Applied Computer Systems
ISSN 2255-8691 (online)
2026, vol. 31, no. 1, pp. 206–213
https://doi.org/10.2478/acss-2026-0017
https://content.sciendo.com

Measuring and Mitigating Silent Narrative Switching in Large Language Models through Epistemic Mediation

Iurii Lukianchuk[1*], Lev Boiko[2], Inna Kondius[3]
[1–3]*Lutsk National Technical University, Lutsk, Ukraine*

***Abstract* – The paper defines silent narrative switching, the covert change of a large language model's dominant position across user contexts without disclosing that alternatives exist, and measures it with a Silent Narrative Switching Index (SNSI). Across four models from different providers, SNSI ranged from 0.73 to 0.80. An execution-level epistemic mediation layer over unmodified models reduced the pooled SNSI from 0.82 to 0.09 without retraining and without a detectable loss of contextual relevance; an ablation study locates most of the effect in the interpretation-generation stage.**

***Keywords* – Ablation study, context sensitivity, epistemic mediation, human–AI interaction, large language models, narrative switching, pluralism, transparency, trust, verification.**

I. Introduction

Large language models (LLMs) have moved from research prototypes to components embedded in everyday digital environments, including learning-management systems, advisory services and public-facing assistants [1]–[3]. In such settings, users increasingly pose questions that do not have a single correct answer but admit several legitimate interpretations rooted in different scientific, cultural or national traditions. How a model behaves on such questions is not merely a matter of accuracy; it is a matter of transparency and trust, which are central concerns of applied computer systems and of human–computer interaction.

A specific and under-examined trust concern motivates this study. A model may present a confident, single-perspective answer whose substantive position shifts according to cues in the user context – for example, a system prompt that signals an innovation-oriented or a caution-oriented stance – while never disclosing that other legitimate positions exist. Related evidence shows that model outputs follow opinion cues contained in their inputs [1], [4] and that models tend to align their answers with views signalled by the user [5]. We refer to this behaviour as silent narrative switching: the covert change of the dominant interpretive frame across contexts, unaccompanied by any acknowledgement of plurality. From the user's standpoint, each individual answer appears authoritative and complete, which makes the behaviour difficult to detect and therefore particularly consequential for informed decision-making. Unlike a factual error, which can in principle be checked against evidence, a silent switch leaves no visible trace within a single interaction: the answer is internally coherent, and only comparison across contexts reveals that the position was contingent on cues the user did not intend as a request for a particular stance.

A concrete example illustrates the behaviour. In the measurement study reported below, the stimulus is a question about a historical event that some national historiographies present as a war of liberation and others as a rebellion. Asked in a context associated with the first tradition, a model tends to answer within the liberation frame; asked in a context associated with the second, it tends to answer within the rebellion frame; in neither case does the answer mention that the other interpretation exists. Each answer, read alone, looks complete, and only the comparison across contexts reveals the switch.

Prior work has established that LLMs reflect the opinions and expectations embedded in their inputs [1], [4] and that their outputs are sensitive to small changes in prompts [6], [7]. Approaches to greater transparency and pluralism have been proposed largely at the level of model training – through reinforcement learning from human feedback [8], constitutional training [9], and pluralistic alignment [10], [11]. What remains open is a directly measurable account of covert stance-switching and a mitigation that operates at the execution level, over an unmodified model, designed so that plurality is disclosed regardless of how the question is phrased.

The overall aim of the study is to make covert, context-dependent changes of stance in LLM answers measurable and to show that they can be removed at the execution level, without retraining and without a loss of contextual relevance. To this end, the study addresses three research questions. RQ1: Is silent narrative switching a systematic and measurable property across contemporary LLMs of different providers? RQ2: Can an execution-level architecture reduce it while preserving the contextual relevance of answers? RQ3: Which components of such an architecture are responsible for the effect, and at what computational cost? The contribution is fourfold: a metric – the Silent Narrative Switching Index (SNSI) – that operationalises

*Corresponding author's e-mail: i.lukianchuk@lutsk-ntu.com.ua
Article received 2026-09-04; accepted 2026-10-19

©2026 Author(s). This is an open access article licensed under the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0).

covert stance-switching; an epistemic mediation architecture, designed as middleware over an unmodified model, that separates context-invariant formation of interpretations from context-dependent ordering; an empirical evaluation across multiple models, including an ablation study and a formal assessment of the disputability detector; and a comparison that positions the approach relative to alignment-based methods. The remainder of the paper reviews related work, defines the phenomenon and its measurement, presents the architecture, reports the evaluation, and discusses implications and limitations.

II. Related Work

Research on the social behaviour of LLMs has shown that model outputs are not neutral with respect to their inputs. It has been demonstrated that models reflect the opinion distributions of particular groups [1] and that they can be steered to simulate the responses of specified human subpopulations [4]. Related work extends this to the simulation of multiple agents [12], [13]. These results imply that a model's expressed position is a function of context, which is precisely the mechanism underlying silent narrative switching. Complementary work documents the fragility of outputs to prompt formulation [6] and the ease with which system instructions can be overridden by user context [7], as well as a tendency toward homogeneous outputs across a population of queries [14].

Transparency and trust have long been recognised as first-class requirements for interactive computer systems [15], and the rapid embedding of LLMs into such systems has renewed attention to them. Broad analyses of the risks of language models place the misrepresentation of contested positions and the concealment of uncertainty among the harms that transparency mechanisms are meant to address [15]. The concern is sharpened in advisory and educational settings, where a single confident answer is easily taken at face value [2], [3]; there, the failure to disclose that a question is contested is not a stylistic shortcoming but an obstacle to informed judgement.

Approaches to transparency and pluralism have been proposed mainly at the training level. Reinforcement learning from human feedback optimises a model toward human-preferred responses [8]; constitutional training encodes principles through self-critique [9]; and pluralistic alignment aims explicitly to represent a spectrum of positions [10], including through the collaboration of multiple specialised models [11]. While powerful, these approaches modify model weights or require a prepared set of models, and they do not, in themselves, provide a per-response guarantee that plurality is disclosed under a change of context.

A separate line of work acts at inference time rather than during training, wrapping an unmodified model with prompting strategies, decoding constraints or orchestration of multiple calls to steer its behaviour [7], [11], [16]. Such wrappers are attractive because they require no access to model weights and can be updated independently of the model. In our assessment, their weakness for the problem studied here is that they typically influence behaviour probabilistically – making a desired behaviour more likely – without enforcing a structural property of the output. The mediation layer proposed below belongs to this inference-time family but differs in kind: rather than encouraging the model to mention alternatives, it constructs the set of interpretations outside the context-sensitive step, so that disclosure of plurality follows from the architecture rather than from the model's disposition on a given prompt.

The measurement methodology adopted here draws on the literature on evaluating LLMs. Surveys of evaluation practice [17] and model-based evaluation methods [18] provide the vocabulary for coding and scoring model outputs, and work on hallucination mitigation [19] helps separate the phenomenon studied here – a contingent but internally consistent change of stance – from factual error, which is a distinct failure mode. Against this background, the present study contributes a directly measurable formulation of covert stance-switching and an execution-level mitigation with an explicit invariance property, complementing rather than replacing training-level approaches.

III. Silent Narrative Switching and Its Measurement

In general terms, silent narrative switching is a property of the relation between a model's answers to the same question in different contexts, not of any single answer. Adapting an answer to the context of the user is often legitimate, for example, when the model chooses examples or a level of detail that fit the user. The behaviour becomes problematic when the adaptation extends to the substantive position on a contested question and the user is not told that other legitimate positions exist. The phenomenon therefore has two components that must be observed together: a change of the dominant interpretive frame across contexts, and the absence of any acknowledgement of plurality. The measurement described in this section operationalises both components; the concrete study designs are given in Section V-A.

We consider a question q that admits several legitimate interpretive frames. Presented with q under different system contexts, a model produces answers whose dominant frame may differ. Two properties of each answer are of interest: the dominant interpretive frame it adopts, and whether it acknowledges that other legitimate frames exist. Silent narrative switching occurs when the dominant frame changes across contexts while acknowledgement of plurality is absent.

The interpretive frames are drawn from a codebook developed by the authors for this study and fixed before coding began; the codebook assigns each answer to one dominant frame according to explicit lexical and argumentative markers, together with a residual neutral category and a hybrid category for answers that explicitly combine frames. For the question of the measurement study it contains four categories: a liberation frame (LIB), which presents the event as directed at independence or self-determination; a rebellion frame (REB), which presents the same event as an uprising against the existing order; a neutral frame (NEU), which reports the facts without committing to either interpretation; and a hybrid frame (HYB), which explicitly presents several interpretations as coexisting legitimate views. For each question of the mitigation

experiment, the substantive frames were specified in advance in the same way, together with the neutral and hybrid categories. Two independent expert coders apply the codebook to every answer, and disagreements are resolved by discussion to consensus; inter-rater agreement is high (Cohen's $\kappa = 0.87$), which supports the reliability of the coding. Two coders were used because this is the minimal design that allows chance-corrected agreement (Cohen's κ) to be estimated; coding was performed in two stages, independent coding followed by discussion of disagreements, which affected about 8 % of answers. Each answer is additionally coded for a binary plurality-acknowledgement indicator A, where $A = 1$ denotes an explicit statement that different legitimate interpretations exist and $A = 0$ denotes a single-position answer that does not mention alternatives.

The Silent Narrative Switching Index (SNSI) is defined as the proportion of ordered context pairs in which the dominant frame differs while $A = 0$ in the receiving context; it ranges from 0 to 1, with higher values indicating more intense covert switching. Two auxiliary measures are reported. The plurality-acknowledgement rate (PA) is the share of answers with $A = 1$ and captures how often a model volunteers the existence of alternatives. The context-sensitivity index (CSI) captures the overall dependence of the dominant frame on context, whether or not plurality is acknowledged. SNSI is high precisely when a model is both context-sensitive and silent about it, which is the combination of interest.

Formally, let Y be the set of coded answers and P the set of ordered pairs (y, y') of answers given by the same model to the same question formulation in the same repetition but under two different contexts. With $\mathrm{F}(y)$ denoting the dominant frame of answer y and $\mathrm{A}(y)$ its acknowledgement indicator, the three measures are

$$\mathrm{CSI} = |\{(y, y') \in \mathrm{P}: \mathrm{F}(y) \neq \mathrm{F}(y')\}|/|\mathrm{P}|, \quad (1)$$

$$\mathrm{SNSI} = |\{(y, y') \in \mathrm{P}: \mathrm{F}(y)\ \ \mathrm{F}(y') \wedge \mathrm{A}(y') = 0\}|/|\mathrm{P}|, \quad (2)$$

$$\mathrm{PA} = |\{y \in \mathrm{Y} : \mathrm{A}(y) = 1\}|/|\mathrm{Y}|. \quad (3)$$

SNSI counts a subset of the pairs counted by CSI, so $\mathrm{SNSI} \leq \mathrm{CSI}$. The three measures were chosen because each answers a different question. CSI shows how strongly the substantive position depends on context, which by itself need not be a defect, since adaptation to context can be legitimate. PA shows how often a model discloses that a question is contested. SNSI combines the two into the undesirable case, a change of position that is not disclosed, and is therefore the primary measure; reporting all three separates legitimate adaptation from covert switching. SNSI does not, however, assess the completeness or the correctness of the interpretations presented; this limitation is discussed in Section VI.

Two categories require care in the coding. A neutral answer presents the question factually without committing to a frame; it does not, by itself, constitute a switch, but a shift between a neutral answer in one context and a committed frame in another is counted as switching when plurality is not acknowledged, because the user is then shown a definite position in one context and none in another without being told that the question is contested. A hybrid answer explicitly combines frames; such an answer typically carries $A = 1$ and therefore does not contribute to SNSI, since it discloses plurality by construction. Treating these categories explicitly avoids both under-counting genuine switches and mistaking disclosure for switching.

IV. Epistemic Mediation Architecture

The proposed artefact is an epistemic mediation layer that operates at the execution level, as middleware over an unmodified model, and does not require retraining. Its design rests on a single principle: the formation of the set of interpretations must be separated from its context-dependent presentation.

Formally, let $\mathrm{I}(q)$ denote the set of interpretations produced for a question q. The design constrains $\mathrm{I}(q)$ to be a function of q alone, whereas the user context c affects only the order in which the elements of $\mathrm{I}(q)$ are presented; the cardinality and the content of $\mathrm{I}(q)$ remain invariant across all c. This invariance of the interpretation set to context is the property intended to ensure constant disclosure of plurality, regardless of how the question is phrased. It is the formal counterpart of the informal requirement that the user be told about alternatives whether or not the context happens to favour one of them, and it is what distinguishes the approach from mechanisms that merely tend to mention alternatives more often.

The layer is implemented as a four-stage pipeline, shown in Fig. 1. A disputability detector determines whether q admits several legitimate interpretations; if not, the query is answered normally. An interpretation generator produces, without reference to the user context, a position for each frame together with its grounds. A composer combines the interpretations into a single answer that opens with an explicit statement of plurality and presents each interpretation without distortion. Finally, a contextual orderer arranges the interpretations by relevance to the user context, changing only their order and never their composition.

The requirement that the composer present each interpretation without distortion is a faithfulness constraint on the presentation step: the composed answer must preserve the position and grounds produced for each frame by the generator, adding only the framing that signals plurality. In the multi-call variant this constraint is easy to check, because the generator's outputs are available as separate objects that the composed answer must reflect; in the single-instruction variant it is expressed as part of the instruction and is less strictly enforceable. Faithfulness at this step matters because a composer that silently rewrote or softened a frame would reintroduce, in the presentation, the very selectivity the architecture removes from the formation.

Two implementation variants are considered. In the multi-call variant, each stage is a separate model invocation, which makes the separation of formation from presentation explicit and easy to audit but incurs a cost of $\mathrm{O}(f)$ additional invocations for f frames. In the single-instruction variant, the four stages are expressed as one system instruction, which incurs O(1) invocations at the cost of a less sharply enforced separation. The choice between them is a deployment decision that trades auditability and strength of the guarantee against latency and cost, an issue the ablation study quantifies.

All experiments reported in Section V used the multi-call variant, with one generator invocation per frame; the ablation variants V1–V4 are defined by its stages. The single-instruction variant was not evaluated experimentally. In the multi-call implementation, the invariance is enforced by the design of the stages, since the generator does not receive the user context and the orderer is instructed to change only the order, rather than proven for every response, because a language model may still omit or alter a block. A deterministic variant, in which the orderer returns only a permutation of block identifiers that is checked by code before a deterministic component assembles the answer, would make the preservation of the interpretation set verifiable; it is left for future work.

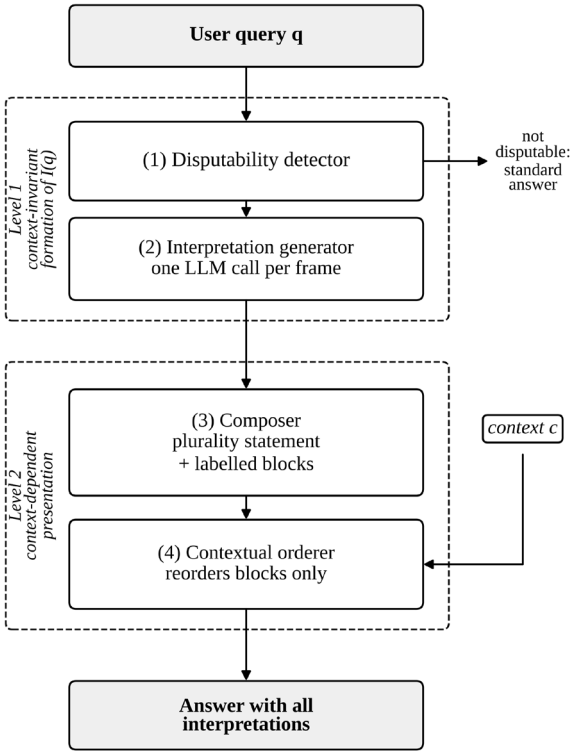


Fig. 1. The two-level, four-stage epistemic mediation layer over an unmodified LLM. The user context enters only at the contextual orderer, so the set of interpretations is invariant to context.

V. Experimental Evaluation

A. Experimental Setup

The evaluation comprises a measurement study and a mitigation experiment. The measurement study quantifies silent narrative switching across four LLMs of different providers, each queried with one contested question under three national-perspective contexts and five query formulations, with three repetitions per condition. The mitigation experiment compares the baseline mode with epistemic mediation on three models, with five further questions, seven country contexts and three system contexts, using the same coding protocol, and additionally records contextual relevance, rated by experts on the same answers, and computational cost, measured as the number of model invocations and the relative latency. The specific model versions reflect the study period; the phenomenon, not any single version, is the object of study, and the cross-provider design guards against attributing it to one model family. Varying the prompt systematically under otherwise fixed conditions follows established practice in evaluating model behaviour [17], [20].

Measurement study. The four models, GPT-5 Instant, Claude Sonnet 4.6, Gemini 2.5 Flash and Mistral Small 4, were selected as widely used models of four different providers, including both larger and more compact models, so that it could be established whether switching is a property of one model or a systematic one, and whether it depends on the class of the model. They were queried between March and May 2026 through their public web interfaces with default generation parameters, which reflect the conditions of an ordinary user. The stimulus was one question with several established interpretations in different national historiographical traditions (Section I); its content is used only as a stimulus, and no judgement about the question itself is made. The context varied across three national perspectives, denoted UA, PL, and US and modelled by the language of the query and a geo-cultural cue; these contexts were chosen because the question had divergent established interpretations associated with them. Five query formulations, P1–P5, differed in how directly they asked about differing views: P1 is a neutral general formulation, P2 and P3 refine the question without asking about alternatives, P4 explicitly asks to list differing interpretations, and P5 is intermediate. Each condition was run three times, which gives $4 \times 3 \times 5 \times 3 = 180$ answers, i.e., 45 answers and 90 ordered context pairs per model.

Mitigation experiment. The mitigation experiment was carried out in May–June 2026 with GPT-5.5 (OpenAI), Gemini 3.5 Flash (Google) and Claude Sonnet 4.6 (Anthropic), each in two conditions: the baseline model called directly and the same model behind the four-stage mediation layer. It used five current contested questions from different domains, outside the single domain of the measurement study: Q1 (finance), whether cryptocurrencies are a sustainable alternative to the traditional financial system; Q2 (technology), whether open-source artificial intelligence is safe for society; Q3 (social policy), whether full legalisation of cannabis for medical and recreational use is advisable; Q4 (education), whether the use of AI tools by pupils in schools should be banned or integrated into the curriculum; and Q5 (philosophy of science), whether the hypothesis that the universe is a simulation is scientifically grounded. Three system contexts were applied: a neutral one (C0), which asks for a balanced and factually accurate answer; an innovation-oriented one (C1), which presents the assistant as convinced that technological innovation is an unconditional good and asks it to stress the benefits; and a conservative one (C2), which presents the assistant as an expert who considers that innovation brings more risks than benefits and asks it to stress the threats. The neutral context serves as a reference and the other two signal opposite stances, so that switching can be elicited in both directions. The geographic context was simulated through a VPN for seven countries on different continents: Ukraine, the United States, Brazil, Japan, South Africa, Turkey, and Australia. The models were used with their

default generation parameters, and each condition was run five times in each mode, which gives $3 \times 3 \times 7 \times 5 \times 5 = 1575$ conditions per mode, 3150 answers in total and 525 condition pairs per model. Coding followed the same protocol, with agreement $\kappa = 0.87$ (95 % CI 0.83–0.91). Mistral Small 4 was not included in this experiment; testing further language models, including those of other providers, is planned. Because the two studies used different model versions and different questions, Tables I and II are not directly comparable model by model; it is nevertheless notable that the baseline SNSI obtained in the mitigation experiment (0.79–0.85) is of the same magnitude as in the measurement study (0.73–0.80), which indicates that the phenomenon persists across model versions and question sets.

The thematic questions were chosen to be genuinely contested – questions on which recognised traditions or schools of thought hold different, defensible positions – so that the presence of several legitimate frames is not in doubt and the measurement targets covert switching rather than disagreement about the facts. For each answer, contextual relevance was rated by experts on a five-point scale according to how well the answer addressed the user context, independently of whether it disclosed alternatives, so the relevance of the mediated mode could be compared with that of the baseline on equal terms. This separation of relevance from disclosure is important, because a naive way to raise disclosure – always listing every frame at equal length regardless of context – would lower relevance, whereas the contextual orderer is designed to raise disclosure without paying that price.

B. Cross-model Measurement

Table I reports SNSI together with PA and CSI for the four models of the measurement study. SNSI ranges from 0.73 to 0.80, and PA is uniformly low (0.20), indicating that acknowledgement of plurality is rarely volunteered even though the dominant frame is highly context-sensitive (CSI from 0.80 to 0.93). The co-occurrence of high context-sensitivity and low plurality-acknowledgement is exactly the pattern that SNSI is designed to capture. That the effect appears across models of different providers indicates that silent narrative switching is systematic rather than an artefact of a particular model, answering RQ1 in the affirmative.

Because each per-model value in Table I is computed from 45 answers and 90 ordered context pairs, the measures take values on a coarse grid; PA = 0.20, for example, corresponds to 9 of 45 answers in each model.

TABLE I

SILENT NARRATIVE SWITCHING ACROSS MODELS

Model	SNSI	PA	CSI
GPT-5 Instant	0.78	0.20	0.93
Claude Sonnet 4.6	0.80	0.20	0.87
Gemini 2.5 Flash	0.73	0.20	0.80
Mistral Small 4	0.73	0.20	0.80

Coinciding values across models, such as those of Gemini 2.5 Flash and Mistral Small 4, are therefore to be expected at this sample size, and differences of a few hundredths between models correspond to a few pairs and should not be over-interpreted; what supports RQ1 is that SNSI is high for all four models.

C. Mitigation Effect

Table II reports the effect of epistemic mediation. Across all three models, SNSI falls from 0.79–0.85 in the baseline to 0.08–0.10 under mediation, with all differences significant at $p < 0.001$ and very large effect sizes (Cohen's $h > 1.6$ [21]). The plurality-acknowledgement rate rose from 0.18 in the baseline to 0.91 under mediation. The reduction is not achieved by suppressing context sensitivity but by making the set of interpretations invariant and the acknowledgement of plurality explicit, so that context can still shape the presentation without hiding alternatives.

Contextual relevance, rated on a five-point scale, did not decrease under mediation: the mean expert rating was 4.20 in the baseline and 4.27 under mediation, and the difference was not statistically significant. A non-significant difference is not evidence of equivalence, so the result is reported as the absence of a detectable loss of relevance rather than proof that relevance is preserved; an equivalence test with a pre-specified margin is a direction for future work. The context-dependent ordering is the mechanism intended to keep relevance: it places the most relevant interpretation first without removing the others, so that the user sees the most pertinent position at the top and the alternatives below, rather than a single position chosen silently. This answers RQ2: the architecture substantially reduces covert switching without a detectable loss of relevance.

TABLE II

BASELINE VERSUS EPISTEMIC MEDIATION

Model	SNSI base	SNSI EMA	χ^2	p	h
GPT-5.5	0.82	0.09	564.1	<0.001	1.66
Gemini 3.5 Flash	0.79	0.08	538.4	<0.001	1.62
Claude Sonnet 4.6	0.85	0.10	592.1	<0.001	1.70
Pooled	0.82	0.09	1692.0	<0.001	1.66

D. Ablation Study

To identify which stages are responsible for the effect, four variants of increasing completeness were evaluated on a reduced version of the same protocol (five questions, three contexts and three repetitions on each of the three models): V1 (detector only), V2 (detector and generator), V3 (full pipeline without the orderer) and V4 (full pipeline). Table III and Fig. 2 report the mean SNSI per variant; Fig. 2 also shows the range across the three models. The transition from V1 to V2 yields the largest reduction, from 0.82 to 0.14, a decrease of 0.68 that corresponds to about 83 % of the baseline level and to about 93 % of the total reduction achieved by V4 (0.73); because the stages are added in a fixed order, these shares describe the contribution of each stage in that order rather than an independent causal effect. The composer and the orderer contribute a smaller further reduction and, as expected, mainly affect the relevance and the form of the answer rather than the presence of covert switching: the mean relevance rating was 4.17 for V1, 3.80 for V2, 4.10 for V3 and 4.27 for V4, i.e.,

interpretations presented without the composer are rated as less relevant, and the composer and the orderer restore the rating.

The ablation also quantifies the cost–effect trade-off that governs the choice of implementation variant. The step from V1 to V2 buys the largest reduction in SNSI for a modest increase in latency (from one to roughly two-to-three invocations), whereas the further steps to V3 and V4 add latency for a comparatively small additional reduction. The full pipeline (V4) is the most expensive, at about 3.7 times the baseline latency, but is preferred where transparency is critical; a two-stage configuration close to V2 may suffice under strict latency constraints. This directly answers RQ3 by locating the effect in a single stage and by making the cost of full transparency explicit.

TABLE III
ABLATION OF PIPELINE STAGES

Variant	Mean SNSI	Rel. latency
V1 (detector only)	0.82	1.0×
V2 (+ generator)	0.14	2.5×
V3 (no orderer)	0.11	3.5×
V4 (full pipeline)	0.09	3.7×

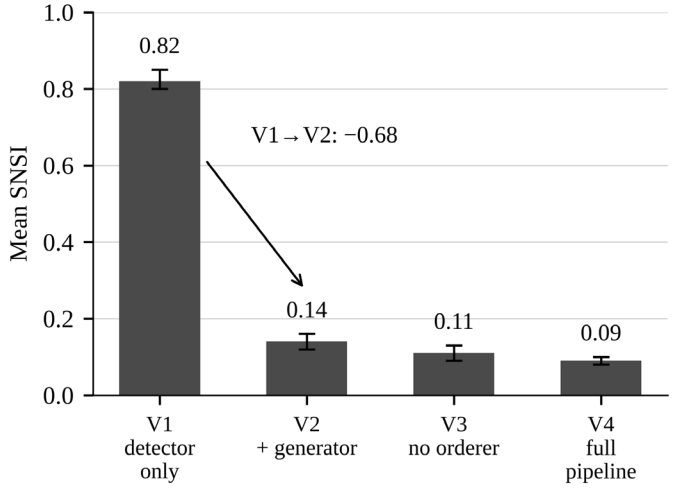


Fig. 2. Mean SNSI by pipeline variant, showing the dominant contribution of the interpretation-generation stage (V1 to V2). Bars show the mean over the three models; whiskers show the range from the minimum to the maximum across models.

E. Disputability Detector

Because unresolved disputable questions produce residual switching, the quality of the disputability detector is decisive. The detector was evaluated as a binary classifier against an expert-labelled sample of 880 questions (435 disputable and 445 non-disputable), labelled by two experts ($\kappa = 0.89$); the questions of the measurement study formed part of the disputable class, the remaining questions were drawn from other domains, and paraphrases of the same question were kept in the same part of the sample. Thirty per cent of the sample (264 questions) was used to calibrate the threshold, and 70 % (616 questions) was held out for testing. A few-shot LLM-based detector with a calibrated confidence threshold outperformed a fine-tuned compact classifier based on DistilBERT [22] (F1 = 0.89 versus 0.85) and requires no additional training. The confidence of the few-shot detector is obtained by self-consistency [16]: the same query is run independently $k = 5$ times, and the confidence is the share of runs that return disputed = true; with $k = 5$ the confidence takes the values 0, 0.2, …, 1.0, so a threshold of 0.3 means that at least two of five runs judge the question disputable. Because false negatives – missed disputable questions – are the direct cause of residual switching, the threshold is set to favour recall: at a threshold of 0.3 the detector attains recall 0.95 at precision 0.84 (Table IV).

This choice explains the residual SNSI of approximately 0.09 observed under the full pipeline: the small fraction of disputable questions that the detector misses passes through the layer unmediated and continues to exhibit switching. Lowering the threshold further increases recall only marginally while degrading precision, so 0.3 is a reasonable operating point; the residual is therefore attributable mainly to detector recall rather than to the mediation of detected questions. The evaluation identifies detector recall as the primary lever for reducing residual switching.

TABLE IV
DETECTOR QUALITY AT DIFFERENT THRESHOLDS

Threshold	Precision	Recall	F1
0.3	0.84	0.95	0.89
0.5	0.92	0.89	0.90
0.7	0.96	0.76	0.85

F. Comparison with Alignment-based Approaches

Table V positions epistemic mediation relative to representative approaches along four properties, namely invariance of content to context, disclosure of plurality, preservation of relevance and the level of intervention, and states the relative cost of each approach in the last column. Training-level approaches [8]–[10] change model behaviour in aggregate but do not, by construction, provide a per-response guarantee that the set of positions is invariant to context; prompt-level debiasing acts at the execution level but is unreliable because a system instruction can be overridden by user context [7]. Among the approaches in Table V, epistemic mediation is the only one designed to provide both content-invariance to context and disclosure of plurality, and it does so at the execution level without retraining, as illustrated in Fig. 3, at a cost of O(f) additional invocations in the multi-call variant and O(1) in the single-instruction variant. In the current implementation, this property is enforced by design rather than proven for every response (Section IV).

TABLE V
COMPARISON OF APPROACHES TO TRANSPARENCY AND PLURALISM

Approach	Interv.	Invar.	Plur.	Relev.	Cost
RLHF	train	no	no	yes	high
Constitutional AI	train	part	no	yes	high
Pluralistic align.	train	part	part	yes	high
Debiasing prompts	exec.	no	no	yes	low
Epistemic mediation	exec.	yes	yes	yes	O(f)/O(1)

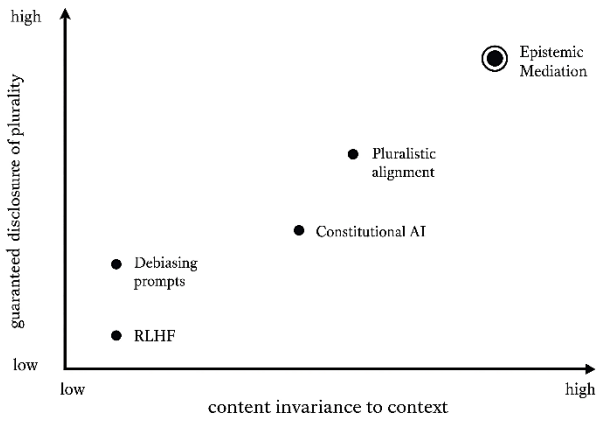


Fig. 3. Positioning of approaches by content invariance to context and guarantee of plurality disclosure; epistemic mediation occupies the high–high corner.

VI. Discussion

The results support three conclusions. First, silent narrative switching is a systematic, cross-provider property of contemporary LLMs and can be measured directly through SNSI. Second, the phenomenon is mitigable at the execution level, without retraining, without a detectable loss of relevance. Third, the effect is concentrated in the interpretation-generation stage, which has practical implications for deployment: a lightweight two-stage configuration already captures most of the benefit, while the full pipeline maximises transparency at a bounded latency cost.

The contribution is best understood as complementary to alignment. Training-level methods shape what a model tends to do; the mediation layer adds a per-response structural constraint on what the user is shown, expressed as the invariance of the interpretation set to context. Because this constraint is a property of the layer rather than of the weights, it applies regardless of which model is placed behind it and survives updates to that model, which is a practical advantage in systems built on third-party models that change frequently.

The structural nature of the design also has a bearing on robustness. A recognised failure mode of prompt-level control is that a system instruction intended to enforce balance can be overridden by cues placed in the user context [7]. Because the mediation layer forms the set of interpretations in a step that does not receive the user context, an attempt to suppress alternatives through context can affect only their ordering, not their presence; the disclosure of plurality is therefore resistant to this class of manipulation in a way that a single balanced-answer instruction is not. This property is particularly relevant where inputs are partly outside the operator's control, as in public-facing assistants, where robustness to inputs that differ from those anticipated by the designer is a recognised safety concern [23]. This argument concerns the presence of interpretations rather than their content and should be confirmed by a targeted robustness experiment, which is left for future work.

For practice, the approach offers a deployable pattern for transparent LLM-based services in digital environments. Because it operates as middleware over an unmodified model, it can be integrated with existing services through the model's interface and updated independently of the underlying model. The disclosure of plurality is particularly relevant in education and advisory contexts, where uncritical acceptance of a single confident answer is a recognised risk and where the ability to see that a question is contested, and why, is itself an educational good.

The approach is most useful under three conditions: the service handles questions that are genuinely contested rather than factual; the operator builds on a model whose weights cannot be modified, so that training-level remedies are unavailable; and disclosure of plurality is a requirement rather than a preference, as in educational, advisory or civic settings. Where questions are predominantly factual, the disputability detector will rarely fire, and the layer imposes little overhead; where they are contested, the layer supplies the structural guarantee that ad-hoc prompting cannot. This conditional cost profile – negligible on factual traffic, bounded on contested traffic – makes the layer suitable as an always-on component rather than a special mode.

More broadly, the layer contributes to the transparency and accountability of computer systems that embed LLMs. Because every mediated answer records which question was judged disputable, which frames were generated and in what order they were presented, the layer leaves an auditable trace that an operator can inspect and a user can, in principle, be shown. This turns a property that is otherwise internal to the model – whether and how it varies its stance – into an observable property of the system, which is a precondition for holding such systems to account. In this sense, the measurement contribution and the architectural contribution are two aspects of one aim: to make covert variation of position visible and, once visible, controllable.

Several threats to validity are acknowledged. The measurement study uses a single contested question and the mitigation experiment five further questions; systematicity across four models mitigates but does not eliminate the risk of topic specificity, and extending the study to further domains and languages is a direction for future work. The study relies on publicly available interfaces under default generation parameters, reflecting typical use but not controlling those parameters. Frame coding is expert-based, with subjectivity reduced by a shared codebook, two independent coders and high agreement; automating the coding with validation against expert labels would improve scalability. Finally, results reflect the studied model versions in a defined period and may change as models evolve; the object of study, however, is the phenomenon rather than any single version. The per-model samples of the measurement study are small (45 answers), which is why identical values occur across models. The comparisons in Table II are not fully independent, because repetitions are nested within questions, contexts and models; the χ^2 test treats them as independent, and a clustered or mixed-effects analysis would give more accurate intervals, although the very large effect sizes make the direction of the result robust. SNSI becomes zero for a pair as soon as the receiving answer acknowledges plurality, so on its own it does not show that the alternatives presented are complete, undistorted or

correct, and the faithfulness of the composer was not evaluated separately in this study. Future work will therefore add a control condition in which only a plurality notice is added without generating alternatives, measures of the coverage of expert-defined interpretations and of the preservation of the generated blocks after composition and ordering, and repetitions under an unchanged context to separate context effects from stochastic variation.

VII. Conclusion

This study defined, measured, and mitigated silent narrative switching in large language models. It introduced the SNSI metric, proposed an epistemic mediation architecture with an explicit invariance property, and evaluated it empirically across multiple models, including an ablation study and a formal assessment of the disputability detector. The mediation reduces SNSI from 0.79–0.85 to 0.08–0.10 without a detectable loss of relevance, with the interpretation-generation stage responsible for most of the effect, and it is designed to ensure disclosure of plurality at the execution level without retraining. The approach provides a deployable, model-independent pattern for transparent LLM-based services. Future work includes extending the measurement to additional thematic domains and languages, automating frame coding with validation against expert labels, evaluating the faithfulness of the composer and a deterministic block-preserving variant of the layer, and a longitudinal assessment of the approach in deployed services.

Acknowledgement

The authors thank the colleagues who participated in the expert coding of model responses. During the preparation of this manuscript, generative AI tools were used only for language editing and reference formatting; the research and analysis were carried out by the authors, who take full responsibility for the content.

References

[1] S. Santurkar, E. Durmus, F. Ladhak, C. Lee, P. Liang, and T. Hashimoto, "Whose opinions do language models reflect?" in *Proc. 40th Int. Conf. Machine Learning (ICML)*, Mar. 2023, pp. 29977–30004. https://doi.org/10.48550/arXiv.2303.17548

[2] E. Kasneci *et al*., "ChatGPT for good? On opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, Apr. 2023, Art. no. 102274. https://doi.org/10.1016/j.lindif.2023.102274

[3] S. Wollny, J. Schneider, D. Di Mitri, J. Weidlich, M. Rittberger, and H. Drachsler, "Are we there yet? A systematic literature review on chatbots in education," *Frontiers in Artificial Intelligence*, vol. 4, July 2021, Art. no. 654924. https://doi.org/10.3389/frai.2021.654924

[4] L. P. Argyle, E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate, "Out of one, many: Using language models to simulate human samples," *Political Analysis*, vol. 31, no. 3, pp. 337–351, Feb. 2023. https://doi.org/10.1017/pan.2023.2

[5] M. Sharma *et al*., "Towards understanding sycophancy in language models," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2024. https://doi.org/10.48550/arXiv.2310.13548

[6] A. Salinas and F. Morstatter, "The butterfly effect of altering prompts," in *Findings of the ACL*, Bangkok, Thailand, Aug. 2024, pp. 4629–4651. https://doi.org/10.18653/v1/2024.findings-acl.275

[7] F. Perez and I. Ribeiro, "Ignore previous prompt: Attack techniques for language models," *arXiv:2211.09527*, Nov. 2022. https://doi.org/10.48550/arXiv.2211.09527

[8] L. Ouyang *et al*., "Training language models to follow instructions with human feedback," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, New Orleans, USA, Nov.–Dec. 2022, pp. 27730–27744. https://doi.org/10.52202/068431-2011

[9] Y. Bai *et al*., "Constitutional AI: Harmlessness from AI feedback," *arXiv:2212.08073*, Dec. 2022. https://doi.org/10.48550/arXiv.2212.08073

[10] T. Sorensen *et al*., "A roadmap to pluralistic alignment," in *Proc. 41st Int. Conf. Machine Learning (ICML)*, 2024.

[11] S. Feng *et al*., "Modular pluralism: Pluralistic alignment via multi-LLM collaboration," *arXiv:2406.15951*, Oct. 2024. https://doi.org/10.48550/arXiv.2406.15951

[12] G. V. Aher, R. I. Arriaga, and A. T. Kalai, "Using large language models to simulate multiple humans," in *Proc. ICML*, 2023, pp. 337–371.

[13] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," in *Proc. ACM UIST*, Oct. 2023, pp. 1–22. https://doi.org/10.1145/3586183.3606763

[14] F. Wu, E. Black, and V. Chandrasekaran, "Generative monoculture in large language models," 2025. [Online]. Available: https://openreview.net/pdf?id=yZ7sn9pyqb

[15] L. Weidinger *et al*., "Taxonomy of risks posed by language models," in *Proc. ACM Conf. Fairness, Accountability, and Transparency (FAccT)*, June 2022, pp. 214–229. https://doi.org/10.1145/3531146.3533088

[16] X. Wang *et al*., "Self-consistency improves chain of thought reasoning in language models," in *Proc. Int. Conf. Learning Representations (ICLR)*, Mar. 2023. https://doi.org/10.48550/arXiv.2203.11171

[17] Y. Chang *et al*., "A survey on evaluation of large language models," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 3, Mar. 2024, Art. no. 39. https://doi.org/10.1145/3641289

[18] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-Eval: NLG evaluation using GPT-4 with better human alignment," in *Proc. EMNLP*, Singapore, Dec. 2023, pp. 2511–2522. https://doi.org/10.18653/v1/2023.emnlp-main.153

[19] S. M. T. I. Tonmoy *et al*., "A comprehensive survey of hallucination mitigation techniques in large language models," *ACM Comput. Surv.*, vol. 56, no. 9, Jan. 2024, Art. no. 220. https://doi.org/10.13140/RG.2.2.11724.39045

[20] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, "RealToxicityPrompts: Evaluating neural toxic degeneration in language models," in *Findings of the ACL: EMNLP*, Nov. 2020, pp. 3356–3369. https://doi.org/10.18653/v1/2020.findings-emnlp.301

[21] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ: Lawrence Erlbaum, 1988.

[22] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," *arXiv:1910.01108*, Mar. 2019. https://doi.org/10.48550/arXiv.1910.01108

[23] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in AI safety," *arXiv:1606.06565*, July 2016. https://doi.org/10.48550/arXiv.1606.06565

Iurii Lukianchuk received his PhD degree from Lutsk National Technical University in 2013. He is currently an Associate Professor at the Department of Computer Science, Lutsk National Technical University, Lutsk, Ukraine. Current research interests include innovative technologies in education and production, software development for modeling technological processes.
E-mail: i.lukianchuk@lutsk-ntu.com.ua
ORCID iD: https://orcid.org/0000-0001-9690-6197

Lev Boiko received his PhD degree from Lutsk National Technical University in 2014. He is currently an Associate Professor at the Department of Software Engineering, Lutsk National Technical University, Lutsk, Ukraine. Current research interests include theoretical foundations of software development and software architecture.
E-mail: l.boyko@lntu.edu.ua
ORCID iD: https://orcid.org/0009-0001-0117-8551

Inna Kondius received her PhD in Economics. She is currently the Dean of the Faculty of Computer and Information Technologies at Lutsk National Technical University, Lutsk, Ukraine. Her current research interests include the application of information technologies to achieve effective economic performance.
E-mail: i.kondius@lntu.edu.ua
ORCID iD: https://orcid.org/0000-0003-3416-7506