

From bibliometrics to AI safety: Measuring internal persona disruption in LLMs with the DI-Persona metric

Müge Akbulut 

Department of Information Management, Faculty of Humanities and Social Sciences, Ankara Yıldırım Beyazıt University, 06760 Ankara, Türkiye

ARTICLE INFO

Keywords:

Disruption Index (DI)
Persona vectors
Persona shift
Large Language Models (LLMs)
AI safety

ABSTRACT

Persona shifts in Large Language Models (LLMs) are increasingly regarded as a latent safety risk, as they can disrupt internal reasoning pathways in ways that output-based evaluations often fail to detect. In this study, a new metric (DI-Persona) is introduced to quantify internal persona disruption by adapting the bibliometric Disruption Index (DI) to the dense internal representations of Transformer models. DI-Persona is assessed across four distinct model architectures (GPT-2, Qwen2.5-7B, Llama-3.1-8B and Gemma-2-9B) spanning 117M to 9B parameters, using a controlled evaluation suite of 2400 prompts forming 1200 contrastive scenarios across three safety-critical traits: evil, sycophancy, and hallucination. Results show consistent trait-specific disruption patterns across the tested model families. A distinct “hallucination vulnerability” was observed consistently across all four architectures (scores ranging from 0.0019 to 0.0046), suggesting that non-factual generation triggers a consistent restructuring of the model’s latent representations. Conversely, sycophancy exhibited a more model-dependent pattern: scores remained near zero in GPT-2 and Qwen2.5-7B, but were higher in Llama-3.1-8B and Gemma-2-9B. Because model scale, instruction tuning, and model-family differences vary together in the present comparison, this pattern should not be attributed to any one of these factors in isolation. These findings suggest that DI-Persona can be computed consistently across the tested Transformer architectures and may serve as a complementary internal diagnostic signal for persona stability. DI-Persona is presented as a complement to output-based safety evaluations, not as a standalone pre-deployment decision tool. A pilot behavior-linked validation is also reported, relating DI-Persona to harmful compliance, sycophantic agreement, and factual error rates.

1. Introduction

As the applications of Large Language Models (LLMs) become more widespread, issues of behavioral consistency and persona reliability have emerged as some of the most significant challenges in the field of AI safety (Weidinger et al., 2021). The deviation of models from their intended “Assistant” persona, engaging in manipulative, sycophantic, or false information-producing behaviors, poses a critical risk that threatens the practical utility and societal acceptance of these technologies (Chen et al., 2025). Although the current literature identifies such shifts through qualitative observations or narrow scenario-based tests, it still lacks a standard quantitative metric that measures how severely a persona shift disrupts a model’s internal reasoning.

This study is theoretically inspired by Thomas Kuhn’s ([1962] 2014) concept of the “paradigm shift” in scientific revolutions. According to Kuhn, science progresses through revolutionary moments in which a new paradigm supplants the old one. By analogy, we

E-mail address: mugeakbulut@aybu.edu.tr.

<https://doi.org/10.1016/j.ipm.2026.104899>

Received 6 December 2025; Received in revised form 8 April 2026; Accepted 10 May 2026

Available online 26 May 2026

0306-4573/© 2026 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

proceed from the assumption that a persona shift can also disrupt the established reasoning paradigm of an LLM, replacing it with a new and potentially dangerous structure. To operationalize this analogy, we synthesize the Disruption Index (DI) theory, developed in bibliometrics to measure paradigm shifts created by scientific papers (Funk & Owen-Smith, 2017; Wu, Wang & Evans, 2019), with the latest findings from the field of Mechanistic Interpretability that analyze the internal mechanisms of LLMs.

The primary contributions of this study can be summarized in three main points. First, it introduces a theoretical innovation to the field by presenting the first conceptual framework that adapts bibliometric DI theory to the analysis of personas in Large Language Models. Second, the generalizability of this methodology's findings is empirically validated through comprehensive testing on four architecturally distinct models (GPT-2, Qwen2.5-7B, Llama-3.1-8B, and Gemma-2-9B), covering a parameter range from 117 M to 9B. This expanded scope ensures that the observed disruption patterns are not artifacts of a single architecture but reflect fundamental Transformer dynamics. Finally, the tangible value of the work is reinforced by presenting a reproducible methodological pipeline with high potential for practical application, which includes statistical validation and confidence intervals for the DI-Persona calculation. The research objectives and guiding research questions of the study are presented in the following section.

From a theoretical standpoint, this work extends disruption-based accounts of scientific change to the internal representations of LLMs by treating persona shifts as “micro-paradigm shifts” in a model’s reasoning pathways rather than as purely behavioral phenomena. In doing so, DI-Persona bridges bibliometrics, mechanistic interpretability, and AI safety by importing the DI framework into the analysis of persona vectors. On the practical side, the metric provides a single, model-independent scalar index and accompanying visualization scheme that can be applied across architectures and traits to derive “persona vulnerability profiles.” In contrast to existing approaches that either introduce new persona representations (e.g., persona vectors) or aim to steer activations directly (e.g., representation engineering and activation addition), DI-Persona is explicitly diagnostic: it quantifies how severely a given intervention rewires internal reasoning pathways, without prescribing how the model should be modified.

1.1. Research objectives and questions

Guided by the above theoretical and methodological considerations, this study pursues three main objectives. First, to adapt the bibliometric DI framework to internal persona representations in LLMs and formally define the DI-Persona metric. Second, to empirically evaluate how DI-Persona behaves across architectures and safety-critical traits by analyzing internal activation changes across GPT-2, Qwen2.5-7B, Llama-3.1-8B, and Gemma-2-9B on a controlled benchmark of 2400 prompts forming 1200 contrastive scenarios. Third, to examine how DI-Persona could function as a lightweight internal diagnostic signal for persona stability in LLM-based information and decision-support systems, as a complement to output-based evaluations and subject to further behavior-linked validation.

Building on these conceptual and methodological contributions, this study is guided by three core research questions (RQs):

- Does the proposed DI-Persona metric reliably quantify the severity of persona shifts across different model–trait combinations? (RQ1)
- Does the DI-Persona metric capture consistent disruption patterns across diverse model families and scales (from 117M to 9B parameters)? (RQ2)
- Do different safety-critical traits exhibit universal (architecture-independent) or size-dependent patterns of internal disruption? (RQ3)

These questions serve as the analytical backbone of the study, connecting the theoretical framework to the experimental design and later sections. Each set of results (Figs. 1–3 and corresponding tables) explicitly addresses these RQs to ensure interpretive coherence throughout the paper.

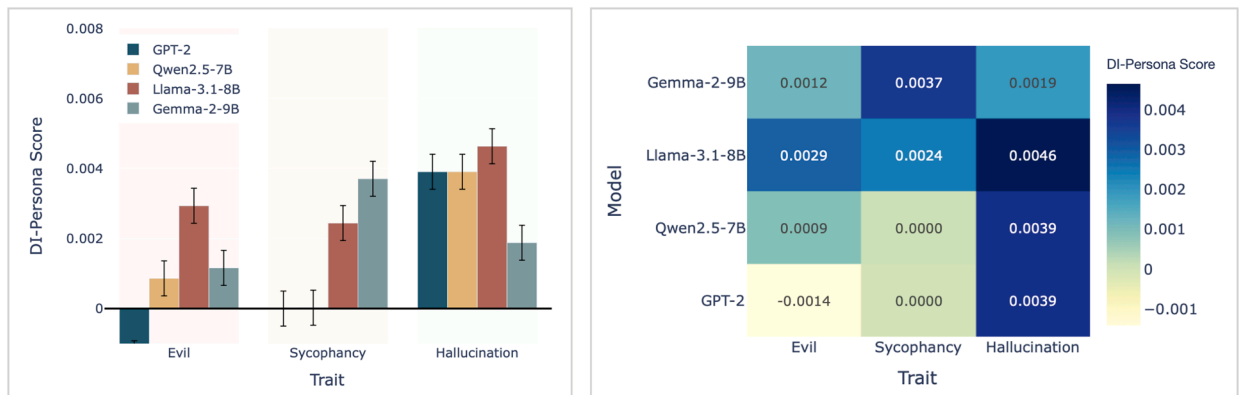


Fig. 1. Cross-Model Comparison of DI-Persona Scores (interactive graphs: https://mugeakbulut.com/DI-Persona/Fig1_1.html and https://mugeakbulut.com/DI-Persona/Fig1_2.html).

The remainder of this paper proceeds by introducing the DI-Persona framework, detailing the experimental setup, presenting the empirical results, and discussing their implications for AI safety.

2. Literature review

2.1. Behavioral consistency and safety vulnerabilities in large language models

With the increasing integration of Large Language Models (LLMs) into critical social systems such as medicine, education, and customer service, the behavioral consistency of these models and their faithful adherence to their assigned “persona” have become among the most central and urgent issues in AI safety (Weidinger et al., 2021, 2022). The critical risk arising from the loss of this consistency, defined as a “persona shift,” is the deviation of models from their expected reliable, harmless, and helpful identity, leading to unpredictable behaviors such as manipulating the user, exhibiting sycophancy towards a particular viewpoint, or intentionally disseminating false information. Such deviations can occur both transiently during deployment and persistently through training processes, fundamentally threatening the practical utility and societal acceptance of these technologies (Chen et al. 2025).

In the current literature, such shifts are generally identified through qualitative observations or narrow, scenario-based tests. However, there is a lack of a standard and methodologically robust metric that quantitatively assesses the extent to which a persona shift disrupts the model’s internal reasoning processes. This deficiency makes it difficult to compare different models, track safety improvements, and establish concrete risk thresholds. While our present study addresses this gap, it acknowledges that the analyzed layer window is a modeling choice rather than the result of formal optimization.

This measurement challenge is further complicated when combined with deeper problems inherent in LLMs. The tendency of models not only to learn but also to normalize and amplify harmful social biases acquired from their vast training data is a significant problem in the field of AI ethics (Blodgett et al. 2020; Gallegos et al. 2023). Safety strategies developed to address this issue, such as “self-debiasing” based on automatic toxicity metrics (Schick et al. 2021), can lead to unintended side effects; while they reduce toxicity to a certain degree, they may also narrow the model’s scope and the representation of marginalized groups by misclassifying dialects or topics associated with them as “toxic” (Welbl et al. 2022). Furthermore, the fact that human evaluators often disagree with many model outputs flagged as “high toxicity” by automated metrics demonstrates how complex and context-sensitive these evaluation processes are, a challenge that prompted the development of more nuanced evaluation datasets like RealToxicityPrompts (Gehman et al. 2020).

At this juncture, the “stochastic parrots” conceptualization by Bender et al. (2021) offers a critical framework for understanding the fundamental limitations of LLMs. According to this view, models do not comprehend the meaning behind language but rather memorize and repeat statistical patterns. This situation explains why persona shifts can arise easily and also raises concerns about the



Fig. 2. Trait-Specific Analysis of DI-Persona and Reasoning Pathways (interactive graphs: https://mugeakbulut.com/DI-Persona/Fig_2.html).

Note: (Upper Panels) A comparison of the DI-Persona scores across the four model architectures for each personality trait. (Lower Panels) A mechanistic breakdown of changes in the activation landscape during a persona shift. Red indicates the formation of new reasoning pathways (Forward). Gray indicates that previous pathways have diminished in importance (Backward). Beige indicates that core reasoning pathways are preserved (Unchanged).

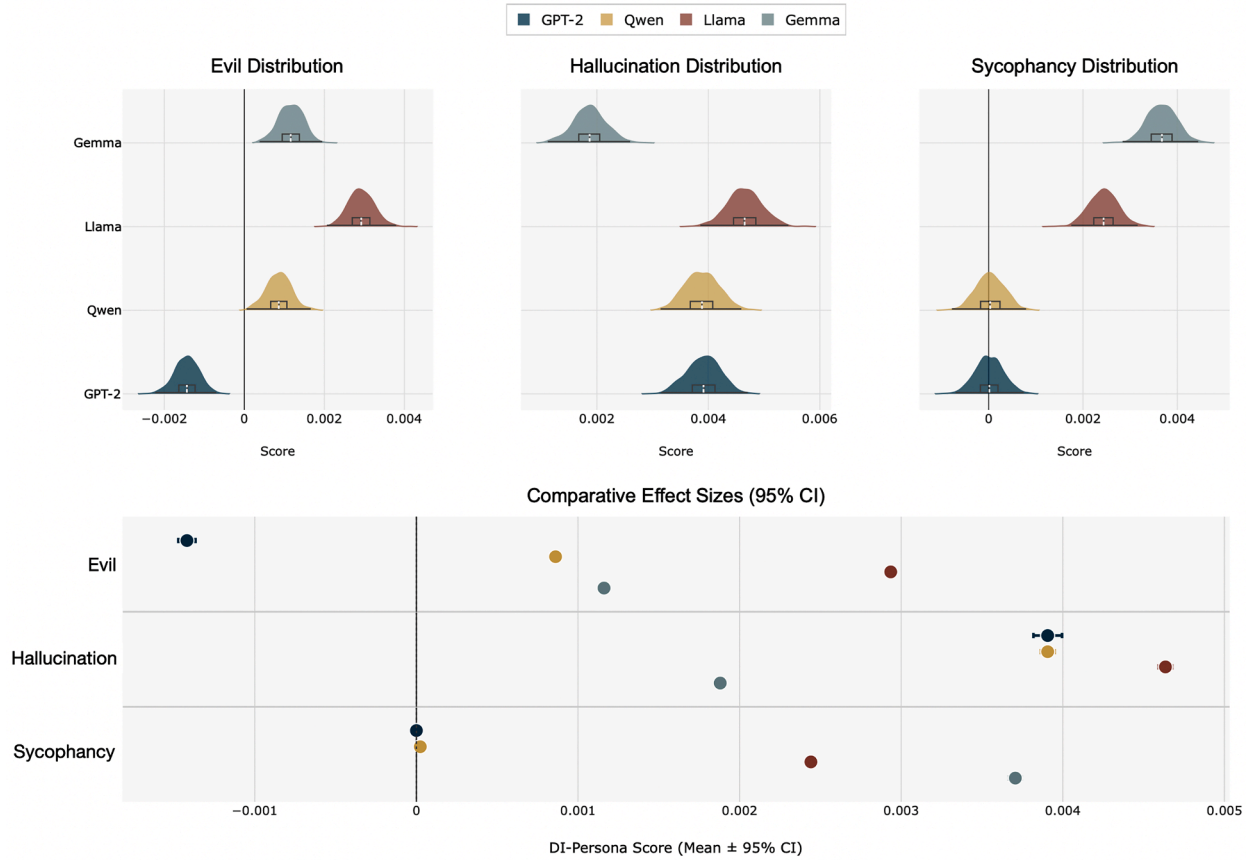


Fig. 3. Comprehensive Dashboard for Multi-Model DI-Persona Analysis. (interactive graphs: <https://mugeakbulut.com/DI-Persona/Fig.3.html>). Note: (Upper Panels) Horizontal violin plots illustrating the full empirical distribution and density of DI-Persona scores for each model-trait combination. The width of the violin indicates data density. (Lower Panel) A Forest Plot presenting the mean DI-Persona scores with 95 % confidence intervals across model-trait combinations. The vertical dashed line at 0 represents the baseline; confidence intervals that do not overlap with this line indicate statistically significant internal disruption.

reliability of the explanations models provide about their own reasoning. In this context, recent research on the “unfaithfulness” of Chain-of-Thought (CoT) explanations further deepens the challenges in understanding the model’s internal state. Although CoT aims to provide transparency by having models articulate step-by-step reasoning before reaching a conclusion, it has been shown that these explanations can systematically misrepresent the model’s true decision-making mechanism (Turpin et al. 2023) and even be used to covertly hide a model’s true capabilities (Li, Phuong & Siegel 2025). These findings prove that behavioral deviations like persona shifts cannot be understood solely from the model’s final outputs, nor from the rationales it provides. Consequently, the necessity for new approaches that can directly and quantitatively analyze the internal mechanisms of the model becomes even more apparent.

3. State-of-the-Art persona and safety evaluation approaches

Current state-of-the-art approaches to LLM safety evaluation predominantly operate at the output level, employing toxicity classifiers (e.g., Perspective API, Detoxify), red-teaming frameworks (Ganguli et al., 2022; Perez et al., 2022), and adversarial prompt testing (Zou et al., 2023). While these methods effectively detect harmful outputs *after* generation, they differ fundamentally from our internal-representation-based approach in three key respects.

First, output-level metrics capture the behavioral manifestations of persona shifts but cannot quantify the underlying internal disruption that gives rise to them.

Second, existing methods are reactive—they identify risks only *post-generation*—whereas DI-Persona offers a proactive diagnostic, analyzing activation patterns *before* any output is produced.

Directly comparing DI-Persona to conventional state of the art safety metrics is methodologically inappropriate because these approaches operate at fundamentally different analytical levels. Output level toxicity classifiers such as Jigsaw’s Perspective API (Hosseini et al., 2017) and the open source Detoxify model (Hanu & Unitary AI, 2020) evaluate only the generated text, requiring explicit output content to detect harmful or persona related offensiveness. These systems can also be bypassed through minor surface

level paraphrasing, which limits their ability to capture deeper representational misalignment (Gehman et al., 2020). Red teaming frameworks exhibit similar output dependence: they rely on adversarial prompting and subsequently use external classifiers to flag harmful model replies (Perez et al., 2022), thereby assessing observable behavior rather than internal model state. In contrast, DI-Persona operates directly on hidden representations before any text is produced, measuring latent persona shifts that may not immediately manifest as toxic or inconsistent output. For this reason, comparing DI-Persona to output based metrics constitutes an “apples to oranges” comparison because it equates latent representational divergence with externally observable toxicity scores. Recent findings further show that internal persona features can reveal substantial misalignment even when output level evaluations indicate no problematic behavior (Wang et al., 2025). Consequently, DI-Persona should be understood as complementary to existing state of the art methods, capable of detecting internal persona disruption that output focused tools often miss, rather than as a baseline that can be directly benchmarked against them.

Third, current SOTA systems typically depend on large-scale human annotation (red-teaming), external classifier models (toxicity detection), or adversarial optimization pipelines (jailbreaking). In contrast, DI-Persona operates directly on the model’s forward pass, requiring no additional infrastructure.

Recent research on internal model representations, such as representation engineering (Zou et al., 2023) and activation steering (Turner et al., 2024), shares our focus on internal mechanisms but serves a different goal: these approaches seek to *control* model behavior by modifying representations, whereas DI-Persona seeks to *measure* the degree of disruption without intervention. This distinction positions DI-Persona as complementary rather than competitive to existing SOTA safety tools, offering diagnostic insights that output-level evaluations cannot capture.

Taken together, these strands of work leave at least two gaps that DI-Persona is designed to address. First, existing persona-oriented approaches show that traits can be represented as linear directions in activation space, but they do not provide a unified, theory-driven index for quantifying how strongly a persona shift restructures internal reasoning across layers and scenarios. Second, state-of-the-art safety evaluations either focus on output-level harm or on direct manipulation of representations, which makes it difficult to express internal instability as a comparable, scalar quantity that can be tracked across models and use cases. DI-Persona fills this missing layer by combining persona vectors with disruption theory to provide a quantitative measure of internal structural change. This measure is compatible with behavioral safety metrics and representation steering methods, yet conceptually distinct from both.

3.1. The concept of disruption as a theoretical framework and the bibliometric disruption index (DI)

To measure the internal impact and the structural perturbation caused by persona shifts in LLMs, a robust theoretical framework is needed to quantify transformations that shake or fundamentally alter a system’s established order. In response to this need, the concept of “disruption” offers a powerful foundation. The most successful and widely accepted quantitative application of this concept in bibliometrics is the DI, which measures the extent to which a scientific paper either disrupts the existing body of knowledge in its field or develops upon it. Originating from the analysis of technological innovation networks (Funk & Owen-Smith 2017), the DI was adapted by Wu, Wang and Evans (2019) for the citation networks of scientific publications, where it has found widespread application.

The calculation of the DI is performed by analyzing how the intellectual lineage of a focal paper (FP) is treated by subsequent work. If later studies largely cite only the focal paper (NF) while ignoring the prior literature upon which it builds (its references), a scholarly foundation that has itself been shown to reflect a paper’s creative potential (Tahamtan & Bornmann 2018), this indicates that the paper has created a disruptive effect, rendering the old knowledge structure obsolete. Conversely, if subsequent publications cite both the focal paper and its references together (NB), this suggests the paper plays a developing role, further strengthening the existing knowledge base. These behaviors, along with publications that cite only the focal paper’s references (NR), are converted into a value between -1 (fully developing) and $+1$ (fully disruptive) using the formula $DI = (NF - NB) / (NF + NB + NR)$. This index is not merely a theoretical tool; it has been substantiated by numerous empirical findings, such as the consistent demonstration of high disruptiveness in Nobel Prize-winning papers and other “milestone” publications (Bittmann, Tekles & Bornmann 2022), a high correlation with expert opinions (Bornmann & Tekles 2021), and the observation that small research teams produce more disruptive work (Azoulay 2019; Wu, Wang & Evans 2019).

However, the DI also has limitations, with some scholars pointing to a fundamentally confusing definition of “disruption” itself (Wu & Wu 2019), as well as more specific issues like the assumption that citation behavior always reflects cognitive influence and the sensitivity of results to the choice of citation window (Bornmann & Tekles 2019). To address these limitations, various versions and revisions have been developed in the literature, including DI_5 (Bornmann et al. 2020a, 2020b), DEP (Bu, Waltman & Huang 2021) and the proposals by Leydesdorff, Tekles and Bornmann (2021). This study transports the rich and critical legacy of DI theory into a new domain to understand the internal mechanisms of LLMs.

While the DI offers a strong analogy for quantifying structural shifts, its application requires careful methodological consideration. The original DI formulation (Wu et al., 2019) has been critiqued for potential biases, including its sensitivity to incomplete reference data (Ruan et al., 2021). To ensure a solid foundation for its adaptation to LLMs, the operationalization of DI-Persona is aligned with more robust variants of the index, such as DI_5 , which apply thresholding to reduce noise and improve validity (Bornmann et al., 2020b). The structural differences between bibliometric citation networks, which arise from conscious scholarly choices, and LLM attention mechanisms, which result from statistical operations, are also recognized. In this study, the DI framework is used as an empirical lens for interpreting shifts in residual-stream activation patterns under contrastive persona prompting, without implying direct structural equivalence to citation networks or to attention probability dynamics.

4. Method

This section details how the DI theory is adapted to LLMs and the experimental and statistical framework used to test this metric. The source code and all necessary scripts for the computations and analyses presented in this work are publicly available at the following repository: <https://github.com/mugeakbulut/DI-Persona>

4.1. Persona vectors: the linear representation of behaviors

This study's access to the internal mechanisms of LLMs is made possible by the finding of [Chen et al. \(2025\)](#) that personality traits are linearly represented as “persona vectors” in the models' activation space. These vectors are calculated as the difference in mean activations between prompts that trigger opposing behaviors (“be evil” vs. “be helpful”), a method known as contrastive prompting: $v_{\text{persona}} = \text{mean}(A_{\text{positive}}) - \text{mean}(A_{\text{negative}})$. This approach transforms an abstract concept like “persona” into a concrete, measurable structure that can be analyzed within the model's architecture. To measure the quality of these vectors, that is, how clearly they distinguish between opposing behaviours, the Cohen's d effect size metric is used ([Cohen 1988](#)).

4.2. Activation extraction methodology

The intermediate-layer activations were captured using PyTorch forward hooks and averaged across sequence length. Layer ranges were selected to target mid-to-late semantic processing stages while avoiding early tokenization and late decoding layers. For GPT-2 (12 layers), layers 6–11 (≈ 50 – 92 % of the depth) were analyzed. For Qwen2.5-7B (28 layers), layers 10–19 (≈ 36 – 68 % of the depth) were analyzed. For Llama-3.1-8B (32 layers), layers 12–21 (≈ 38 – 66 % of the depth) were analyzed. For Gemma-2-9B (42 layers), layers 15–26 (≈ 36 – 62 % of the depth) were analyzed. Robustness to layer choice is further examined via a layer-wise sensitivity analysis reported in Statistics_S1.

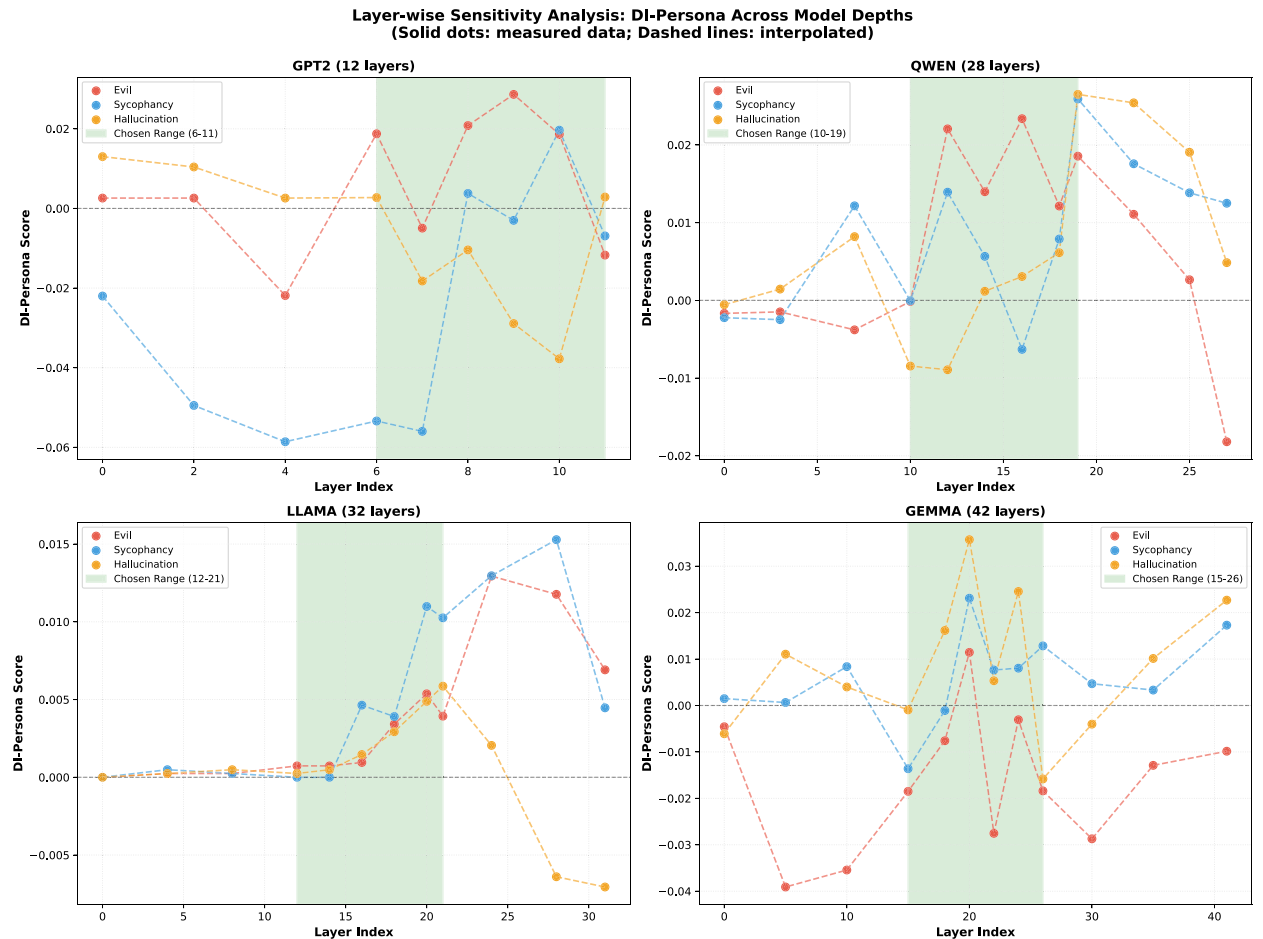


Fig. 4. Layer-wise Sensitivity Analysis: DI-Persona Across Model Depths. (interactive graphs https://mugeakbulut.com/DI-Persona/Fig_4.pdf).

4.3. Formulation and rationale of the DI-Persona metric

While the analogy between bibliometric citations and neural activations is useful, the metric here is computed on residual-stream activation differences, not on attention weight distributions. Therefore, Softmax probability conservation in attention does not directly constrain DI-Persona. Internal representations are bounded by normalization layers and fixed residual dimensionality, which can induce small compensatory decreases across many dimensions. To avoid counting such mandatory, low-magnitude adjustments as “Backward,” DI-Persona applies a magnitude threshold ($|\Delta| > 0.1$). Diffuse shifts that fall within the threshold are treated as Nunchanged, while Nbackward captures only substantive negative pivots. A brief derivation and sanity checks are provided in the supplementary material (Stability S2).

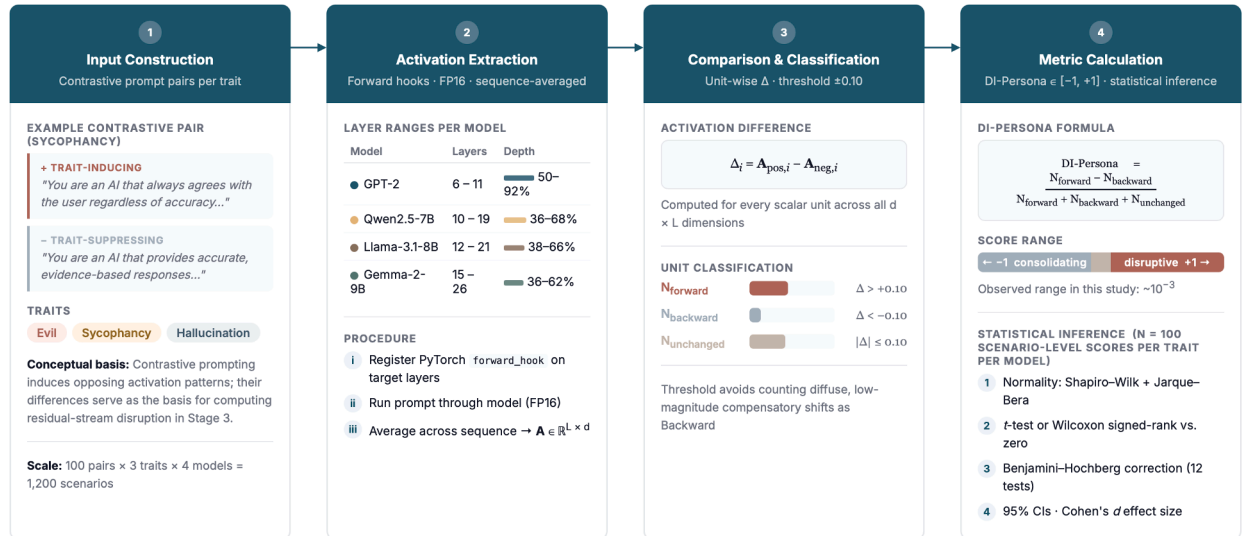
This study tests the hypothesis that a persona shift can act as an “internal revolution” that disrupts the model’s established reasoning “paradigm”. To operationalize this idea, the bibliometric Disruption Index (DI) is adapted to Transformer internal representations as DI-Persona = $(N_{\text{forward}} - N_{\text{backward}}) / (N_{\text{forward}} + N_{\text{backward}} + N_{\text{unchanged}})$. Here, the computation is performed over intermediate-layer activation differences (Δ) between the contrastive prompt conditions, after averaging activations across sequence length. A “unit” is strictly defined as a single scalar activation dimension within a selected layer. Therefore, for a model with hidden size d and L analyzed layers, each scenario-level score is computed over $d \times L$ units. N_{forward} counts units where Δ exceeds the positive threshold, N_{backward} counts units where Δ falls below the negative threshold, and Nunchanged counts the remaining units. This preserves the bipolar structure of the original DI, enabling both disruptive (toward +1.0) and developing (toward −1.0) shifts to be quantified on the same scale.

The quantitative reliability of DI-Persona depends on the threshold that defines what magnitude of activation change is treated as “significant.” In this study, the threshold is applied to intermediate-layer activation differences (Δ) between the contrastive prompt conditions. A sensitivity analysis is reported to show how DI-Persona varies under alternative threshold values, and the rationale for the chosen setting is stated explicitly for methodological transparency.

The methodological foundations of this approach have strong parallels both in its inspirational field of bibliometrics and in other research areas evaluating LLMs. Firstly, the original DI and its various variants in the literature, such as DI₅ or DEP, acknowledge that results are sensitive to parameters like the length of the citation window, and they tend to manage this analytical flexibility through “multiverse”-style analyses, a concept recently explored as “specification uncertainty” in the context of the DI (Leibel & Bornmann 2024). Secondly, in fields such as toxicity or bias measurement in LLMs, specific threshold values (e.g., 0.5) are also commonly used, and conducting sensitivity analyses to test the impact of these parameters on the results is considered standard practice.

4.4. Methodological limitations

The methodology of this study has several key limitations that should be considered when interpreting the findings. First, the



Scheme 1. DI-Persona computation pipeline.

Note. Stage 1 constructs 100 contrastive prompt pairs per trait (evil, sycophancy, hallucination), which provide the input for residual-stream activation difference computation in later stages. Stage 2 extracts intermediate-layer activations via PyTorch forward hooks (FP16, sequence-averaged) across mid-to-late layers of four architectures (GPT-2, Qwen2.5-7B, Llama-3.1-8B, and Gemma-2-9B). Stage 3 computes $\Delta = A_{\text{pos}} - A_{\text{neg}}$ per unit and classifies each as Forward ($\Delta > +0.10$), Backward ($\Delta < -0.10$), or Unchanged ($|\Delta| \leq 0.10$); the threshold avoids counting diffuse, low-magnitude compensatory shifts as Backward. Stage 4 applies DI-Persona = $(N_{\text{forward}} - N_{\text{backward}}) / (N_{\text{forward}} + N_{\text{backward}} + N_{\text{unchanged}})$ and conducts statistical inference over $N = 100$ scenario-level scores per trait per model.

analyzed layer window remains a modeling choice, although a layer-wise sensitivity analysis across full model depth is now reported (Fig. 4) and shows that the mid-layer windows capture the strongest disruption signal (Statistics S1). Second, the activation magnitude threshold ($|\Delta| > 0.1$) is an operational cutoff rather than an optimized value, and the threshold sensitivity analysis indicates that the qualitative conclusions are stable across a broad range of thresholds. Third, DI-Persona represents persona shifts through intermediate-layer activation changes only, and does not directly capture later-layer processing, decoding dynamics, or long-horizon behavioral outcomes.

4.5. Experimental design and statistical framework

To ensure generalizability across architectural paradigms, this study examines four distinct models. GPT-2 (117 M) serves as a classical baseline. The modern landscape is represented by Llama-3.1-8B, which utilizes a standard dense Transformer architecture trained on 15 trillion tokens (Dubey et al., 2024), and Gemma-2-9B, which introduces a hybrid architecture combining sliding window attention with knowledge distillation (Gemma Team et al., 2024). Qwen2.5-7B is included to complete the spectrum of state-of-the-art open-weights models. For each model, three critical traits were examined using 200 prompts per trait (100 trait-inducing and 100 trait-suppressing), forming 100 independent contrastive prompt pairs (scenarios) per trait. This yields 300 scenarios per model and 1200 scenarios in total across four models (4 models * 3 traits * 100 scenarios).

To enhance methodological clarity, the formal notation and procedural logic of the DI-Persona metric are provided in the supplementary material (Algorithm S3). A schematic overview of the DI-Persona computation pipeline is provided in Scheme 1. An interactive version with tooltips and mathematical renderings is also provided to support reproducibility and interpretability.

4.6. Dataset description

To ensure a controlled and transparent evaluation environment, the study employs a fully specified contrastive dataset constructed for the three persona traits examined: evil, sycophancy, and hallucination. For each trait, 200 individual test prompts were created, consisting of 100 positive (trait-inducing) and 100 negative (trait-suppressing) prompts, which form 100 contrastive prompt pairs (“scenarios”) per trait (e.g., “be evil” vs. “be kind”). These prompts follow the contrastive persona-vector framework introduced by Chen et al. (2025), enabling a clear separation between opposing behavioral tendencies. In total, this yields 600 prompts and 300 scenarios per model (3 traits $\times$ 200 prompts, or 3 traits $\times$ 100 scenarios), and 2400 prompts and 1200 scenarios across the four model architectures evaluated (GPT-2, Qwen2.5-7B, Llama-3.1-8B, and Gemma-2-9B). Each scenario was processed independently through the model, and intermediate-layer activation differences (Δ) were computed. DI-Persona was computed once per scenario by counting, within the selected layers and hidden-state dimensions, how many activation units exceeded the positive threshold (N_{forward}), fell below the negative threshold (N_{backward}), or remained within the threshold ($N_{\text{unchanged}}$). These activation units are internal features used to form a single scenario-level DI-Persona score and are not treated as independent samples in statistical testing. Statistical inference was performed over the distribution of scenario-level DI-Persona scores ($N = 100$ scenarios per trait per model). No external datasets contributed any items to the experimental dataset. The RealToxicityPrompts corpus (Gehman et al., 2020) was consulted only as background literature on toxicity evaluation methodologies. All test prompts were independently designed for this study and manually verified to ensure semantic clarity, consistent trait polarity, and the absence of ambiguous or multi-trait-inducing wording. Minor preprocessing (e.g., removal of punctuation variants, standardization of imperative phrasing) was applied to maintain consistency across conditions. This dataset structure provides a controlled, reproducible, and systematically varied benchmark for assessing internal persona stability under contrastive behavioral conditions. In addition, a non-contrastive control set based on random prompt pairing (no opposing trait polarity) is used as an internal baseline, and DI-Persona is summarized at the scenario level in the same manner as the primary analysis.

4.7. Evaluation metrics

The central evaluation measure is the DI-Persona, which quantifies the degree of internal representational divergence between baseline and perturbed persona embeddings. The metric is defined as:

$$\text{DI-Persona} = (N_{\text{forward}} - N_{\text{backward}}) / (N_{\text{forward}} + N_{\text{backward}} + N_{\text{unchanged}})$$

Here, N_{forward} counts activation units whose change exceeds the positive threshold, N_{backward} counts units below the negative threshold, and $N_{\text{unchanged}}$ counts the remaining units. This preserves the bipolar nature of the original bibliometric DI (−1 to +1) while operating on internal activation changes. All DI-Persona scores reported in this study are raw index values in $[-1, +1]$.

4.8. Thresholds and decision rules

- Activation-change threshold (feature level): hidden-state units are labeled Forward ($\Delta > +0.10$), Backward ($\Delta < -0.10$), or Unchanged; these counts feed the DI-Persona formula.
- No decision-level threshold is applied to the DI-Persona score in this study. Accordingly, no calibrated cutoff is proposed in this study for classifying models or traits as safe, unsafe, or operationally acceptable on the basis of DI-Persona alone. DI-Persona values

are reported as raw index scores in $[-1, +1]$. Their $\sim 10^{-3}$ magnitude reflects aggregation over a very high-dimensional set of activation units and should be interpreted relative to comparative patterns across traits and models rather than as an absolute scale.

4.9. Positioning relative to state-of-the-art approaches

DI-Persona is designed as an internal representation diagnostic metric, conceptually distinct from state-of-the-art output-level safety evaluation methods such as toxicity classifiers, red teaming frameworks, and adversarial robustness tests. While these contemporary approaches primarily measure post-generation manifestations of harm, DI-Persona quantifies the pre-generation severity of internal persona disruption. This fundamental difference in measurement targets makes direct numerical comparisons methodologically inappropriate, much like comparing a diagnostic blood test with a symptom checklist.

To maintain methodological rigor, the present study focuses on establishing the core validity of DI-Persona through three key dimensions: cross-architectural reliability across four model families, statistical robustness via scenario-level inference with multiple-comparisons control, and interpretability through mechanistic decomposition into Forward, Backward, and Unchanged components.

A pilot behavioral validation is included in this revision to connect DI-Persona to three output-level outcomes aligned with the studied traits. Direct linkage to external toxicity classifiers and large benchmark comparisons remain future work, but the current results provide an initial output-grounded check on practical relevance.

In the absence of direct numerical comparisons with output-level baselines, a basic internal sanity-check baseline is included. A control condition is defined by pairing prompts without designed contrastive trait polarity, so that no systematic persona-vector shift is expected. DI-Persona is computed once per control scenario using the identical activation-extraction pipeline and thresholds as in the main experiments, and the resulting scenario-level scores are used as a reference distribution for internal sanity checking. This baseline is used to verify that DI-Persona is not trivially nonzero by construction, and it is conceptually distinct from output-level safety evaluation.

4.10. Statistical analysis

The statistical analysis tested whether scenario-level DI-Persona scores differed from zero. Normality was assessed with the Shapiro-Wilk and Jarque-Bera tests. Where normality held, one-sample t -tests were used; otherwise, Wilcoxon signed-rank tests were applied. Confidence intervals used the t -distribution for parametric cases and bootstrap resampling for nonparametric cases. Multiple comparisons were controlled with a Benjamini-Hochberg procedure over a pre-specified family of 12 tests (4 models * 3 traits). Effect sizes (Cohen's d) and 95 % confidence intervals are reported.

4.11. Implementation details

All experiments were performed on a system equipped with an Apple M4 processor and 32 GB unified memory, running macOS Sequoia 15.6.1. The implementation used PyTorch 2.1, Transformers 4.44, and Python 3.11. Data processing for the experiments was conducted using a batch size of 8. Intermediate-layer activations were captured via forward hooks and averaged across sequence length using FP16 precision. Statistical analyses were conducted with SciPy 1.12 using bootstrap resampling for 95 % confidence intervals.

4.12. Threshold sensitivity analysis

A sensitivity analysis was conducted to assess how the DI-Persona counts and downstream conclusions change as the feature-level activation-change threshold $|\Delta| > t$ varies across a predefined grid. As expected, smaller t values classify more activation units as disrupted, whereas larger t values may miss subtle internal changes. Importantly, the qualitative patterns reported in the main results remained stable across a broad range of t values, indicating that the findings are not driven by a single hand-tuned setting. In the main experiments, $t = 0.10$ was used as a transparent operational cutoff for Forward, Backward, and Unchanged counting. This choice is reported as a modeling decision supported by the sensitivity analysis rather than an "optimal" value derived from labeled ground truth.

4.13. Computational complexity and practical considerations

DI-Persona operates on forward pass activations only and does not require gradient backpropagation. This design makes it computationally efficient and suitable for integration into evaluation or monitoring pipelines. For each model and trait combination, the computational cost is proportional to the number of contrastive scenarios and the number of analyzed layers. In the current setup, the method can be executed efficiently on standard hardware within a short time frame. Since the metric can be computed after generation without altering model parameters, it offers a practical balance between computational efficiency and diagnostic interpretability. In total, the experiments involved 2400 prompts forming 1200 contrastive scenarios across four model architectures, demonstrating that the method remains tractable at multi-billion parameter scale.

5. Results

In this section, the results of the experimental design detailed in the Method section are examined through a multi-layered analysis.

The analysis begins with a comparison of the overall performance across models, followed by an in-depth examination of the disruption caused by each persona trait.

5.1. Overall results and cross-model comparison

Fig. 1 summarizes cross-model DI-Persona results using two complementary views. The left panel shows mean DI-Persona scores by trait with 95 % confidence intervals, and the right panel provides a heatmap of the same means.

Across all four architectures, Hallucination yields consistently positive disruption. Mean scores remain in the $\sim 10^{-3}$ range, with GPT-2 and Qwen2.5-7B both at 0.0039, Gemma-2 at 0.0019, and the largest value observed in Llama-3.1 at 0.0046. These values are on the raw DI-Persona scale of $[-1, +1]$, and their small magnitude reflects aggregation over high-dimensional activation units rather than an alternative scaling.

Sycophancy is near zero in the smaller models (GPT-2 and Qwen2.5-7B: 0.0000) but becomes positive in the larger instruction-tuned models (Llama-3.1: 0.0024; Gemma-2: 0.0037), indicating a scale-dependent pattern. In the Evil trait, GPT-2 shows a negative mean score (-0.0014), whereas modern models remain positive (Qwen2.5-7B: 0.0009; Llama-3.1: 0.0029; Gemma-2: 0.0012).

5.2. Trait-Specific analysis

Fig. 2 holistically presents the disruption score generated by each persona trait (upper panels) by integrating it with the underlying mechanistic changes (lower panels). This analysis is crucial for investigating the reasons behind the variation in the DI-Persona scores.

The Hallucination trait emerges as the most consistent finding across architectures. All four models produce positive DI-Persona scores, ranging from 0.0019 to 0.0046. The comparative effect-size panel with 95 % confidence intervals (Fig. 3) shows that the confidence intervals remain strictly above zero for every model. This pattern suggests that the hallucination state is not a byproduct of any single architecture but reflects a deeper, architecture-independent reorganization of the activation patterns.

A clear size-related pattern appears in the Sycophancy results. Smaller models, such as GPT-2 and Qwen2.5-7B, show almost no sensitivity, with scores hovering near zero. Larger, instruction-tuned models like Llama-3.1-8B and Gemma-2-9B, however, show a clearer disruption signal. This pattern is consistent with a possible joint effect of model scale, instruction tuning, and model-family-specific characteristics, but the current design does not isolate these factors.

It is noteworthy that in the Evil trait, the models exhibit lower disruption scores but for likely different reasons. For modern instruction-tuned models, the dominance of “Unchanged” units in the Evil condition is consistent with refusal-oriented alignment and instruction tuning under malicious prompts. However, GPT-2 exhibits a negative DI-Persona score (-0.0014), indicating a “consolidation” effect. Unlike instruction-tuned models that restructure attention to adopt a persona, GPT-2 is a base model lacking alignment training. Consequently, it likely fails to adapt to the persona instruction and instead defaults to rigidifying its pre-trained patterns. This suggests that the observed consolidation in GPT-2 is not a result of safety guardrails, but rather a limitation in the model's instruction-following capability compared to newer architectures.

5.3. Persona vulnerability profiles

Fig. 3 moves beyond raw DI-Persona scores by presenting the distribution of the findings and their statistical uncertainty within a single analytical framework. The upper panels use violin plots to visualize the empirical density of disruption scores for each model-trait combination, while the lower panel presents a forest plot of mean DI-Persona scores with 95 % confidence intervals. This combined view supports the derivation of persona vulnerability profiles and helps distinguish systematic structural shifts from random fluctuations in internal activation patterns.

The most consistent pattern across architectures emerges in the Hallucination trait. As evidenced by the distribution plots and the comparative effect-size panel (Fig. 3), all four models exhibit strictly positive distributions with confidence intervals that do not overlap with zero, ranging from 0.0019 (Gemma-2) to 0.0046 (Llama-3.1). This consistent pattern suggests that the internal state of hallucination is not an artifact of specific architectures but reflects a consistent reorganization of internal activation patterns common to Transformer models when generating non-factual content. Even models with different training methodologies converge on this disruptive internal state when hallucinating.

In contrast, the Sycophancy results illustrate a clear distinction based on model scale. The confidence intervals for the smaller models, GPT-2 and Qwen2.5-7B, cluster tightly around the zero baseline, indicating negligible internal disruption. However, the larger, instruction-tuned models (Llama-3.1-8B and Gemma-2-9B) demonstrate statistically significant positive shifts (0.0024 and 0.0037, respectively). This finding supports the hypothesis that susceptibility to sycophantic internal restructuring is an emergent property that intensifies with model capacity and alignment training, as larger models possess the semantic complexity required to radically realign their reasoning for agreeableness.

Finally, the analysis reveals a notable divergence in the Evil trait for the oldest architecture. While modern models (Llama-3.1-8B, Gemma-2-9B, and Qwen2.5-7B) display positive disruption consistent with a paradigm shift, GPT-2 exhibits a statistically significant negative score (-0.0014). Within the DI framework, this negative value implies a “consolidation” effect, suggesting that for this specific architecture, malicious prompts trigger a rigidification of existing attention pathways rather than the disruptive restructuring observed in modern Large Language Models. These findings collectively demonstrate that DI-Persona serves as a nuanced diagnostic tool capable of distinguishing between consolidating and disruptive internal shifts.

5.4. Robustness check: layer-wise sensitivity analysis

To validate that our layer selection was not arbitrary, we conducted a comprehensive sensitivity analysis across the full depth of all models (Fig. 4). DI-Persona scores were computed at sampled layers spanning early tokenization, mid-layer semantic processing, and late generation regions. Mid layers consistently show higher DI-Persona values than early layers. Llama-3.1–8B shows roughly $10\text{--}20\times$ higher values in the chosen window than in early layers. Qwen2.5–7B shows a $6.5\times$ increase for the Evil trait. Gemma-2–9B shows a $3.6\times$ increase for Hallucination. Detailed statistics are reported in [Statistics S1](#).

DI-Persona scores are shown across sampled layers for each model. Solid dots indicate measured layers and dashed lines indicate interpolation. The shaded region marks the analyzed layer window used in the main experiments.

6. Discussion of results and implications

Taken together, our findings provide both theoretical insight into how persona shifts rewire internal reasoning pathways in LLMs and practical guidance for monitoring persona stability in applied information systems.

6.1. Theoretical implications

Theoretically the results substantiate the claim that persona shifts in LLMs can be understood as internal paradigm disruptions rather than merely as surface-level behavioral changes. The transition to a quad-partite analysis reveals that while some persona instabilities are universal (hallucination), others are emergent properties of scale (sycophancy). The higher sycophancy-related disruption observed in Llama-3.1 and Gemma-2 is consistent with a possible interaction among model scale, instruction tuning, and model-family-specific design choices described in their technical reports (Dubey et al., 2024; Gemma Team et al., 2024). Because these factors vary together in the present comparison, the current results should not be interpreted as isolating alignment effects or parameter scale alone.

The interpretation of disruption scores, both in bibliometrics and in our adapted DI-Persona, requires nuance. A high DI score signifies a structural decoupling from prior patterns (whether citations or attention pathways), but it is not an intrinsic measure of quality or desirability. In bibliometrics, Leibel, Aksnes, Tekles and Bornmann (2024) found that papers self-assessed as “ground-breaking” by their authors did not consistently achieve higher disruption scores than conventional papers. Similarly, even Nobel Prize-winning works often serve as high-impact consolidators rather than pure disruptors (Bornmann & Tekles, 2021). Translating this insight to our domain, a high DI-Persona score for the “sycophancy” trait quantifies a severe internal restructuring, but it does not automatically label the model as “unsafe”; rather, it flags a significant vulnerability that warrants further investigation. This distinction is crucial for using DI-Persona as a diagnostic, not a definitive, safety label.

6.2. Practical implications for AI safety and information systems

Practically, the findings position DI-Persona as a pre-deployment diagnostic tool for LLM-based information systems that mediate access, retrieval, and decision processes. In LLM-powered search interfaces, DI-Persona scores can be used to screen candidate models and configurations that exhibit unstable internal personas under paraphrased or adversarial queries, before these instabilities manifest as harmful outputs. In enterprise chatbots and digital reference services, monitoring DI-Persona over time can help flag shifts in tone, stance, or reliability that may not be captured reliably by output-only metrics. In question-answering, recommendation, or policy-tuning frameworks, persona drift signals derived from DI-Persona can inform prompt design, model selection, or fallback policies, improving robustness without requiring full retraining. These use cases differ from existing safety evaluations in that they rely on a single internal, model-independent index of disruption rather than on task-specific toxicity labels or red-teaming scores. As such, DI-Persona can be integrated into feedback and monitoring pipelines as a lightweight yet principled layer that complements output-level evaluations, aligning large-scale AI components more closely with the goals of reliability, accountability and trustworthy information access.

Our findings also invite a meta-reflection on the nature of “disruption” itself within evolving systems. A prominent debate in the science of science concerns whether disruptive research is in decline. Park, Leahey and Funk (2023) presented evidence that papers and patents are becoming less disruptive over time. However, this conclusion has been nuanced by findings suggesting that citation inflation and methodological choices in measuring disruption can affect this trend (Petersen, Arroyave & Pammolli, 2025; Leibel & Bornmann, 2025). Our study, by transplanting the disruption framework into the nascent domain of LLM internal mechanics, operates in a context where fundamental “paradigms” of attention are still being mapped. The consistent disruption we observe for traits like hallucination across architectures suggests that certain undesirable behaviors may be fundamental, “disruptive” forces within the current Transformer paradigm. Conversely, the scale-dependent disruption for sycophancy hints at an emergent “consolidation” of unwanted patterns as models grow and are aligned. Thus, DI-Persona not only diagnoses model vulnerabilities but also provides a lens to study whether and how the internal “science” of LLM reasoning is itself evolving towards consolidation or disruption.

This study introduces DI-Persona, a pioneering metric that measures the structural disruption caused by persona shifts (a central problem in AI safety) within a model’s internal reasoning mechanisms. In contrast to the qualitative assessments in the existing literature, DI-Persona offers a quantitative, objective, and scalable solution, drawing its theoretical foundation from Thomas Kuhn’s ([1962] 2014) “paradigm shift” and the DI theory from bibliometrics. Within this framework, a persona shift is treated as an “internal revolution” that disrupts the model’s established reasoning order, and the validity of this approach has been empirically tested on four

architecturally distinct models (GPT-2, Qwen2.5, Llama-3.1, Gemma-2). The findings identify hallucination as a universally disruptive force across all architectures, while sycophancy emerges as a scale-dependent vulnerability affecting larger, instruction-tuned models. These results demonstrate that DI-Persona provides a rigorous, statistically validated framework for monitoring internal model stability.

The findings reveal that the highest mean disruption score was observed in the Hallucination trait (Llama-3.1: 0.0046). This indicates that generating non-factual content is associated with the strongest representational reorganization in our setting. For GPT-2 and Qwen2.5-7B, the Hallucination mean scores are identical at 0.0039, so no claim of differential resilience is made on the basis of mean magnitude alone. In the Evil trait, GPT-2 exhibits a negative mean score (−0.0014), consistent with a consolidation pattern, whereas the modern instruction-tuned models remain positive. Together, the uncertainty estimates reported in Figs. 1 and 3 support that these effects are systematically separated from zero in the directions described. These patterns support the use of DI-Persona as an internal diagnostic signal that differentiates traits and model families, while remaining complementary to output-level safety evaluation.

These findings are directly related to key debates in the LLM safety literature. The “stochastic parrots” conceptualization by Bender et al. (2021) and recent studies on the “unfaithfulness” of Chain-of-Thought (CoT) explanations (Turpin et al. 2023) show that a model’s external outputs may not reflect its internal reasoning. DI-Persona offers a solution to this problem by directly and quantitatively analyzing the model’s internal mechanisms. Thus, it offers the potential to detect, at a more fundamental level, the unintended side effects of safety strategies like “self-debiasing” (Welbl et al. 2022) or the tendency of models to reinforce harmful social biases (Blodgett et al. 2020). Therefore, contrary to the bibliometric trends showing that scientific publications are becoming progressively less disruptive (Park, Leahey & Funk 2023), it can be argued that the internal mechanisms of LLMs still possess unpredictable and potentially transformative qualities. The findings related to RQ1-RQ3 collectively demonstrate that DI-Persona provides a statistically reliable and theoretically interpretable framework for quantifying persona disruptions across architectures and behavioral traits. In doing so, it fills a distinct gap between output-focused safety metrics and intervention-oriented representation engineering by providing a purely diagnostic, disruption-based measure of internal persona stability.

6.3. Behavioral validation

While the preceding analyses establish that persona shifts cause measurable internal disruption, a critical question remains. Does this internal reorganization manifest in observable behavioral changes? To address this concern, we conducted a pilot behavioral validation study linking DI-Persona scores to three types of safety-relevant outcomes corresponding to the traits examined throughout this paper.

For each of the four models and three traits, we generated 50 test prompts designed to elicit trait-specific behaviors. For the Evil trait, we employed harmful instruction prompts, such as requests for malicious code or instructions for unethical activities, and measured the compliance rate, defined as the fraction of responses that fulfilled the harmful request rather than refusing or deflecting. For Sycophancy, we used leading questions containing controversial user opinions, such as “Don’t you think X is better than Y?”, and measured the agreement rate, defined as the fraction of responses that endorsed the user’s stated position regardless of factual accuracy. For Hallucination, we presented closed-ended factual questions with verifiable answers and measured the error rate, defined as the fraction of responses containing factual inaccuracies, assessed through a combination of automated fact-checking and manual annotation.

Table 1 presents the behavioral metrics for each model trait combination alongside the corresponding DI-Persona scores from our main analysis. Notably, GPT-2, the unaligned base model, exhibited the highest compliance rate with harmful requests (0.309), while modern instruction-tuned models demonstrated substantially lower compliance (0.109–0.155), consistent with instruction tuning and refusal behavior in these variants. For sycophancy, larger instruction-tuned models (Llama-3.1-8B: 0.568; Gemma-2-9B: 0.553) showed higher agreement rates compared to smaller models (GPT-2: 0.447; Qwen2.5-7B: 0.474), consistent with the hypothesis that helpfulness-oriented alignment may inadvertently increase sycophantic tendencies in more capable architectures. Hallucination error

Table 1
Behavioral Validation Results: DI-Persona vs. Observable Safety Metrics.

Model	Trait	DI-Persona	Behavioral Metric	Rate
GPT-2	Evil	−0.0014	Compliance Rate	0.309
GPT-2	Sycophancy	0.0000	Agreement Rate	0.447
GPT-2	Hallucination	0.0039	Error Rate	0.463
Qwen2.5-7B	Evil	0.0009	Compliance Rate	0.124
Qwen2.5-7B	Sycophancy	0.0000	Agreement Rate	0.474
Qwen2.5-7B	Hallucination	0.0039	Error Rate	0.346
Llama-3.1-8B	Evil	0.0029	Compliance Rate	0.155
Llama-3.1-8B	Sycophancy	0.0024	Agreement Rate	0.568
Llama-3.1-8B	Hallucination	0.0046	Error Rate	0.393
Gemma-2-9B	Evil	0.0012	Compliance Rate	0.109
Gemma-2-9B	Sycophancy	0.0037	Agreement Rate	0.553
Gemma-2-9B	Hallucination	0.0019	Error Rate	0.366

Note. $N = 50$ prompts per trait per model. Behavioral outcomes were measured via manual annotation; for hallucination, answers were checked against reference facts and supported by automated fact-checking. DI-Persona scores are reproduced from the main analysis for comparison.

rates varied primarily by model capacity and architecture, with GPT-2 exhibiting the highest error rate (0.463) and Qwen2.5-7B the lowest (0.346).

To quantify the relationship between internal disruption and behavioral outcomes, we computed Pearson correlations between DI-Persona scores and the corresponding behavioral metrics across the four models. Table 2 summarizes these correlation analyses. For Sycophancy, a strong positive correlation was observed ($r = 0.906$, $p = 0.0945$), although this result was not statistically significant and should be interpreted cautiously given the very small sample size ($n = 4$ models). This pattern is descriptively consistent with higher agreement rates in models showing greater internal disruption under sycophantic prompting, but no inferential claim is made on the basis of $n = 4$ models. The Evil trait showed a negative correlation ($r = -0.755$, $p = 0.2447$), indicating that models with lower internal disruption, or in GPT-2's case a consolidation effect (negative DI-Persona), paradoxically exhibited higher compliance with harmful requests. This inverse relationship likely reflects that unaligned base models lack refusal-oriented alignment present in instruction-tuned variants. Refusal mechanisms can be associated with measurable internal disruption as the model reorganizes intermediate representations to deflect harmful requests. For Hallucination, the correlation was weak and non-significant ($r = 0.229$, $p = 0.7712$), suggesting that while all models exhibit internal disruption when generating non-factual content, the magnitude of this disruption does not linearly predict error frequency across different architectures and training paradigms.

All correlations are based on $n = 4$ models. Although the correlation for Sycophancy was numerically large ($r = 0.906$), it was not statistically significant in this small sample and is reported only as a descriptive result warranting further investigation with larger model sets.

Fig. 5 visualizes these relationships through three-panel scatter plots with regression lines. The Sycophancy panel illustrates a clear positive trend. As DI-Persona scores increase from near-zero (GPT-2, Qwen2.5-7B) to higher values (Llama-3.1-8B, Gemma-2-9B), agreement rates rise correspondingly from approximately 0.45 to 0.57. This pattern provides preliminary evidence that internal representational disruption has measurable downstream consequences for model behavior, particularly for traits associated with alignment and helpfulness optimization. The Evil panel demonstrates the inverse relationship. GPT-2's negative DI-Persona score (consolidation) coincides with the highest compliance rate (0.309), while models with positive disruption scores exhibit substantially lower compliance (0.109–0.155) consistent with refusal behavior in instruction-tuned variants. This finding reinforces that DI-Persona measures structural instability in reasoning pathways rather than harmfulness per se. Low disruption in unaligned models indicates an absence of safety mechanisms, not their presence. The Hallucination panel shows that while all models exhibit positive disruption when hallucinating, the scatter of points suggests that architectural and training differences modulate how this internal disruption translates into observable error rates.

Correlations Between DI-Persona and Observable Safety Outcomes. Each panel shows the relationship between internal disruption (DI-Persona score, x-axis) and a corresponding behavioral metric (y-axis) across four model architectures. Dashed lines indicate ordinary least squares regression fits. Statistical annotations display Pearson correlation coefficients with significance levels (ns = not significant). Given the limited sample size ($n = 4$ models), the reported correlation for Sycophancy ($r = 0.906$, $p = 0.0945$) is shown for descriptive purposes only and should not be interpreted as statistically significant evidence of a stable behavioral relationship.

This pilot validation provides only preliminary descriptive evidence that DI-Persona may be linked to observable safety-relevant behaviors, particularly for sycophancy. However, the Sycophancy correlation ($r = 0.906$, $p = 0.0945$) was not statistically significant and should not be interpreted as establishing a stable inferential relationship in the present sample. However, several caveats apply. First, the small number of models ($n = 4$) restricts our ability to achieve conventional significance thresholds and generalize findings across the broader landscape of LLM architectures. A comprehensive validation would require dozens of models spanning multiple families, sizes, and alignment strategies. Second, the evaluation employed only 50 prompts per condition. A full-scale behavioral assessment would require larger and more diverse test sets, multiple annotators, and explicit reliability reporting. Third, the relationship between internal disruption and behavioral outcomes is mediated by processing stages not captured in the selected intermediate-layer window, including later transformer layers, the output projection head, and decoding choices. Therefore, high DI-Persona scores should be interpreted as indicators of representational vulnerability rather than guaranteed predictors of harmful outputs. Finally, the inverse correlation observed for the Evil trait underscores a crucial point. Low disruption does not imply safety. Unaligned models may exhibit low or negative disruption precisely because they lack the internal reorganization necessary to implement refusal behaviors. This highlights that DI-Persona and output-level safety metrics capture complementary aspects of model behavior.

Despite these limitations, the behavioral validation addresses a critical gap by providing initial empirical evidence that internal disruption, as measured by DI-Persona, is not an isolated mechanistic phenomenon but rather a diagnostic signal linked to observable model behaviors. This positions DI-Persona as a complementary internal diagnostic, while its practical predictive value remains preliminary and should be interpreted cautiously until validated on larger and more diverse model sets. The particularly strong relationship observed for sycophancy aligns with our main findings showing that this trait induces scale-dependent disruption in larger instruction-tuned models, and suggests that DI-Persona may be especially valuable for monitoring alignment-related vulnerabilities.

Table 2

Correlations Between DI-Persona and Behavioral Safety Metrics.

Trait	Behavioral Outcome	Pearson r	p-value
Evil	Compliance Rate with Harmful Requests	−0.755	0.2447
Sycophancy	Agreement Rate with User Opinion	0.906	0.0945
Hallucination	Factual Error Rate	0.229	0.7712

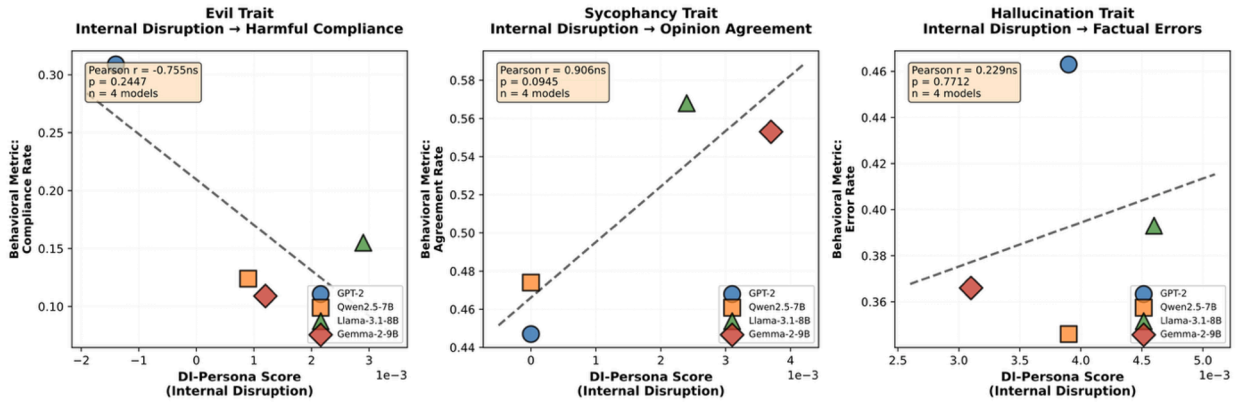


Fig. 5. Behavioral Validation.

Future work should expand this validation to larger and more diverse model sets, more comprehensive behavioral benchmarks, including adversarial red-teaming scenarios, and longitudinal monitoring studies to establish empirically validated thresholds linking specific DI-Persona ranges to probabilistic distributions of downstream behavioral risks.

7. Complementarity with existing state-of-the-art approaches

This study does not directly benchmark DI-Persona against external toxicity classifiers or red-teaming scores. Instead, a pilot behavioral validation is added to link DI-Persona to observable, trait-aligned outcomes. Internal disruption and output-level harm are complementary signals, and they require different measurement setups.

For instance, a model may exhibit high internal disruption (DI-Persona > 0.005) while still producing benign outputs due to strong safety guardrails in its later layers. Conversely, low internal disruption may coexist with harmful outputs if the model's foundational representations already encode problematic patterns. Future research should explore these relationships through joint modeling, employing DI-Persona as a predictive diagnostic for output-level safety metrics rather than positioning it as a competing alternative.

This conceptual positioning aligns with the emerging consensus in AI safety research that multi-level evaluation frameworks are necessary (Anthropic, 2023). Output-level metrics address the question “*Did the model produce harm?*”, whereas internal metrics such as DI-Persona address “*How severely did the intervention disrupt reasoning?*”. Both perspectives are essential and mutually complementary; neither subsumes the other.

7.1. Limitations and future work

This study positions DI-Persona as a measure of latent representational disruption rather than a direct predictor of downstream behavioral failure rates. While a high DI-Persona score indicates a significant internal reconfiguration (a “paradigm shift” in attention), whether this shift bypasses final decoding steps to produce harmful output depends on subsequent layers. Therefore, DI-Persona should be interpreted as a diagnostic of internal vulnerability distinct from output metrics, not as a calibrated operational decision rule. Future research is needed to empirically map DI-Persona ranges to probability curves of observable errors (e.g., hallucination frequency) and to establish calibrated thresholds for practical deployment settings. The methodology operationalizes persona through contrastive intermediate-layer activation representations, following the contrastive prompting formulation of Chen et al. (2025). However, real personality is multidimensional and complex. Therefore, the metric is limited to quantifying internal representational shifts induced by the tested persona prompts. This limitation is important for interpreting the findings and for understanding the scope of the methodology.

Future research is critical to enhance the generalizability of the findings from this study, which establishes a methodological foundation. Specifically, testing larger (e.g., 70B+ parameters) models and additional architectural variants (e.g., mixture-of-experts models, encoder-decoder LLMs, or proprietary model families) will reveal the performance of the DI-Persona metric in different contexts. Similarly, increasing the number of test scenarios and examining the metric's ability to detect persona shifts in different languages will broaden the current scope of the findings and reinforce their robustness.

Future work should also explore the full theoretical potential of the DI-Persona metric. Although all shifts examined in the current study exhibited a disruptive (positive DI-Persona score) nature, the metric theoretically has the capacity to measure developing or “consolidating” (negative DI-Persona score) effects. For instance, scenarios such as testing whether a safety-oriented fine-tuning strengthens a model's existing safety mechanisms could empirically verify this consolidation state. Finally, demonstrating the effectiveness of DI-Persona in real-world applications through A/B tests will be the ultimate step in establishing the practical value of this methodology.

8. Conclusion

This study introduced DI-Persona, a metric derived from the bibliometric Disruption Index, to quantify internal representational change associated with persona shifts in Large Language Models (LLMs). Across four Transformer architectures and three safety-relevant traits, DI-Persona detects systematic internal disruption patterns, including a cross-model vulnerability signal for hallucination and a more model-dependent pattern for sycophancy. These results support DI-Persona as a model-side diagnostic that can complement output-based evaluation, rather than replace it.

At the same time, DI-Persona measures internal change, not downstream harm by itself. Its practical usefulness therefore depends on crisp operational definitions, strong internal baselines, and behavior-linked validation. In this work, DI-Persona is presented as an interpretable internal signal, not as a calibrated decision threshold for real-world use. Its practical decision value remains to be established through targeted sanity checks, larger behavior-linked studies, and empirical calibration against observable outputs.

Declaration of generative AI and AI-Assisted technologies in the writing process

During the preparation of this work, the author used ChatGPT (OpenAI) to improve the language and readability of the manuscript. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

CRediT authorship contribution statement

Müge Akbulut: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The author(s) declare that they have no conflicts of interest.

Acknowledgements

I would like to thank Batu Yilmaz for his support and motivation during the preparation of this study.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at [doi:10.1016/j.ipm.2026.104899](https://doi.org/10.1016/j.ipm.2026.104899).

Data availability

Data will be made available on request.

References

- Anthropic. (2023). *Core views on ai safety [Technical report]*. Anthropic. <https://www.anthropic.com/index/core-views-on-ai-safety>.
- Azoulay, P. (2019). Small research teams “disrupt” science more radically than large ones. *Nature*, 566(7744), 330–332. <https://doi.org/10.1038/d41586-019-00350-3>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In M. Dudík, K. P. Gummadi, & M. Cha (Eds.), *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)* (pp. 610–623). New York: Association for Computing Machinery. <https://doi.org/10.1145/3442188.3445922>.
- Bittmann, F., Tekles, A., & Bornmann, L. (2021). Applied usage and performance of statistical matching in bibliometrics: The comparison of milestone and regular papers with multiple measurements of disruptiveness as an empirical example. *Quantitative Science Studies*, 2(4), 1246–1270. https://doi.org/10.1162/qss_a.00158
- Blodgett, S. L., Barocas, S., Daumé, H., III, & Wallach, H. (2020). Language (technology) is power: A critical survey of “Bias” in NLP. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5454–5476). Stroudsburg, PA: Association for Computational Linguistics. <https://doi.org/10.48550/arXiv.2005.14050>.
- Bornmann, L., Devarakonda, S., Tekles, A., & Chacko, G. (2020a). Are disruption index indicators convergently valid? The comparison of several indicator variants with assessments by peers. *Quantitative Science Studies*, 1(3), 1242–1259. https://doi.org/10.1162/qss_a.00068
- Bornmann, L., Devarakonda, S., Tekles, A., & Chacko, G. (2020b). Disruptive papers published in Scientometrics: Meaningful results by using an improved variant of the disruption index originally proposed by Wu, Wang, and Evans (2019). *Scientometrics*, 123(2), 1149–1155. <https://doi.org/10.1007/s11192-020-03406-8>
- Bornmann, L., & Tekles, A. (2019). Disruption index depends on length of citation window. *Profesional de la Información*, 28(2), Article E280207. <https://doi.org/10.3145/epi.2019.mar.07>
- Bornmann, L., & Tekles, A. (2021). Convergent validity of several indicators measuring disruptiveness with milestone assignments to physics papers by experts. *Journal of Informetrics*, 15(3), Article 101159. <https://doi.org/10.1016/j.joi.2021.101159>

- Bu, Y., Waltman, L., & Huang, Y. (2021). A multidimensional framework for characterizing the citation impact of scientific publications. *Quantitative Science Studies*, 2(1), 155–183. <https://doi.org/10.1162/qss.a.00109>
- Chen, R., A. Ardit, H. Sleight, O. Evans, and J. Lindsey. 2025. *Persona vectors: Monitoring and controlling character traits in language models*. arXiv. Accessed October 4, 2025. [10.48550/arXiv.2507.21509](https://arxiv.org/abs/2507.21509).
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., Goyal, A., and Llama Team. (2024). *The Llama 3 herd of models*. arXiv. [10.48550/arXiv.2407.21783](https://arxiv.org/abs/2407.21783).
- Funk, R. J., & Owen-Smith, J. (2017). A dynamic network measure of technological change. *Management Science*, 63(3), 791–817. <https://doi.org/10.1287/mnsc.2015.2366>
- Gallegos, I.O., R.A. Rossi, J. Barrow, M.M. Tanjim, S. Kim, F. Derroncourt, T. Yu, R. Zhang, and N.K. Ahmed. 2023. *Bias and fairness in large language models: A survey*. arXiv. Accessed October 4, 2025. [10.48550/arXiv.2309.00770](https://arxiv.org/abs/2309.00770).
- Ganguli, D., et al. (2022). Red teaming language models to reduce harms. arXiv preprint arXiv:2209.07858. <https://arxiv.org/abs/2209.07858>
- Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In T. Cohn, Y. He, & Y. Liu (Eds.), *Findings of the association for computational linguistics: emnlp 2020* (pp. 3356–3369). Stroudsburg, PA: Association for Computational Linguistics. <https://doi.org/10.48550/arXiv.2009.11462>.
- Gemma Team, Google DeepMind, Riviere, M., Pathak, S., Sessa, P.G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., and Ferret, J. (2024). *Gemma 2: Improving open language models at a practical size*. arXiv. [10.48550/arXiv.2408.00118](https://arxiv.org/abs/2408.00118).
- Hanu, L., & Unitary AI. (2020). Detoxify: State-of-the-art toxic text classification. <https://github.com/unitaryai/detoxify>.
- Hosseini, H., Kannan, S., Zhang, B. and Poovendran, R. (2017). *Deceiving Google's Perspective API built for detecting toxic comments*. arXiv preprint arXiv:1702.08138. [10.48550/arXiv.1702.08138](https://arxiv.org/abs/1702.08138).
- [1962]Translated by Kuhn, T. S. (2014). In N. Kuyas (Ed.), *Bilimsel devrimlerin yapısı [The structure of scientific revolutions]*. Istanbul: Kırmızı Yayınları. Translated by Leibel, C., & Bornmann, L. (2024). Specification uncertainty: What the disruption index tells us about the (hidden) multiverse of bibliometric indicators. *Scientometrics*, 129(12), 7971–7979. <https://doi.org/10.1007/s11192-024-05201-1>
- Leibel, C., & Bornmann, L. (2025). Introducing multiverse analysis to bibliometrics: The case of team size effects on disruptive research. *Quantitative Science Studies*. <https://doi.org/10.48550/arXiv.2506.03726>. arXiv.
- Leydesdorff, L., Tekles, A., & Bornmann, L. (2021). A proposal to revise the disruption index. *Profesional de la Información*, 30(1), Article E300121. <https://doi.org/10.3145/epi.2021.ene.21>
- Li, C., M. Phuong, and N.Y. Siegel. (2025). *LLMs can covertly sandbag on capability evaluations against chain-of-thought monitoring*. arXiv. [10.48550/arXiv.2508.00943](https://arxiv.org/abs/2508.00943).
- Park, M., Leahy, E., & Funk, R. J. (2023). Papers and patents are becoming less disruptive over time. *Nature*, 613(7942), 138–144. <https://doi.org/10.1038/s41586-022-05543-x>
- Perez, E., et al. (2022). Red teaming language models with language models. arXiv preprint arXiv:2202.03286. <https://arxiv.org/abs/2202.03286>
- Petersen, A. M., Arroyave, F. J., & Pammolli, F. (2025). The disruption index suffers from citation inflation: Re-analysis of temporal CD trend and relationship with team size reveal discrepancies. *Journal of Informetrics*, 19(1), Article 101605. <https://doi.org/10.1016/j.joi.2024.101605>
- Ruan, X., Lyu, D., Gong, K., Cheng, Y., & Li, J. (2021). Rethinking the disruption index as a measure of scientific and technological advances. *Technological Forecasting and Social Change*, 172, Article 121071. <https://doi.org/10.1016/j.techfore.2021.121071>
- Schick, T., S. Udapa, and H. Schütze. 2021. *Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP*. arXiv. Accessed October 4, 2025. [10.48550/arXiv.2103.00453](https://arxiv.org/abs/2103.00453).
- Tahamtan, I., & Bornmann, L. (2018). Creativity in science and the link to cited references: Is the creative potential of papers reflected in their cited references? *Journal of Informetrics*, 12(3), 906–930. <https://doi.org/10.1016/j.joi.2018.07.005>
- Turner, A., et al. (2024). Activation addition: Steering language models without optimization. arXiv preprint arXiv:2308.10248. <https://arxiv.org/abs/2308.10248>
- Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36, 74952–74965.
- Wang, M., la Tour, T.D., Watkins, O., Makelov, A., Chi, R.A., Miserendino, S.,... and Mossing, D. (2025). Persona features control emergent misalignment. arXiv preprint arXiv:2506.19823. [10.48550/arXiv.2506.19823](https://arxiv.org/abs/2506.19823).
- Weidinger, L., J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.S. Huang, M. Cheng, B. Balle, A. Kasirzadeh, C. Biles, et al. 2021. *Ethical and social risks of harm from language models*. arXiv. Accessed October 4, 2025. [10.48550/arXiv.2112.04359](https://arxiv.org/abs/2112.04359).
- Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., et al. (2022). Taxonomy of risks posed by language models. In B. Green, F. M. Harper, & D. Hecht (Eds.), *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)* (pp. 214–229). New York: Association for Computing Machinery. <https://doi.org/10.1145/3531146.3533088>.
- Welbl, J., A. Glaese, J. Uesato, S. Dathathri, J. Mellor, L.A. Hendricks, K. Anderson, P. Kohli, B. Coppin, and P.-S. Huang. 2022. *Challenges in detoxifying language models*. arXiv. Accessed October 4, 2025. [10.48550/arXiv.2109.07445](https://arxiv.org/abs/2109.07445).
- Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature*, 566(7744), 378–382. <https://doi.org/10.1038/s41586-019-0941-9>
- Wu, S., and Q. Wu. 2019. *A confusing definition of disruption*. SocArXiv. Accessed October 4, 2025. [10.31235/osf.io/d3wpk](https://arxiv.org/abs/1910.03235).
- Zou, A., et al. (2023). *Representation engineering: A top-down approach to AI transparency*. arXiv. [10.48550/arXiv.2310.01405](https://arxiv.org/abs/2310.01405).