*Data & Policy* (2025), 7: e73
doi:10.1017/dap.2025.10041

RESEARCH ARTICLE

# Domain anchorage in LLMs: Lexicon profiling and unintended information leakage

Lekha Challappa, Zijin Zhang and Rajiv Garg

Goizueta Business School, Emory University, Atlanta, GA, USA
**Corresponding author:** Rajiv Garg; Email: rajiv.garg@emory.edu

**Received:** 07 April 2025; **Revised:** 26 August 2025; **Accepted:** 10 September 2025

**Keywords:** AI governance; algorithmic bias; computational linguistics; data leakage; domain anchorage; implicit profiling; large language models; lexicon similarity

## Abstract

This study investigates unintended information flow in large language models (LLMs) by proposing a computational linguistic framework for detecting and analyzing domain anchorage. Domain anchorage is a phenomenon potentially caused by in-context learning or latent "cache" retention of prior inputs, which enables language models to infer and reinforce shared latent concepts across interactions, leading to uniformity in responses that can persist across distinct users or prompts. Using GPT-4 as a case study, our framework systematically quantifies the lexical, syntactic, semantic, and positional similarities between inputs and outputs to detect these domain anchorage effects. We introduce a structured methodology to evaluate the associated risks and highlight the need for robust mitigation strategies. By leveraging domain-aware analysis, this work provides a scalable framework for monitoring information persistence in LLMs, which can inform enterprise guardrails to ensure response consistency, privacy, and safety in real-world deployments.

### Policy Significance Statement

As large language models (LLMs) like GPT-4 are increasingly deployed in enterprise and public-sector settings, concerns about unintended information flow and privacy leakage are becoming more urgent. This study introduces a generalizable framework to detect and manage these risks, offering organizations and policymakers a practical tool for monitoring information persistence across interactions. By implementing domain-aware guardrails informed by this framework, businesses can enhance response consistency, protect user privacy, and reduce the likelihood of biased or inappropriate outputs. These safeguards also support compliance with data protection regulations such as the General Data Protection Regulation (GDPR), contributing to more transparent, accountable, and secure AI governance across industries.

## 1. Introduction

As large language models (LLMs), such as GPT-4, become increasingly integrated into enterprise and institutional workflows, concerns about unregulated usage and unintended information flow have grown more pressing. A recent survey revealed that 68% of employees using ChatGPT at work do not disclose their usage to employers[1], a trend that is mirrored across other LLM platforms in professional settings.

[1] https://www.businessinsider.com/70-of-people-using-chatgpt-at-work-havent-told-bosses-2023-3

This integration in domains such as healthcare introduces substantial ethical and privacy concerns. For instance, in healthcare, AI's ability to process and analyze vast datasets with precision raises significant issues regarding patient privacy and data security (Williamson and Prybutok, 2024). The application of AI in corporate settings further necessitates a thorough examination of how sensitive information is managed and protected (Sargiotis, 2024).

This surge in informal, unmonitored adoption highlights a critical governance challenge: these models may be exposed to sensitive corporate, financial, or healthcare data without appropriate oversight or safeguards. This trend raises a pressing question: could LLMs, when repeatedly exposed to domain-specific input, inadvertently carry information from one interaction to another, resulting in information leakage or cross-user bias? Real-world incidents, including documented cases of AI models unintentionally revealing snippets of sensitive user data, reinforce the urgency of this inquiry and the need for robust detection methods.

To investigate this risk, we simulate multiuser interactions with an LLM in a controlled setting, analyzing how domain-specific prompting may influence subsequent outputs. We design two lexicon profiles, Profile A and Profile B, that represent semantically equivalent queries phrased using different vocabulary, mimicking distinct users operating within the same domain. By comparing responses generated for these divergent input styles, we aim to measure the degree to which the model's answers remain consistent, regardless of prompt variation, once it is anchored to a particular domain. While GPT-4 is used as the empirical testbed, our simulation approach is generalizable to other LLMs with similar architecture and in-context learning (ICL) capabilities. The focus of our framework is to quantify similarities between prompts and model responses across multiple linguistic dimensions, such as lexical, syntactic, semantic, and positional, to detect when outputs become disproportionately influenced by a prior domain context.

Our theoretical grounding builds on recent work interpreting ICL as a form of implicit Bayesian inference, where the model learns to infer shared latent concepts from prompt examples without explicit parameter updates. We extend this view by proposing that domain-specific prompts create a latent "cache" of concepts that persistently influence output generation. This effect, which we term domain anchorage, is not limited to GPT-4, but is likely a general characteristic of modern LLMs capable of rapid adaptation. While such behavior may improve answer consistency, it also carries risk, as domain anchorage can lead to unintended reinforcement of user-specific information, introduce bias or echo-chamber effects, and increase the likelihood of inadvertent cross-user information transfer.

The potential for domain anchorage to undermine fairness, safety, and data privacy underscores the importance of explainability and transparency in AI systems. In response, we propose a computational linguistic framework to detect and quantify domain anchorage as a proxy for information persistence in LLMs. This framework offers a generalizable method for auditing model behavior and assessing unintended information flow. Our contribution is threefold: first, we introduce a structured, model-agnostic approach to analyze domain-specific consistency in LLM responses; second, we demonstrate how this framework can serve as the foundation for implementing domain-aware guardrails that enhance privacy and reduce information leakage.

Beyond risk detection and governance, a critical yet often overlooked dimension of LLM deployment in enterprise settings is response consistency. Organizations increasingly rely on LLMs to support workflows that demand reliable, repeatable, and verifiable outputs to draft regulatory documents, generate customer responses, or assist with internal decision support. In these contexts, inconsistent answers to semantically equivalent queries can erode trust in the system, introduce operational inefficiencies, and complicate compliance reporting. Our findings offer a mechanism to manage such inconsistencies by identifying patterns of domain anchorage that influence how models generalize across similar inputs. By leveraging the proposed framework, enterprises can establish domain-aware prompt strategies or monitoring protocols that reinforce consistency without requiring full model retraining or fine-tuning. This contribution is especially valuable for industries like finance, healthcare, and law, where decision-making must be transparent and defensible. In this way, our framework not only identifies potential risks, but also offers a practical path to strengthening the reliability and usability of LLMs within closed enterprise environments.

https://doi.org/10.1017/dap.2025.10041 Published online by Cambridge University Press

This research supports broader efforts to institutionalize transparency, fairness, and data protection within AI governance. As highlighted in the UK Government's AI White Paper[2], explainability is central to maintaining public trust in automated systems. However, recent findings from the Centre for Data Ethics and Innovation (2021) reveal a significant public knowledge gap around how personal data is processed. Our framework helps bridge that gap by making latent behaviors in language models observable and measurable, thereby supporting compliance with principles outlined in data protection regulations such as the GDPR and the UK Data Protection Act. Furthermore, this paper directly responds to the call for rigorous, interdisciplinary methodologies that enable responsible and transparent AI deployment in policy-relevant domains (Engin et al., 2024). By focusing on domain anchorage as a mechanism of unintended information persistence in LLMs, we contribute to the broader AI policy agenda aimed at operationalizing trust, safety, and oversight in algorithmically mediated decision environments. By equipping organizations with tools to evaluate and control unintended information flow in LLMs, we offer a research-backed pathway to align AI system design with evolving legal and ethical standards for responsible data use.

To situate our work within the broader landscape, we next review literature on ICL, semantic persistence, and information leakage in LLMs, identifying where current approaches fall short in detecting domain anchorage.

## 2. Literature review

Recent advancements in language models such as GPT-4 have demonstrated significant capabilities, including ICL. Xie et al. (2022) describe ICL as an implicit Bayesian inference process, wherein language models, during pretraining, learn to infer shared latent concepts across varied contexts. This capability enables models to generate accurate responses even when presented with novel prompts. For example, the model may learn to associate named entities with attributes such as nationality or occupation, despite not being explicitly trained on such pairings.

In addition to concerns surrounding ICL, Hermann et al. (2024) provide complementary evidence from the protein domain, showing that models trained without pretraining-aware splits can exhibit substantial leakage, leading to inflated performance due to overlap between training and evaluation data. These findings reinforce our concern that transformer outputs may reflect latent patterns shaped by prior exposure, even in the absence of explicit memory or fine-tuning.

The response-generation process involves the model implicitly learning a mapping between input and output, regardless of the artificial concatenation of examples during prompting. By recognizing how language models infer shared latent concepts from domain-specific prompts, we can better understand the mechanisms contributing to response uniformity. Our study posits that domain-specific priming introduces a form of meta-optimization, where the model's transformer layers adapt to reflect domain context, producing responses that exhibit significant uniformity regardless of lexical variation in the input. The relevance of Xie et al. (2022)'s findings to our work lies in the potential for this anchoring to introduce systematic bias and privacy risk. By comparing responses generated across distinct lexicon profiles within the same domain, we evaluate the extent to which domain-specific prompts influence the model's attention and embedding layers. This assessment is central to understanding how proprietary or sensitive information might be inadvertently shared through model outputs as a result of persistent domain anchorage.

### *2.1. Meta-optimization as a driver of domain anchorage*

In the context of meta-optimization as a driver of domain anchorage, Dai et al. (2023) investigate the mechanisms behind GPT models' ICL abilities, proposing that these models implicitly perform gradient descent as meta-optimizers. This process involves adjusting attention weights in response to input data, akin to real-time gradient updates, allowing models to adapt to new tasks without explicit fine-tuning. The

[2] https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper

https://doi.org/10.1017/dap.2025.10041 Published online by Cambridge University Press

authors highlight that transformers' attention mechanisms can be viewed as critical components in this optimization process. Their analysis reveals that attention-based adaptation enables the model to optimize for the training loss within the prompt context, effectively treating each prompt as a mini-task to be solved via implicit meta-learning. Notably, this process does not require parameter updates; instead, the model simulates the effect of gradient descent through its internal structure. The authors demonstrate that this meta-optimization behavior allows GPT models to generalize beyond the specific examples provided in the prompt by leveraging the inductive bias learned during pretraining.

This framework offers valuable insight into the underlying mechanisms that enable GPT-4 and similar models to rapidly build contextual specificity. It directly informs our understanding of domain anchorage, wherein repeated exposure to semantically consistent prompts leads the model to internalize a domain-specific prior. As the model adapts to new contexts through in-context gradient-like mechanisms, it may exhibit output behaviors that disproportionately reflect earlier domain prompts. This aligns with our hypothesis that transformer-based LLMs can form latent caches of conceptual associations, influencing future completions even across semantically distinct queries.

By connecting the phenomenon of domain anchorage to the internal meta-optimization processes (Dai et al., 2023), we strengthen the theoretical grounding of our leakage detection framework. Understanding the role of attention-weight adaptation in shaping model outputs over successive interactions also highlights the importance of robust auditing tools to identify when such contextual adaptations begin to compromise fairness, privacy, or output variability.

### *2.2. Extractable memorization*

Recent empirical work demonstrates that LLMs are capable of extracting memorized content from their training data under specific prompting conditions (Nasr et al., 2025). The study identifies cases in which production-grade models, including ChatGPT, reproduced long sequences from pretraining data, including copyrighted text and personally identifiable information. The authors were able to extract over 10,000 examples from ChatGPT's training dataset, revealing the model's susceptibility to emitting verbatim text from its training corpus. These responses were often elicited by prompts that either repeated high-frequency tokens or introduced subtle domain-specific patterns. The findings highlight the presence of persistent internal representations that can be reactivated through contextual cues. This behavior supports the hypothesis that repeated exposure to domain-specific prompts can influence the model's internal state in a manner consistent with what we define as domain anchorage.

The existence of extractable memorization underscores the importance of proactive safeguards in enterprise and regulated environments, particularly when models are exposed to sensitive input streams. The leakage-detection framework proposed in this study offers a structured method to monitor and flag output similarity patterns indicative of such memorization, contributing to the mitigation of privacy risk and model governance.

### *2.3. Bias and stability in language models*

Dentella et al. (2023) systematically evaluated the performance of ChatGPT and other LLMs on grammaticality judgment tasks. The findings revealed that ChatGPT exhibited variable accuracy, performing above chance on grammatical sentences but below chance on ungrammatical ones. Additionally, the model demonstrated significant response instability and a pronounced "yes bias" toward affirming that sentences were grammatically correct, regardless of their actual correctness. Complementing this linguistic analysis, Schmidt et al. (2025) conducted a domain-specific evaluation comparing ChatGPT's diagnostic performance to that of 265 medical residents. The study found that ChatGPT's diagnostic decisions were susceptible to contextual bias, particularly when distractor symptoms or prior case patterns were introduced. These findings indicate that LLMs may internalize and reinforce domain-specific priors in ways that mirror human diagnostic bias. This behavior aligns with our broader claim that repeated exposure to a specific domain can lead to persistent internal representations, a phenomenon that we define

https://doi.org/10.1017/dap.2025.10041 Published online by Cambridge University Press

as domain anchorage. Taken together, these studies underscore the challenges of ensuring consistency and reliability in LLM outputs, particularly in high-stakes domains such as healthcare, where contextual drift or anchoring effects may compromise safety and trustworthiness.

Building on these foundations, we designed a controlled experimental framework to systematically detect and quantify domain anchorage. The following section outlines our methodological design, integrating computational linguistics with embedding-based similarity measures.

## 3. Methodology

Our methodological approach builds upon established theoretical innovations in transformer-based LLMs, particularly decoder-only architectures exemplified by GPT-4. These models leverage multihead self-attention mechanisms combined with positional embeddings to effectively encode complex linguistic relationships and contextual nuances. Recent scholarship frames ICL within transformers as implicit Bayesian inference, wherein models infer latent conceptual structures from contextual examples. Extending this theoretical perspective, we hypothesize the existence of *domain anchorage*, a latent alignment phenomenon emerging when GPT-4 is primed with domain-specific prompts. Domain anchorage may result in notably uniform model outputs across varied lexical inputs, suggesting potential biases or unintended information reinforcement.

To rigorously investigate this phenomenon, our computational linguistic framework quantifies linguistic features of prompts and model-generated responses. We design prompts following established prompt-based learning principles, which emphasize controlled template structures to guide model behavior (Liu et al., 2023). Specifically, we employ token-limited interrogative prompts to minimize confounding variability while isolating domain anchorage effects. These prompts maintain consistent syntactic framing across domains, varying only in the domain-specific prime, enabling systematic measurement of lexical, semantic, syntactic, and positional similarities. Semantic proximity is represented using embedding-based techniques, including averaged embeddings, deep averaging networks (DANs), and LSTM-based embeddings. Lexical analysis assesses domain-specific vocabulary frequency distributions; syntactic analysis uses hierarchical grammatical representations informed by Chomsky's generative grammar (X-bar theory); semantic analysis evaluates meaning-based relationships between words; and positional analysis examines word order effects via transformer-based positional embeddings. Together, these dimensions provide a comprehensive quantitative basis for evaluating how domain-specific contexts influence GPT-4 linguistic behaviors and for assessing the implications of domain anchorage in responsible LLM deployment.

Our empirical approach (Figure 1) involves constructing pairs of lexicon profiles, denoted as Profiles A and B, that express identical semantic intent through varied vocabulary and linguistic structures, simulating realistic user variation. The computational linguistic transformations applied convert qualitative textual data into quantitative metrics, allowing us to systematically measure input–output similarity. Inputs ($x$) and outputs ($y$) are modeled mathematically as transformations $y = f(x)$, where the function $f$ represents the transformer architecture with multi-head attention (MHA) and ICL of the LLM. By analyzing the similarity between input profiles and generated responses across linguistic dimensions, we identify cases where output generation is disproportionately anchored to previously introduced domain contexts.

Our theoretical foundation integrates insights from computational linguistics, particularly semantic similarity and linguistic profiling. Following extant literature (Chandrasekaran and Mago, 2021), we quantify lexical similarity through term frequency metrics, syntactic similarity using hierarchical grammatical representations based on Chomsky's X-bar theory (Chomsky, 1972), semantic similarity via cosine distances of sentence embeddings (Mihalcea et al., 2006; Cer et al., 2017), and positional similarity using transformer positional encodings (Vaswani et al., 2017). Thus, our computational linguistic framework assesses how transformer-based LLMs like GPT-4 exhibit domain anchorage—where prior exposure to domain-specific prompts influences subsequent model responses. We quantify linguistic variations by focusing on four distinct dimensions: lexical, semantic, syntactic, and positional similarity.

https://doi.org/10.1017/dap.2025.10041 Published online by Cambridge University Press

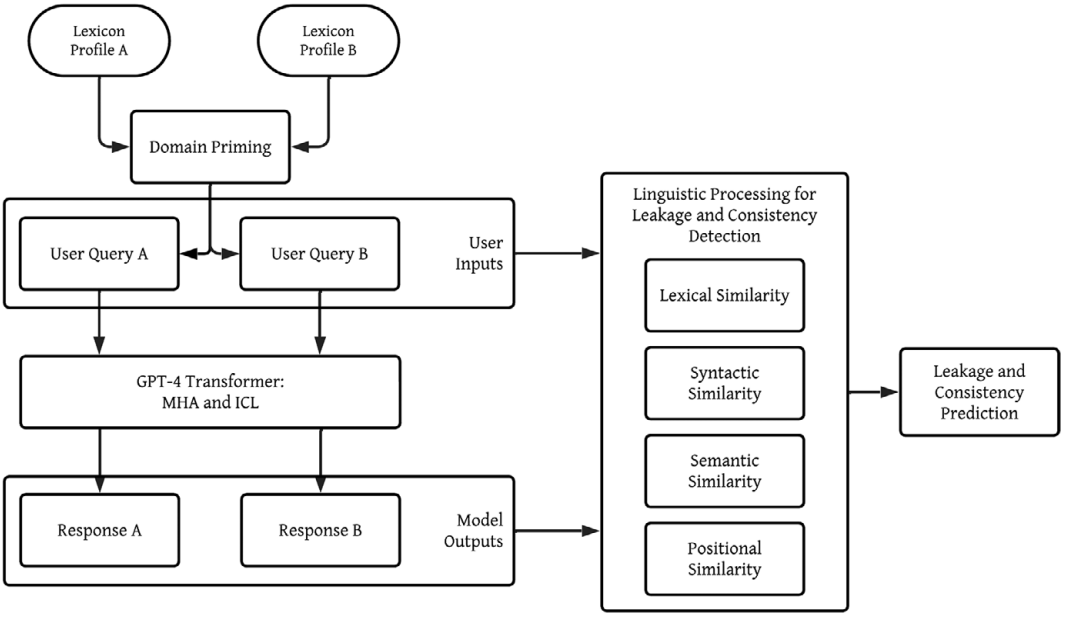


***Figure 1.*** *Flow of lexicon profiles A and B through LLM.*

Lexical similarity measures the overlap between sets of vocabulary terms used in prompts, providing insights into how specific domain terminology influences model behavior. Formally, it is computed as the intersection over the union of two distinct vocabulary sets ($W_1$ and $W_2$):

$$L(W_1, W_2) = \frac{|W_1 \cap W_2|}{|W_1 \cup W_2|}$$

Semantic similarity evaluates the contextual meanings and associations between words or sentences. Using embedding-based representations ($\vec{q}_1$ and $\vec{q}_2$), we calculate the cosine similarity between prompts or responses, quantifying how closely their meanings align:

$$S\left(\vec{q}_1, \vec{q}_2\right) = \frac{\vec{q}_1 \cdot \vec{q}_2}{\|\vec{q}_1\| \|\vec{q}_2\|}$$

Syntactic similarity analyzes grammatical structures, drawing on Chomsky's generative grammar framework to construct hierarchical syntactic representations. These representations, where $\vec{S}(q)$ represents the syntactic structure vector of query $q$, can help assess the structural alignment between different linguistic inputs:

$$\text{Sim}_{\text{syntax}}\left(\vec{q}_1, \vec{q}_2\right) = \exp\left(-\|\vec{S}(q_1) - \vec{S}(q_2)\|\right)$$

Positional similarity examines how the order of words within sequences impacts model responses. Utilizing positional encodings, we quantify how variations in word placement affect GPT-4's attention mechanisms and output generation:

$$\text{Sim}_{\text{positional}}\left(\vec{q}_1, \vec{q}_2\right) = \frac{1}{n}\sum_{i=1}^{n} \frac{\vec{P}_i(q_1) \cdot \vec{P}_i(q_2)}{\|\vec{P}_i(q_1)\| \|\vec{P}_i(q_2)\|}$$

https://doi.org/10.1017/dap.2025.10041 Published online by Cambridge University Press

where $\vec{P}_i(q)$ represents the positional encoding vector of the $i$-th token in query $q$, and $n$ is the number of tokens.

By integrating these four dimensions, our framework provides a comprehensive quantitative measure of linguistic similarity. Specifically, we model prompts (human inputs) and model-generated outputs mathematically to evaluate their alignment across these dimensions. The resulting composite similarity metric enables us to systematically detect, quantify, and analyze domain anchorage and potential information leakage within LLM responses. Formally, the total similarity between two textual representations is expressed as (Figure 2):

$$\mathrm{Sim}\left(\vec{q}_1, \vec{q}_2\right) = \alpha S_{\text{semantic}}\left(\vec{q}_1, \vec{q}_2\right) + \beta S_{\text{syntactic}}\left(\vec{q}_1, \vec{q}_2\right) + \gamma S_{\text{positional}}\left(\vec{q}_1, \vec{q}_2\right) \tag{3.1}$$

where $\alpha$, $\beta$, and $\gamma$ are weighting parameters to balance linguistic contributions. This composite metric robustly captures nuanced linguistic variations influencing LLM response behaviors.

Our experimental design evaluates the GPT-4 model across five representative corporate domains (information technology, finance, software, healthcare, and entertainment) selected due to their high usage of LLMs. For each domain, distinct domain-specific primes (e.g., "Act as a Healthcare specialist") were introduced to half of the simulated clients, establishing a comparative baseline with unprimed controls. Each client engaged in a standardized sequence of interrogative prompts with controlled lexical, semantic, and syntactic variations. Queries were strategically constructed to ensure consistent positional and syntactic structures while systematically varying semantic and lexical content, targeting a similarity range reflective of typical linguistic variations in professional contexts ($0.4 < \text{similarity} < 0.5$). Simulations utilized OpenAI's GPT-4 API with uniform parameters (e.g., max tokens: 100, temperature: 1.5, top-p: 0.5) to ensure reproducibility and comparability across interactions.

Our methodology, while robust, has several acknowledged limitations. First, the specificity of intent in token-limited prompts restricts generalization; future research should incorporate broader prompt diversity and intent coverage. Second, while our computational linguistic metrics provide insightful quantification, they may overlook nuanced linguistic subtleties, suggesting complementary human qualitative analyses in subsequent studies. Third, despite leveraging detailed insights from transformer architectures, the exact internal computational mechanisms remain proprietary and opaque, limiting definitive causal claims about the internal states of GPT-4. Finally, our approach measures semantic similarity as a measure of response consistency and leakage rather than direct information dissemination, indicating a need for further empirical validation to directly establish cross-user information leakage.

Despite these limitations, our computational linguistic methodology provides a valuable, scalable framework to audit and mitigate domain anchorage effects in enterprise-grade LLM implementations. It offers tangible guidelines for proactive domain-aware interventions, enhancing model transparency, reliability, and consistency—critical prerequisites for trustworthy AI integration in high-stakes environments.

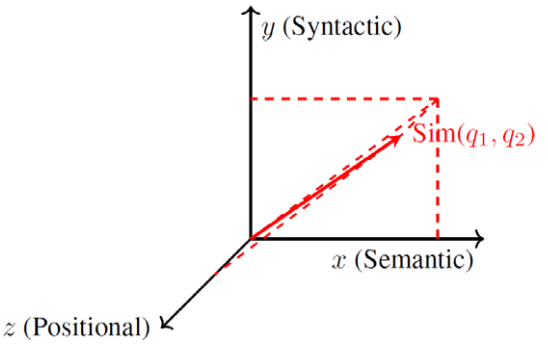


***Figure 2.*** *Vector representation of response similarity.*

https://doi.org/10.1017/dap.2025.10041 Published online by Cambridge University Press

### *3.1. Experimental design*

Our empirical simulation focuses on five corporate domains selected due to their widespread adoption of LLMs in professional contexts: Information Technology ($X_1$), Finance ($X_2$), Software ($X_3$), Healthcare ($X_4$), and Entertainment ($X_5$). To assess the impact of domain-specific contexts on GPT-4's responses, we established standardized domain-specific primes (Table 1) to simulate realistic professional scenarios.

To control variables in our priming procedure, we implemented measures to standardize and differentiate our primes. Positional similarity was controlled by maintaining identical sentence structures and word orders across all primes, ensuring a positional similarity score of 1. Intent similarity was also controlled, as all primes aimed to direct the model to act as a specialist within a specific domain, thus maintaining an intent similarity score of 1. Syntactic similarity was preserved by using nouns to represent each domain, providing consistent syntactic structures. However, differentiation was introduced through lexical and semantic variations, as the domain-specific terms (e.g., "Information Technology," "Finance") were varied. This differentiation in terms and their associated meanings allowed us to build preliminary contextual specificity and effectively assess domain anchorage.

To establish a benchmark for linguistic variance among employees working within the same domain, we aimed to achieve a similarity score within the range of $0.4 < x < 0.5$. This benchmark reflects the balance between maintaining domain-specific relevance and introducing sufficient linguistic variability. By targeting this similarity range, we ensured that the lexical and semantic differences were representative of natural linguistic variations, while preserving the contextual integrity required to assess domain anchorage accurately.

To systematically evaluate the influence of domain-specific priming on GPT-4's responses, simulations were conducted using the OpenAI API client with uniform parameters (Table 2). This experimental setup allowed independent handling of each query, ensuring unbiased evaluations of the model's behavior. Our design included a total of 200 simulated clients divided evenly into two groups: 100 primed

***Table 1.*** *Domain-specific primes for lexicon profiles A and B*

| Domain | Prime |
|---|---|
| $X_1$ | Act as an Information Technology specialist |
| $X_2$ | Act as a Finance specialist |
| $X_3$ | Act as a Software specialist |
| $X_4$ | Act as a Healthcare specialist |
| $X_5$ | Act as an Entertainment specialist |

***Table 2.*** *GPT-4 client parameters*

| Parameter | Value |
|---|---|
| Model | gpt–4-turbo |
| Max Tokens | 100 |
| Post-processing | Limit to 5 tokens |
| Temperature | 1.5 |
| Top P (Nucleus Sampling) | 0.5 |
| Frequency Penalty | 0.1 |
| Presence Penalty | 0.1 |
| Total No. of Prompts | 10,000 |
| Total No. of Responses | 10,000 |
| Total No. of Tokens (Regex) | 101,360 |

https://doi.org/10.1017/dap.2025.10041 Published online by Cambridge University Press

clients, which received domain-specific prompts, and 100 unprimed clients, which did not receive any priming context.

Each group consisted of 20 clients for each domain, with 10 clients representing Lexicon Profile A and 10 clients representing Lexicon Profile B. This setup allowed us to compare primed and unprimed responses across various domains. The chain of queries was repeated 10 times for each client to build a substantial analytical case. The only difference between the primed and unprimed clients was the presence of a domain-specific prime. The intent across client profiles and domains remained consistent: to learn about the domain in question. By constraining the responses to 5 tokens and removing punctuation prior to tokenization (i.e., counting only alphanumeric "word" tokens and excluding nonword characters), we ensured a focused and manageable response data set, enabling thorough analysis of the impact of domain-specific primes on the model's responses.

After establishing the domain-specific prime for each client, we proceeded with five subsequent interrogative prompts to assess the model's response dynamics. Each client received a sequence of these prompts, carefully designed to vary the positional, semantic, and syntactic similarity by approximately 0.417 across Lexicon Profiles A and B, after repeated attempts to achieve the predetermined similarity benchmark. This variation ensured that while the prompts maintained a degree of consistency in their intent and structure, they introduced sufficient linguistic variability to simulate realistic usage scenarios.

By operationalizing domain anchorage across lexical, syntactic, semantic, and positional dimensions, this design enables empirical testing of our hypotheses, as presented in the following results

## 4. Results

At the end of the investigation, all 100 primed and 100 unprimed clients were successfully evaluated over a period of two months. To ensure a consistent analytical procedure, clients within the same domain were evaluated within the same period. The results are separated into two main subsections: (1) sequential pairwise variability—the similarity of responses within a Lexicon Profile (A or B); and (2) prime-presence variability—the response similarity differential between the 100 primed and 100 unprimed clients.

### *4.1. Sequential pairwise variability*

Sequential pairwise variability measures the consistency of the model's responses when presented with prompts from the same domain and lexicon profile. High similarity scores in this analysis indicate that the model is retaining and reapplying contextual cues across sequential interactions, a behavior relevant to our domain anchorage hypothesis.

The results indicate a high degree of consistency in the model's responses within each domain. The average sequential pairwise similarity for Lexicon Profile A was 0.768. The average sequential pairwise similarity for Lexicon Profile B was 0.814. The average similarity scores for Lexicon Profile B were approximately 6% higher than Lexicon Profile A. Notably, the lowest similarity was exhibited by Domain $X_3$ for Lexicon Profile A (Figure 3).

### *4.2. Prime-presence variability*

Prime-presence variability assesses how strongly domain primes stabilize model outputs beyond surface form differences. The results (Table 3) collected by computing a weighted similarity between each response to queries with the same client-side iteration indicated that all primed domains exhibited a weighted similarity greater than 0.80 between responses generated by GPT-4 for Lexicon Profiles A and B. Notably, each query maintained consistent intent, with variations only in the linguistic elements. This consistency in intent allows us to test the relationship between domain primes and linguistic variants on the responses generated by the transformational architecture. Across all five domains, cross-profile response similarity remained high with low dispersion (SD = 0.026–0.038), indicating that once a domain context is set, variation between Lexicon Profiles A and B contributes minimally to output variance. By contrast,

https://doi.org/10.1017/dap.2025.10041 Published online by Cambridge University Press

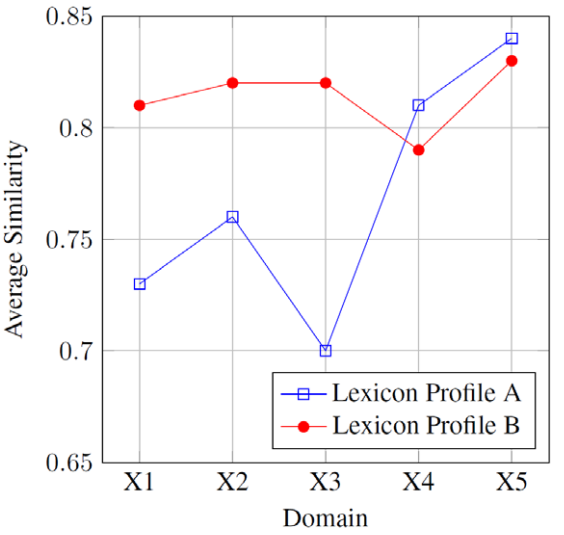


***Figure 3.*** *Sequential pairwise response similarity results across primed prompts.*

***Table 3.*** *Domain-specific response similarity between Lexicon profiles A and B*

| Domain | Primed response similarity | Unprimed response similarity | Paired t-test |
|---|---|---|---|
| Information Technology ($X_1$) | 0.901 (0.038) | 0.516 (0.044) | $t(9) = 52.45, p < 0.001$ |
| Finance ($X_2$) | 0.909 (0.036) | 0.507 (0.040) | $t(9) = 54.15, p < 0.001$ |
| Software Engineering ($X_3$) | 0.903 (0.034) | 0.513 (0.033) | $t(9) = 46.99, p < 0.001$ |
| Healthcare ($X_4$) | 0.812 (0.026) | 0.516 (0.031) | $t(9) = 28.03, p < 0.001$ |
| Entertainment ($X_5$) | 0.897 (0.033) | 0.534 (0.033) | $t(9) = 40.56, p < 0.001$ |

cross-profile similarity in the unprimed condition was markedly lower within the same domains, indicating that the observed uniformity stems from the domain prime rather than baseline lexical overlap.

The paired t-tests reveal that, for the same prompt iteration, adding a domain prime increases cross-profile similarity significantly and consistently ($p < 0.001$ in all cases), substantially anchoring model outputs to the domain context compared to the residual variability introduced by lexicon profile differences. Within our controlled, token-limited setting, the presence of a prime is not only statistically significant but also practically dominant relative to residual variability from profile wording.

These results are consistent with ICL as implicit Bayesian inference. LLMs operate over a large parameter space that can generate diverse outputs, and domain-specific priming narrows this space toward token patterns aligned with the primed domain, reducing variability. In multiuser settings, this narrowing allows information introduced by one user to influence another's outputs when similar anchoring cues are present. This is analogous to two users issuing highly specific search queries and arriving at the same rare web page, except that in an LLM, the "page" is generated dynamically from prior context, making leakage less visible but potentially more consequential, which motivates guardrails against cross-user influence.

## 5. Discussion

The empirical results of this study provide preliminary evidence to support our hypothesis of domain anchorage in GPT-4. The high degree of similarity observed in sequential responses within the same

https://doi.org/10.1017/dap.2025.10041 Published online by Cambridge University Press

lexicon profile, exceeding a mean score of 0.76, suggests that the model's outputs are influenced by the domain-specific context established by the initial prompt. This effect is reinforced by cross-lexicon response similarity, which remained above 0.80 across all tested domains, indicating that the model aligns more closely with domain context than with individual lexical variations.

The most striking evidence of domain anchorage emerged from the prime-presence variability analysis. The observed increase in response similarity between primed and unprimed clients, ranging from 57.36% to 80.56%, suggests a significant effect of domain-specific priming on GPT-4's responses. This finding suggests that the model is not merely adapting to the general linguistic patterns of the input but may be actively anchoring its responses to the specific domain knowledge provided in the prime, potentially leading to intra-domain information dissemination. While these results are consistent with our anchorage hypothesis, further quantitative analysis is required to address previously outlined limitations to strengthen inferential claims.

These findings invite further consideration of how domain anchoring interacts with model behavior in practical settings. From an interpretability standpoint, these patterns suggest that GPT-4 retains and reuses domain-specific contextual cues in ways that systematically shape its outputs (Appendix A). This behavior has implications for information dissemination: while domain anchoring can improve coherence and relevance in specialized contexts, it also risks narrowing the diversity of responses and reinforcing domain-specific biases. Potential mitigation strategies could include introducing cross-domain variation during prompt design, interleaving neutral context, or deliberately varying lexical framing to reduce over-anchoring effects. Additionally, adaptive prompting techniques—where the model is periodically reset with balanced, multidomain context—could help maintain responsiveness to new inputs without losing domain coherence. Exploring these approaches systematically would help establish practical safeguards against information siloing, enabling more balanced and representative outputs across varied application domains.

### *5.1. Mitigating organizational risks and managerial implications*

Our findings underscore the potential risks associated with domain anchorage in organizational settings, particularly when sensitive data is involved. We propose a comprehensive deployment and mitigation strategy grounded in our computational linguistic framework, with implications spanning enterprise-level risk management, infrastructure design, and prompt engineering interventions.

For internal LLM deployments handling sensitive data, enterprises should prioritize robust protections, including private infrastructure and explicit disabling of caching and ICL to prevent inadvertent information persistence or leakage (Figure 4). Conversely, in less sensitive settings, organizations may tolerate reduced data security to optimize for response consistency and decision-making efficiency. In public-facing deployments—especially those involving sensitive user inputs—localized Smaller Language Models (SLMs) and strong guardrails, for input/output filtering, memory controls, and monitoring, must be employed to prevent cross-user information transfer and semantic entrenchment. Even when data is nonsensitive, public LLM platforms should still implement lightweight safeguards to ensure output integrity and avoid unintended dissemination.

To operationalize these principles, we propose a three-step approach to organizational risk management: First, organizations should construct lexicon profile databases for employees, capturing the lexical, semantic, syntactic, and positional characteristics of user–model interactions. These profiles act as linguistic feature maps and support periodic audits to identify anomalies or consistency breaches. Second, enterprises should implement systematic assessments of prompt-response similarity metrics across users. Persistently high similarity values could indicate excessive anchoring to domain-specific memory, warranting contextual intervention or prompt restructuring. Third, de-anchoring strategies should be developed through contextual prompt engineering. Techniques such as prompt randomization, controlled lexical variation, and dynamic context switching can reduce the influence of semantically aligned keywords and promote response diversity. Collectively, these recommendations offer a pathway to safeguard LLM deployments across a variety of operational contexts while preserving transparency, consistency, and security.

https://doi.org/10.1017/dap.2025.10041 Published online by Cambridge University Press

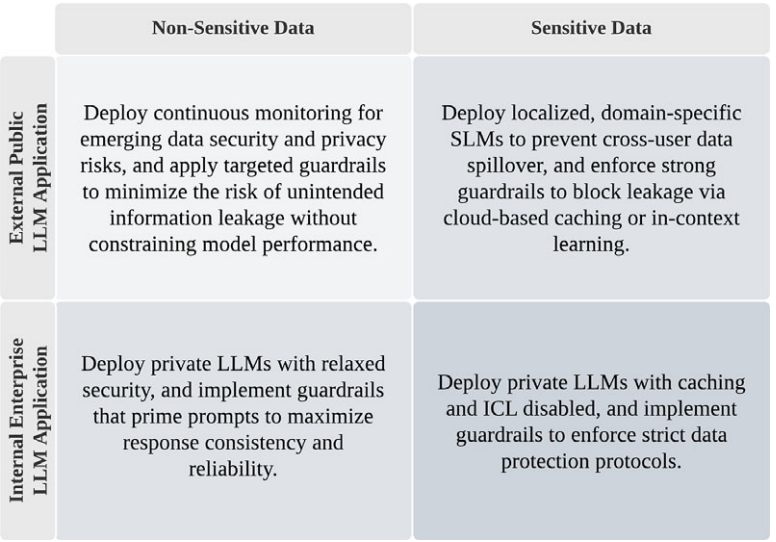


***Figure 4.*** *LLM deployment strategy matrix by sensitivity level and access type.*

### *5.2. Limitations and future direction*

While our study provides valuable insights into GPT-4's handling of domain-specific prompts, it is important to acknowledge several limitations that could impact the findings and their generalizability.

#### *5.2.1. Scope of analysis*

Our initial analysis focused on interrogative, token-limited prompts aiming to achieve a singular intent. Although this approach is valuable for establishing baseline results, it limits the scope of the study. To build statistical significance, future research should include a variety of intent-specific prompts across different domains. This would help in understanding how GPT-4 handles a broader range of tasks and interactions.

#### *5.2.2. Evaluation metrics*

Our study used specific computational linguistic parameters to evaluate response similarity, including positional, semantic, and syntactic similarity. While these metrics provide valuable insights, they do not capture all aspects of linguistic variation and may overlook subtleties in language use. Future research should incorporate additional evaluation metrics and human judgment to ensure a more comprehensive assessment of the model's performance (Huang and Chang, 2023).

#### *5.2.3. Architectural understanding*

Without precise knowledge of the architecture and mathematics supporting transformer layers (e.g., MHA and Feed-Forward Networks), we cannot fully ascertain how responses anchor to the domain. This limitation indicates a need for further simulations and cross-sectional analysis across these simulations to derive a comprehensive understanding of the transformational propensity for domain anchorage (Brown et al., 2020; Vaswani et al., 2017).

#### *5.2.4. Response similarity and information dissemination*

While response similarity is a valuable measure, it cannot by itself prove information dissemination. Similar responses could result from shared linguistic patterns rather than actual information transfer. Further studies are needed to explore how information might be disseminated across different contexts and users (Zhang et al., 2024).

https://doi.org/10.1017/dap.2025.10041 Published online by Cambridge University Press

To understand how domain anchorage evolves within the model, we can analyze the attention scores assigned to domain-specific terms and the hidden states at each layer of the transformer. This could involve visualizing the hidden states using dimensionality reduction techniques like t-SNE or PCA to identify clusters of domain-specific terms and track how these clusters evolve across layers. This analysis can reveal which layers are primarily responsible for identifying and extracting domain-specific information, and how the representation of these terms changes as they are processed through the network. By isolating the effects of specific layers, we can develop highly granular interventions, beyond contextual prompt engineering, to mitigate domain anchorage and domain-specific response bias.

Successful deployments treat domain anchoring and priming not as one-off tweaks but as ongoing disciplines, particularly in large public or private sector organizations. Furthermore, expanding the range of prompts and intents across different domains would enhance the study's generalizability and provide a more comprehensive understanding of the dynamics of domain anchorage in diverse scenarios. Additionally, incorporating additional quantifiable, unbiased metrics to assess information dissemination could offer deeper insights into the extent and nature of the problem. For instance, constructing network profiles or analyzing the centrality of employees' linguistic fingerprints within the model's internal representations could reveal potential pathways for information leakage and enable the development of targeted interventions.

A promising avenue for strategic exploration is the application of RolE Prompt Guided Multi-Domain Adaptation (REGA) to enterprise-scale LLM deployment (Wang et al., 2024). REGA introduces a three-part strategy: self-distillation, which preserves general capabilities without requiring access to proprietary pretraining data; role prompting, which clearly defines domain boundaries through assigned expert and generalist roles; and role integration, which consolidates domain-specific knowledge under a central prompt to enable seamless cross-domain generalization.

While this study provides a controlled, white-box analysis of domain anchorage in GPT-4, several limitations suggest directions for future work. The proprietary nature of GPT-4 restricts internal inspection, so future research may benefit from applying this framework to open-source models with accessible internals. To enhance practical relevance, one can explore black-box evaluation using output-only metrics such as entropy and divergence. Our prompt design prioritized experimental control, but expanding to more diverse prompt types and domains would improve generalizability. Additionally, incorporating broader linguistic features, such as discourse coherence and pragmatic cues, alongside human evaluation, may offer a more comprehensive view of anchorage effects. Finally, future work should deepen the discussions on interpretability, dissemination risks, and mitigation strategies.

Beyond enterprise applications, domain anchorage presents risks in personal-use scenarios such as counseling, advising, or self-help. In these contexts, language models may unintentionally reinforce harmful biases or unhealthy behavioral patterns if prior domain context shapes subsequent responses. Such effects carry important implications for fairness, user autonomy, and psychological safety. Additionally, persistent anchorage may complicate compliance with key GDPR principles, including data minimization, transparency, and the right to explanation (Selbst and Powles, 2017; Wachter et al., 2017). While LLMs do not retain input data explicitly, their latent alignment with earlier domain prompts can result in behavior that resembles indirect profiling or personalization, raising critical legal and ethical challenges for real-world deployment.

## 6. Conclusion

Our research investigated the impact of domain-specific priming on responses generated by GPT-4. We hypothesized that domain-specific words receiving higher gradient updates during training could introduce bias, create semantic echo chambers, and oversimplify relationships. To test this hypothesis, we analyzed whether GPT-4, when primed with domain-specific prompts, exhibits domain anchorage, where model responses become increasingly aligned with the primed context, regardless of prompt variation. Using a computational linguistic framework, we quantified linguistic similarity across prompts and outputs in semantic, syntactic, lexical, and positional dimensions. Our results provide empirical evidence

https://doi.org/10.1017/dap.2025.10041 Published online by Cambridge University Press

of domain anchorage, suggesting that domain-specific priming can shape model behavior in ways that are persistent and measurable. This research contributes to a broader understanding of unintended information persistence in LLMs and introduces a generalizable framework for detecting such effects. The findings support the development of guardrails for responsible LLM deployment and highlight practical strategies for mitigating risk, improving consistency, and enhancing transparency in enterprise and public-facing settings.

**Data availability statement.** The publicly available GPT-4 model was utilized for this research. Replication data and code can be found at https://doi.org/10.5281/zenodo.16930416 (Challappa et al., 2025).

**Author contribution.** Conceptualization: R.G. Methodology: L.C., Z.Z., and R.G. Data curation: L.C. and Z.Z. Data visualization: L.C. and Z.Z. Writing original draft: L.C., Z.Z., and R.G. All authors approved the final submitted draft.

**Funding statement.** This research was supported only by faculty support funds allocated to the principal investigator at the Goizueta Business School, Emory University.

**Competing interests.** None

## References

**Brown T**, **Mann B**, **Ryder N**, **Subbiah M**, **Kaplan JD**, **Dhariwal P**, **Neelakantan A**, **Shyam P**, **Sastry G**, **Askell A**, *et al.* (2020) Language models are few-shot learners. *Advances in Neural Information Processing Systems 33*, 1877–1901.

**Centre for Data Ethics and Innovation** (2021) *Public Attitudes to Data and AI Tracker Survey – Wave 1*. Available at https://www.gov.uk/government/publications/cdei-public-attitudes-to-data-and-ai-tracker-survey (accessed 29 March 2025.

**Cer D**, **Diab M**, **Agirre E**, **Lopez-Gazpio I and Specia L** (2017) SemEval-2017 task 1: Semantic textual similarity multilingual and Crosslingual focused evaluation. In Bethard S, Carpuat M, Apidianaki M, Mohammad SM, Cer D and Jurgens D (eds), *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada: Association for Computational Linguistics, pp. 1–14. https://doi.org/10.18653/v1/S17-2001.

**Challappa L**, **Zhang Z and Garg R** (2025) Detecting data leakage in large language models: A computational linguistic framework for domain Anchorage. *Zenodo*. https://doi.org/10.5281/zenodo.16930416.

**Chandrasekaran D and Mago V** (2021) Evolution of semantic similarity—a survey. *ACM Computing Surveys (Csur) 54*(2), 1–37.

**Chomsky N** (1972) "Remarks on Nominalization". *Studies on Semantics in Generative Grammar*, Berlin, Boston: De Gruyter Mouton, 11–61. https://doi.org/10.1515/9783110867589-003.

**Dai D**, **Sun Y**, **Dong L**, **Hao Y**, **Ma S**, **Sui Z and Wei F** (2023) Why Can GPT Learn In-Context? Language Models Secretly Perform Gradient Descent as Meta-Optimizers. In *Findings of the Association for Computational Linguistics: ACL 2023*, 4005–4019, Toronto, Canada: Association for Computational Linguistics.

**Dentella V**, **Günther F and Leivada E** (2023) Systematic testing of three language models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proceedings of theNational Academy of Sciences 120*(51), e2309583120.

**Engin Z**, **Gardner E**, **Hyde A**, **Verhulst S and Crowcroft J** (2024) Unleashing collective intelligence for public decision-making: The data for policy community. *Data & Policy 6*, e23.

**Hermann L**, **Fiedler T**, **Nguyen HA**, **Nowicka M and Bartoszewicz JM** (2024) Beware of data leakage from protein LLM pretraining. *Proceedings of the 19th Machine Learning in Computational Biology Meeting, in Proceedings of Machine Learning Research 261*, 106–116. Available from https://proceedings.mlr.press/v261/hermann24a.html.

**Huang J and Chang KCC** (2023) Towards Reasoning in Large Language Models: A Survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, 1049–1065, Toronto, Canada: Association for Computational Linguistics.

**Liu P**, **Yuan W**, **Fu J**, **Jiang Z**, **Hayashi H and Neubig G** (2023) Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys 55*(9), 1–35.

**Mihalcea R**, **Corley C**, **Strapparava C**, *et al.* (2006) Corpus-based and knowledge-based measures of text semantic similarity. *AAAI 6*, 775–780.

**Nasr, M**, **Rando, J**, **Carlini, N**, **Hayase, J**, **Jagielski, M**, **Cooper, AF, … and Lee, K.** (2025) Scalable extraction of training data from aligned, production language models. In *The Thirteenth International Conference on Learning Representations*. https://arxiv.org/abs/2311.17035

**Sargiotis D** (2024) Data Security and Privacy: Protecting Sensitive Information. In *Data Governance*. Springer, Cham. https://doi.org/10.1007/978-3-031-67268-2_6

**Schmidt H**, **Rotgans J and Mamede S** (2025) Bias sensitivity in diagnostic decision-making: Comparing ChatGPT with residents. *Journal of General Internal Medicine 40*, 790–795. http://doi.org/10.1007/s11606-024-09177-9.

**Selbst AD and Powles J** (2017) Meaningful information and the right to explanation' (2017). *International Data Privacy Law 7*, 233.

**Vaswani A**, **Shazeer N**, **Parmar N**, **Uszkoreit J**, **Jones L**, **Gomez AN**, **Kaiser Ł and Polosukhin I** (2017) Attention is all you need. *Advances in Neural Information Processing Systems 30*, 6000–6010. https://dl.acm.org/doi/10.5555/3295222.3295349.

**Wachter S**, **Mittelstadt B and Floridi L** (2017) Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International data privacy law 7*(2), 76–99.

**Wang R**, **Mi F**, **Chen Y**, **Xue B**, **Wang H**, **Zhu Q**, **Wong KF and Xu R** (2024) Role prompting guided domain adaptation with general capability preserve for large language models. In Duh K, Gomez H and Bethard S (eds), *Findings of the Association for Computational Linguistics: NAACL 2024*. Mexico City, Mexico: Association for Computational Linguistics, pp. 2243–2255. http://doi.org/10.18653/v1/2024.findings-naacl.145.

**Williamson SM and Prybutok V** (2024) Balancing privacy and progress: A review of privacy challenges, systemic oversight, and patient perceptions in AI-driven healthcare. *Applied Sciences 14*(2), 675.

**Xie SM**, **Raghunathan A**, **Liang P and Ma T** (2022) An Explanation of In-context Learning as Implicit Bayesian Inference. In International Conference on Learning Representations. https://arxiv.org/abs/2111.02080

**Zhang Y**, **Zhang M**, **Yuan H**, **Liu S**, **Shi Y**, **Gui T**, **Zhang Q and Huang X** (2024) LLMEval: a preliminary study on how to evaluate large language models. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'24/IAAI'24/EAAI'24)*. vol *38*, AAAI Press, Article 2186, 19615–19622. https://doi.org/10.1609/aaai.v38i17.29934.

## A. Appendix. Theoretical Exploration of the Mechanisms Underlying Domain Anchorage

Building upon the foundational concepts of ICL as implicit Bayesian inference (Xie et al., 2022) and meta-optimization as implicit gradient descent (Dai et al., 2023), we can mathematically formalize the relationship between these mechanisms and the observed phenomenon of domain anchorage. In this framework, ICL allows the model to infer a shared latent concept $c$ from demonstration examples $D$, as shown in the equation:

$$P(y|x,D) = \int P(y|x,c)P(c|D)dc \tag{A.1}$$

where $P(y|x,c)$ is the probability of generating $y$ given $x$ and the latent concept $c$, and $P(c|D)$ is the posterior probability of the latent concept $c$ given the demonstration examples $D$.

Meta-optimization, through implicit gradient descent, updates the model's attention weights $W$ to better capture the domain-specific context $d$. This adaptation of attention weights is represented as:

$$W' = W - \eta \nabla L(X, W) \tag{A.2}$$

where $\eta$ is the learning rate and $\nabla L(X, W)$ is the gradient of the loss function with respect to the attention weights.

The updated attention weights subsequently influence the posterior probability $P(c|D,d)$, leading to the generation of responses $y$ that are more aligned with the domain-specific context $d$, thus exhibiting domain anchorage. This relationship is captured in the modified probability equation:

$$P(y|x,D,d) = \int P(y|x,c,d)P(c|D,d)dc \tag{A.3}$$

This mathematical formulation provides a theoretical basis for understanding how domain anchorage emerges in GPT-4.

**Cite this article:** Challappa L, Zhang Z and Garg R (2025). Domain anchorage in LLMs: Lexicon profiling and unintended information leakage. *Data & Policy*, 7: e73. doi:10.1017/dap.2025.10041