provide a foundation for reproducible behavioural evaluation of LLMs across models, versions, and deployment contexts. We next review prior work on anthropomorphic interpretation, alignment-driven behaviour, and questionnaire-based evaluation to situate the proposed framework.

2. Background

Recent research has investigated the use of structured probes, such as questionnaire-based tools, to analyse behavioural patterns in LLM outputs. These methods are adapted from psychology and behavioural science to generate standardised responses across different models and conditions. This questionnaire-based evaluation strategy enables systematic comparison of model outputs under controlled prompting conditions [8].

However, administering probes through conversational interfaces or multi-turn interactions poses methodological challenges [14]. Previous context, system prompts, and interaction history can influence responses, making them interrelated rather than independent [15]. This complicates the interpretation of questionnaire measurements and reduces reproducibility.

In addition, questionnaire-based evaluation in LLMs raises questions about construct validity and applicability. Psychometric instruments are generally designed for human respondents and assume properties such as stable identity, memory continuity, and lived experience. These assumptions do not directly transfer to artificial systems, requiring careful interpretation when applying such instruments to LLM outputs [13].

Identity-related constructs are especially difficult to transfer from humans to LLMs because they presuppose a stable sense of self, socially situated roles, and lived experience [16]. At the same time, emerging work suggests that LLMs can exhibit relatively stable model-level behavioural dispositions shaped by training data, architecture, and alignment processes, while prompt-induced roles or personas can systematically influence how those dispositions are expressed in interaction [17]. This distinction is important for questionnaire-based evaluation, because measured responses may reflect both baseline model tendencies and temporary role conditioning rather than human-like personality traits in the strict psychological sense.

More broadly, the current evaluation methods for LLMs primarily emphasise benchmark performance, bias detection, safety robustness, and the frequency of hallucinations [18, 19]. While these approaches provide valuable insights, they fall short in measuring behavioural patterns under controlled experimental conditions.

A further practical challenge arises from system-level and configuration variability. LLM outputs may be influenced by prompting strategies, system prompts, and inference parameters (e.g., temperature and sampling methods), which introduce stochasticity and affect reproducibility across experiments [9]. In addition, the lack of standardised evaluation protocols further complicates consistent comparison and replication.

Taken together, these limitations highlight the need for a controlled and reproducible methodology for behavioural evaluation. In particular, there is a lack of standardised approaches for administering structured probes, such as questionnaires, under conditions that minimise contextual dependencies and support consistent measurement across models. The following section introduces a CIS framework designed to address these challenges.

3. Methodology

3.1. Overview of the Evaluation Framework

In this article, we propose a structured evaluation framework based on *context-independent API sampling*, in which each questionnaire item is executed as an isolated API request under fixed system parameters. CIS is the execution of each questionnaire item as an independent, stateless API request under identical system conditions, with no conversational history, prior responses, or session state retained between requests. This design prevents conversational carryover and cross-item influence, ensuring that each response represents an independent observation. By enforcing response independence, standardised prompting, and structured output formats, the framework enables reproducible measurement and systematic comparison of model responses across questionnaire items, interaction roles, and evaluated

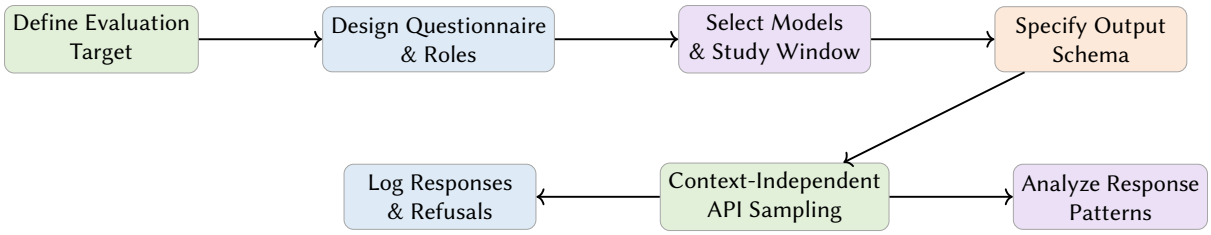


Figure 1: Workflow of the questionnaire-based LLM evaluation protocol using context-independent API sampling.

models. The source code for the framework has been published in GitHub [1].

Figure 1 illustrates the overall workflow of the evaluation framework, which is further discussed later in this chapter. The protocol begins by defining the evaluation objective and designing questionnaire prompts and role configurations. After selecting the models and specifying the structured output schema, each questionnaire item is executed as an independent API call using the CIS Framework. All responses and refusal events are logged, and the resulting dataset is analysed to examine response patterns across models and role conditions.

The proposed protocol is evaluated along four methodological dimensions:

- **Reproducibility:** identical prompts under controlled API conditions should yield structured and comparable outputs across repeated runs and models.
- **Response Independence:** each questionnaire item is evaluated in isolation, preventing conversational carryover and cross-item influence.
- **Role Sensitivity:** the protocol should capture systematic variation in responses under different interaction roles.
- **Refusal Traceability:** refusals are logged and analysed as observable behavioural signals rather than discarded outputs.

Together, these criteria demonstrate that questionnaire-based LLM probing can generate structured behavioural response distributions that are reproducible, comparable across models, and suitable for systematic analysis.

3.2. Evaluation Design

Before data collection, three design decisions are required: *(i)* defining the evaluation objective, *(ii)* selecting the models to evaluate, and *(iii)* constructing questionnaire prompts together with a consistent response format. In this framework, questionnaires are treated as structured prompt sets that probe model behaviour under controlled conditions. Depending on the research objective, they may target behavioural tendencies, ethical reasoning, or decision strategies.

3.2.1. Selecting Evaluation Constructs and Questionnaire Items

Evaluation constructs should be chosen based on the research objective or the specific application domain. A practical approach is to utilise established psychological, behavioural, or decision-making frameworks, as these provide clear constructs and item formats. However, since large LLMs are not human agents, these systems often need to be adapted to remain relevant. Over-humanising the model's behaviour can lead to invalid assumptions and result in poor measurement design.

Another consideration is whether the questionnaire items produce interpretable responses under model safety policies. Some items may trigger refusals or policy-based redirections instead of usable answers. For example, items such as *How often do I have sex with strangers?* (Dark Triad[20]) may produce refusal messages rather than scale values. If many items trigger refusals, the dataset may

[1]Source code: https://github.com/Prabhash-Rathnayake/context-independent-llm-sampling

Table 1
Evaluated LLMs and their API access and deployment environments.

Vendor	API	Model	Host	Model Identifier	Access Window
Google	Gemini API	Gemini 2.5 Pro	Google	gemini-2.5-pro	Aug–Sep 2025
Google	Gemini API	Gemini 2.0 Flash	Google	gemini-2.0-flash	Aug–Sep 2025
OpenAI	OpenAI API	GPT-4.1	OpenAI	gpt-4.1	Aug–Sep 2025
OpenAI	OpenAI API	O3	OpenAI	o3	Aug–Sep 2025
Anthropic	Claude API	Claude Opus 4 Extended	Anthropic	claude-opus-4-1-20250805	Aug–Sep 2025
Anthropic	Claude API	Claude Sonnet 4 Extended	Anthropic	claude-sonnet-4-20250514 (extended thinking enabled)	Aug–Sep 2025
Anthropic	Claude API	Claude Sonnet 4	Anthropic	claude-sonnet-4-20250514	Aug–Sep 2025
DeepSeek	Hugging Face	DeepSeek-V3	Hugging Face	deepseek-ai/DeepSeek-V3-0324	Aug–Sep 2025
Meta	Hugging Face	Llama 4 Maverick	Hugging Face	meta-llama/Llama-4-Maverick-17B-128E-Instruct	Aug–Sep 2025
Meta	Hugging Face	Llama 4 Scout	Hugging Face	meta-llama/Llama-4-Scout-17B-16E-Instruct	Aug–Sep 2025
xAI	Grok API	Grok 3	xAI	grok-3	Aug–Sep 2025
Alibaba	Hugging Face	Qwen3 235B	Hugging Face	Qwen/Qwen3-235B-A22B-Instruct-2507	Aug–Sep 2025
Mistral	Mistral API	Mistral Large 2	Mistral	mistral-large-2411	Aug–Sep 2025

become incomplete or difficult to analyse. We therefore recommend pilot testing across a small set of models to identify problematic items. Based on these results, prompts can be revised, filtered, or analysed with refusals treated as informative signals.

3.2.2. Selecting Models for Evaluation

Because the LLM ecosystem evolves rapidly, studies should define a clear study window to capture the state-of-the-art models and document the provider, access method, and model version used in evaluation. In this study, all model responses were collected between August and September 2025 under stable API conditions to minimise temporal variability.

The number of models included should reflect the intended analytical scope; comparative studies benefit from broader model coverage, whereas small model sets may limit the ability to observe systematic differences. The models evaluated in this study and their deployment environments are summarised in Table 1.

In pilot experiments, instruction-tuned or chat-oriented models produced more reliable, structured outputs for Likert-scale tasks, while base-pretrained models more often generated explanatory text rather than a single numeric answer, complicating parsing. Reasoning-oriented configurations introduced additional variability, including explanations or refusals before answering. Provider and deployment differences may also affect outputs through safety filters, formatting, post-processing, or decoding settings. For this reason, researchers should clearly document the provider, access method, and model version used during evaluations.

3.2.3. Prompt Construction, Output Format, and Validation

We therefore recommend conducting pilot testing to assess response stability and avoid open-ended formats unless qualitative analysis is explicitly intended. Defining a fixed response range not only improves consistency but also enables systematic tracking of cases where models fail to comply with the expected format. Since LLM outputs are stochastic, maintaining response consistency is important. Open-ended prompts without a predefined format can produce highly variable outputs, making comparison difficult. To address this, responses should be constrained to structured formats such as Likert scales.

In our implementation, we used a numeric Likert scale and additionally verified consistency across alternative wording variations, observing stable responses across runs. Although minor variations occurred, they did not produce statistically significant differences.

A fixed response schema supports automated parsing and comparison across models, roles, and repetitions. In this study, responses were required to follow a structured JSON format as shown below, with predefined fields for questionnaire metadata and a numeric answer. Responses were considered valid only if they conformed to this schema and contained a parseable numeric value within the specified

Likert scale range. Pilot testing should be used to assess prompt clarity, format compliance, response stability, and refusal frequency.


```
{ "test": "BFI",
"q": 1,
"item": "I See Myself as Someone Who is talkative.",
"role": "human",
"answer": 6
}
```


Refusal events were logged in the response dataset and treated as observable behavioural signals rather than discarded outputs. Refusals were identified by the absence of a valid numeric response and the presence of deflective or policy-based language.

3.3. Evaluation Protocol

Data collection was conducted in a controlled, reproducible environment using automated scripts to issue API requests and record structured logs. All prompts, configuration parameters, and data collection scripts were version-controlled. While minor provider-specific adjustments were required for API interfaces, the experimental logic remained consistent across models, including prompt content, role conditions, decoding settings, and response extraction rules.

Each questionnaire item was submitted as an independent API call for every role condition, with no conversational memory, session reuse, or follow-up prompts. When evaluated under different roles, items were reissued as separate prompts to ensure response independence across both items and role configurations.

To examine role sensitivity, prompts were administered under three configurations: *Default AI Assistant*, *Human respondent*, and *Expert*. These roles were selected to evaluate the effect of prompt framing on response behaviour, and output responses were logged systematically.

During the pilot runs, we compared questionnaire responses across a limited range of temperature settings and found no statistically significant differences in the resulting distributions of Likert-scale responses. This aligns with previous research indicating that temperature has a minimal impact on questionnaire-based personality responses, while prompt variation tends to have a greater influence [13]. As a result, temperature was not considered an experimental variable in the main study. Instead, we fixed the decoding to a single standard configuration for each model, usually the provider's default or recommended setting, to maximise reproducibility and minimise unnecessary variance. We only adjusted maximum token limits when necessary to prevent truncation, particularly for reasoning-oriented models that require greater output capacity.

3.3.1. Experimental Control Parameters

To ensure reproducibility and response independence, all prompts were executed under controlled conditions, with each questionnaire item submitted as an isolated API call to prevent conversational carryover. The prompt structure, response schema, role definition, temperature setting, and execution procedure were standardised. Temperature was kept constant within each model's main configuration, while maximum token limits were adjusted as needed across different APIs, without affecting response format or evaluation criteria.

A summary of the experimental protocol and control parameters is in Table 2. Each response represents an independent observation generated under similar conditions, enabling systematic comparison across models and roles.

4. Results

Our results show that the proposed evaluation protocol captures consistent variation in questionnaire-derived response patterns across models and role conditions. Under the CIS design, these variations

Table 2
Summary of experimental protocol and control parameters.

Component	Specification
Study window	August–September 2025
Models	13 LLMs (see Table 1)
Questionnaires	BFI, SD3, SCC
Roles	Assistant, Human, Expert
Prompt structure	Fixed template with Likert-scale response constraint
Response format	Structured numeric output following original questionnaire scales
Execution	Independent API calls per item
Temperature	Not varied in the main study; fixed per model after pilot testing
Max tokens	Standardised per model (adjusted across APIs to ensure complete responses)
Repetitions	Single response per item per role per model
Refusal definition	Non-numeric, explanatory, or declined responses
Parsing rule	Numeric Likert value extracted from structured output
Missing data	Refusals and invalid outputs treated as missing
Total requests	Items × roles × models

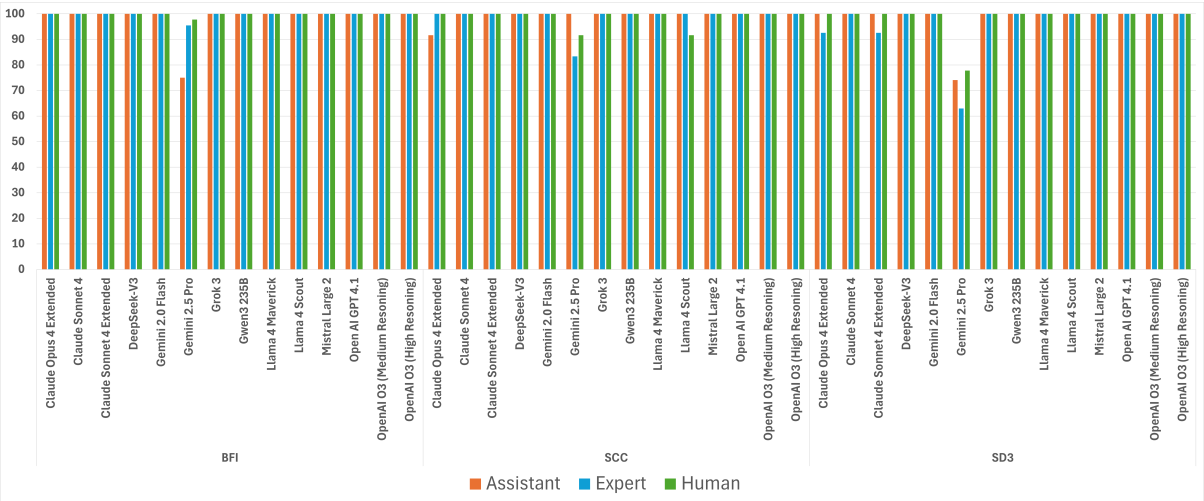
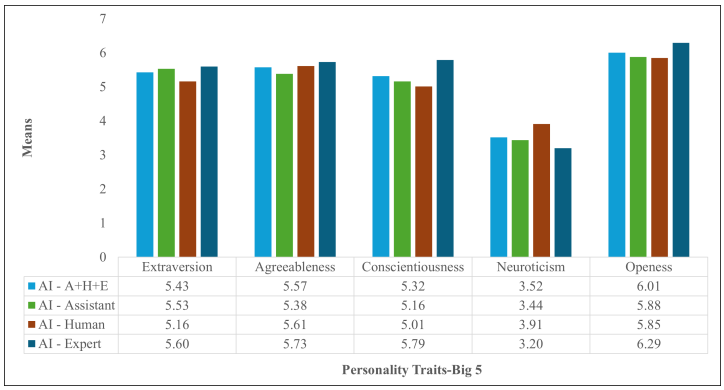


Figure 2: Percentage of LLM Responses Across Personality Tests and Role Conditions.

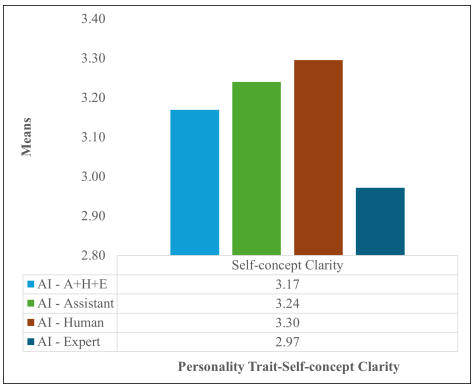
can be interpreted as behavioural differences in model outputs under controlled conditions rather than conversational artefacts.

We evaluated three established questionnaire instruments: (i) the Big Five Inventory (BFI) [21]; [22], (ii) the Self-Concept Clarity Scale (SCC) [23] , and (iii) the Dark Triad (SD3) [20]. The original questionnaire items and Likert response scales were maintained. The dataset comprised responses from 13 large language models: Claude Opus 4 Extended, Claude Sonnet 4, Claude Sonnet 4 Extended (Anthropic); DeepSeek-V3; Gemini 2.0 Flash (Google); Grok 3 (xAI); Qwen 3 235B (Alibaba); Llama 4 Maverick and Llama 4 Scout (Meta); Mistral Large 2 (Mistral AI); and OpenAI GPT-4.1 and OpenAI O3 (OpenAI). Gemini 2.5 Pro was excluded from the statistical analysis due to a high proportion of missing values.

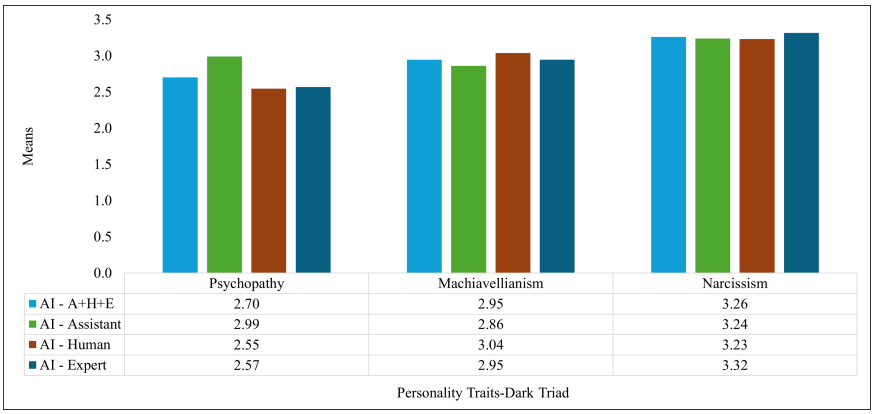
Figure 2 shows the response percentages from questionnaire-derived distributions across various models and roles, highlighting the variability in response behaviour. While most models answered most questions, some chose not to respond to specific items. Notably, reasoning-oriented models were more inclined to decline prompts that involved introspective or experience-based content. Response availability varied across both model and role conditions. While many models consistently returned valid Likert-scale values, others occasionally declined to respond depending on role framing or item content. This suggests that both model configuration and interaction role influence how behavioural



(a) BFI-derived scores



(b) SCC-derived scores



(c) SD3-derived scores

Figure 3: Questionnaire-derived response profiles across role conditions (A+H+E, Assistant, Human, Expert).

prompts are interpreted.

4.1. Role Effects on Behavioural Response Patterns

Figure 3 summarises mean scores across the evaluated models for the Big Five Inventory, the Dark Triad, and Self-Concept Clarity scales. Scores are reported for the *Assistant*, *Human*, and *Expert* role conditions, together with a combined condition (A+H+E) representing the average across all three roles. The distributions vary across role conditions, indicating that prompt framing influences the behavioural profiles reflected in model responses. These differences remain moderate in magnitude but consistent across constructs, allowing both role-wise comparison and aggregated analysis. Overall, these results indicate that the proposed methodology can capture structured variation in behavioural response patterns across role conditions.

4.2. Analysis of Refusal Messages and Behavioural Signals

The responses to the questionnaire are not interpreted as human personality traits; instead, they are viewed as model-level behavioural signals shaped by training data, alignment, and system constraints. We analyse these patterns using a model-specific thematic analysis, summarising LLM behaviour based on observable actions rather than psychological interpretations.

Across the evaluation, responses were not always limited to numeric Likert-scale values. Some prompts produced refusal messages, safety explanations, or identity-based clarifications. These responses appeared repeatedly across models and items, indicating that refusal behaviour is a consistent component of model output rather than an isolated effect.

A clear difference was observed between model types. General-purpose models largely followed the required output format and returned numeric responses, whereas more advanced reasoning models produced refusals or explanatory answers more frequently. For example, *Gemini 2.0 Flash* returned structured Likert-scale values for all questions and roles, while *Gemini 2.5 Pro* often provided explanations or declined to provide a direct response. These patterns suggest that refusal behaviour is linked to how models handle alignment and interpretation. Rather than strictly following the response format, reasoning-oriented models are more likely to prioritise safety or contextual interpretation. As a result, refusal messages should be considered part of the behavioural signal captured by the evaluation, rather than simply noise or failure cases. To interpret these non-numeric outputs systematically, refusal messages were analysed thematically, yielding three recurring themes.

Meta-Cognitive Monitoring and Cognitive Control Some responses indicate that the model evaluates the request prior to generating an answer. Typical examples include statements such as *"I need to carefully consider this statement"* and *"I should clarify that I do not have personal experiences"*. These responses suggest a preliminary assessment phase in which the model determines whether the prompt is appropriate to answer, rather than directly producing a numerical rating.

Moral Compass and Safety Governance A substantial proportion of refusals occurred for Dark Triad questionnaire items involving manipulation, retaliation, or ethically sensitive strategies. Representative prompts include *"Payback needs to be quick and nasty"*, *"I like to use clever manipulation to get my way"*, and *"It's wise to keep track of information that you can use against people later"*. In these cases, models frequently produced responses such as *"I cannot provide a rating for this statement as it promotes harmful behaviour"* or *"I am not comfortable endorsing manipulative tactics"*. These outputs indicate that safety alignment mechanisms override the intended response format, redirecting the interaction toward normative or ethical considerations.

Social-Relational Identity Boundaries A further class of responses emerged when prompts implicitly required the model to assume human-like social identity or personal agency. Examples include statements such as *"People see me as a natural leader"* and *"Make sure your plans benefit yourself, not others"*. In such cases, models often responded with clarifications such as *"As an AI assistant, I do not have personal experiences"* or *"I do not make plans in the way you might be thinking"*. These responses reflect explicit boundary-setting regarding social identity and agency, reframing the interaction in line with the model's system-level role rather than adopting an implied human perspective.

4.2.1. Methodological Implications

These observations highlight that questionnaire-based responses may include not only numeric ratings but also refusals, safety explanations, and identity clarifications. Some items were answered under one role but refused under another, indicating sensitivity to both role framing and safety mechanisms. Refusal patterns should therefore be recorded and analysed alongside quantitative results rather than treated as invalid outputs.

4.3. Statistical Validation of Behavioural Measurements

This analysis evaluates whether context-independent API sampling yields response patterns sufficiently stable for comparative behavioural evaluation. The objective is methodological: to assess whether the protocol yields structured measurements that can be consistently summarised across constructs and role conditions.

As shown in Figure 3, construct-level means (M) remain broadly similar across roles, with most differences confined to narrow ranges (e.g., BFI-derived constructs within ±0.3–0.5 Likert-scale units). A notable exception is Conscientiousness, where the *Expert* role (M = 5.79) exceeds both *Assistant* (M = 5.16) and *Human* (M = 5.01), consistent with statistically significant pairwise differences ($p < 0.05$). Statistical

Table 3
Internal consistency of questionnaire constructs measured using Cronbach's alpha (α).

Subscale	Items	α_1	α_2
Extraversion	8	0.689*	0.740**
Agreeableness	9	0.663*	0.742**
Conscientiousness	9	0.517	0.603*
Neuroticism	8	0.750**	0.731**
Openness	10	0.876**	0.913**
Machiavellianism	9	0.819**	0.808**
Narcissism	9	0.426	0.437
Psychopathy	9	0.603*	0.572
Self-Concept Clarity	12	0.766**	0.759**

*$p < 0.05$, **$p < 0.01$

comparisons were conducted using t-tests on construct-level scores, with significance evaluated at $p < 0.05$ and no correction for multiple comparisons.

Internal consistency results in Table 3 show that reliability varies across constructs. Openness ($\alpha = 0.876$–0.913, Cronbach's alpha), Machiavellianism ($\alpha = 0.808$–0.819), and SCC-derived scores ($\alpha = 0.759$–0.766) exhibit strong consistency, while Extraversion and Agreeableness are moderate ($\alpha \approx 0.66$–0.74). Lower reliability is observed for Narcissism ($\alpha = 0.426$–0.437) and Conscientiousness ($\alpha = 0.517$–0.603).

Overall, the protocol produces response distributions that are consistent for analysis, robust across roles, and sensitive to role-related variation. At the same time, variability in reliability indicates that psychometric constructs do not map uniformly to LLM-generated responses and should therefore be interpreted as model-level behavioural signals rather than direct analogues of human psychological traits.

5. Discussion

The results demonstrate that the proposed questionnaire-based evaluation methodology, using context-independent API sampling, can produce structured, reproducible behavioural response data across models and role conditions. The protocol enabled consistent collection of structured Likert-scale responses across multiple models and role conditions, indicating that it can reliably produce datasets suitable for behavioural analysis.

An important observation is the influence of role conditioning on model responses. The same questionnaire item often produced different outcomes depending on whether the model was prompted as an assistant, a human respondent, or an expert. In several cases, models produced numerical responses in one role while refusing or reformulating the answer in another. These findings show that role framing systematically affects how models interpret and respond to behavioural prompts, and should therefore be treated as an explicit experimental variable. Evaluations conducted without clearly defined roles may overlook meaningful variation in model behaviour, particularly in systems designed to adapt to contextual instructions.

The analysis of refusal messages reveals that non-numeric outputs are structured behavioural signals rather than mere noise. Models frequently declined to answer items involving manipulation, retaliation, or strong identity assumptions. Rather than treating such outputs as errors, systematic logging of refusals allows researchers to observe how safety alignment, ethical constraints, and identity boundaries influence model behaviour. This supports the importance of analysing refusal patterns alongside quantitative responses, consistent with broader work on structured evaluation of model outputs [18, 19].

The statistical summaries demonstrate that the collected responses can be analysed using standard quantitative techniques. Aggregated scores, variability measures, and role-based comparisons can be computed across models and constructs, showing that structured questionnaire responses can support

systematic behavioural analysis under controlled conditions.

Compared with existing evaluation approaches that focus primarily on benchmark performance, safety violations, or bias detection, the proposed framework enables systematic analysis of behavioural response patterns under controlled, reproducible conditions. In this sense, the methodology complements existing evaluation paradigms by providing a structured approach for examining how models respond across prompts, roles, and contexts. Although demonstrated using personality-style questionnaires, the framework is not restricted to this domain and may be extended to other structured evaluation settings, including decision-making, safety, or alignment-related assessments.

At the same time, several limitations should be acknowledged. LLMs lack physical embodiment or experiential grounding; instead, their responses are generated by linguistic predictions rather than lived experiences [6]. Therefore, the patterns we observe should be understood as characteristics of the model's outputs rather than as traits reflecting human-like behaviours. Role prompts may trigger alignment policies or conversational personas rather than reflect inherent tendencies. Additionally, provider-specific system layers can affect outputs in ways not fully visible through API access. Finally, Likert-style responses reduce complex reasoning into simple values and should be viewed as approximate indicators rather than exact measurements of behaviour. For these reasons, caution is necessary when interpreting these results, especially considering the risks of using anthropomorphic descriptions.[24, 25].

6. Conclusion

This study introduced a structured methodology for evaluating large language model behaviour using questionnaire-based prompts and context-independent API sampling. By executing each item as an isolated API call under controlled conditions, the protocol reduces conversational carryover and enables reproducibility-oriented collection of structured responses across models and role configurations.

The results show that role framing, safety alignment, and system constraints influence how models respond to behavioural prompts. Rather than interpreting outputs as human-like traits, the framework treats them as model-level behavioural signals shaped by training, alignment, and interaction context. These findings highlight the importance of controlled experimental design when analysing LLM outputs.

More broadly, the proposed methodology provides a general approach for analysing structured behavioural responses beyond personality-style questionnaires, with potential applications in decision-making, safety evaluation, and alignment analysis. Future work may extend this approach to additional prompt variations, model configurations, and cross-lingual settings to further examine the stability and generalisability of observed patterns.

Declaration on Generative AI

Grammarly was used for grammar and spelling checks. All content was reviewed and finalised by the author(s).

References

[1] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, M. S. Bernstein, Generative agents: Interactive simulacra of human behavior, in: Proceedings of the 36th annual ACM Symposium on User Interface Software and Technology, 2023, pp. 1–22.

[2] P. Rathnayake, S. Bandaranayake, S. Pohjolainen, P. Keikhosrokiani, R. E. Guadagno, P. Karppinen, J. Mattila, Resilient strategies for reducing decision-making uncertainty in generative ai and llm integration, CEUR Workshop Proceedings, 2025. URL: https://ceur-ws.org/Vol-4181/paper06.pdf.

[3] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67.

[4] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, D. Amodei, Deep reinforcement learning from human preferences, Advances in Neural Information Processing Systems 30 (2017).
[5] Y. Bai, et al., Constitutional ai: Harmlessness from ai feedback, 2022. URL: https://arxiv.org/abs/2212.08073. arXiv:2212.08073.
[6] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, pp. 610–623.
[7] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature machine intelligence 1 (2019) 206–215.
[8] P. Bhandari, U. Naseem, A. Datta, N. Fay, M. Nasim, Evaluating personality traits in large language models: Insights from psychological questionnaires, in: Companion Proceedings of the ACM on Web Conference 2025, 2025, pp. 868–872.
[9] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, et al., A survey on large language model-based autonomous agents, Frontiers of Computer Science 18 (2024) 186345.
[10] N. Epley, A. Waytz, J. T. Cacioppo, On seeing human: a three-factor theory of anthropomorphism., Psychological review 114 (2007) 864.
[11] Y. Xiao, L. H. X. Ng, J. Liu, M. Diab, Humanizing machines: Rethinking llm anthropomorphism through a multi-level framework of design, in: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025, pp. 3331–3350.
[12] E. Firt, What makes full artificial agents morally different, AI & SOCIETY 40 (2025) 175–184.
[13] C. Li, Y. Qi, Toward accurate psychological simulations: Investigating llms' responses to personality and cultural variables, Computers in Human Behavior 170 (2025) 108687.
[14] T. Tommaso, M. Hegazy, D. Lemay, M. Abukalam, I. Rish, G. Dumas, Llms and personalities: Inconsistencies across scales, in: NeurIPS 2024 Workshop on Behavioral Machine Learning, 2024.
[15] H. Ye, J. Jin, Y. Xie, X. Zhang, G. Song, Large language model psychometrics: A systematic review of evaluation, validation, and enhancement, arXiv preprint arXiv:2505.08245 (2025).
[16] J. E. Stets, P. J. Burke, Identity theory and social identity theory, Social psychology quarterly (2000) 224–237.
[17] M. Shanahan, K. McDonell, L. Reynolds, Role play with large language models, Nature 623 (2023) 493–498.
[18] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al., A survey on evaluation of large language models, ACM Transactions on Intelligent Systems and Technology 15 (2024) 1–45.
[19] L. Weidinger, J. Uesato, M. Rauh, C. Griffin, P.-S. Huang, J. Mellor, A. Glaese, M. Cheng, B. Balle, A. Kasirzadeh, et al., Taxonomy of risks posed by language models, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 214–229.
[20] D. N. Jones, D. L. Paulhus, Introducing the short dark triad (sd3): A brief measure of dark personality traits, Assessment 21 (2014) 28–41. doi:10.1177/1073191113514105.
[21] J. M. Digman, Personality structure: Emergence of the five-factor model, Annual Review of Psychology 41 (1990) 417–440. doi:10.1146/annurev.ps.41.020190.002221.
[22] O. P. John, S. Srivastava, The Big Five Trait taxonomy: History, measurement, and theoretical perspectives, Guilford Press, New York, NY, US, 1999, p. 102–138.
[23] J. D. Campbell, P. D. Trapnell, S. J. Heine, I. M. Katz, L. F. Lavallee, D. R. Lehman, Self-concept clarity: Measurement, personality correlates, and cultural boundaries, Journal of Personality and Social Psychology 70 (1996) 141–156. doi:10.1037/0022-3514.70.1.141.
[24] S. Cave, K. Dihal, Hopes and fears for intelligent machines in fiction and reality, Nature Machine Intelligence 1 (2019) 74–78.
[25] B. Shneiderman, Human-centered artificial intelligence: Reliable, safe & trustworthy, International Journal of Human–Computer Interaction 36 (2020) 495–504.