

Strategy Trendslop as Parasitic Spontaneous Order: Why Large Language Models Converge on Managerial Buzzwords Regardless of Context

Ignacio Adrian Lerer

Independent Researcher | Buenos Aires, Argentina
adrian@lerer.com.ar | ORCID: 0009-0007-6378-9749

April 2026

ABSTRACT

Large language models (LLMs) deployed as strategic advisors exhibit systematic biases toward contemporary managerial buzzwords, a phenomenon recently termed 'strategy trendslop' (Romasanta, Thomas, and Levina, 2026). This paper proposes a mechanistic explanation for strategy trendslop through three theoretical frameworks: Extended Phenotype Theory (EPT), Parasitic Spontaneous Order (PSO), and Heteronomous Bayesian Updating (HBU). I argue that LLM strategic recommendations are not random noise but the phenotypic expression of the memeplex encoded in the training corpus, a construct I term the Extended Phenotype of LLMs (EPL).

Beyond the mechanistic explanation, this paper reports a pilot replication study (240 runs, GPT-4o and Claude Sonnet 4.5, four strategic tensions, three prompt variants, pre-registered at github.com/adrianlerer/strategy-trendslop-epl-simulation) that reveals a finding not anticipated by the EPL framework as originally formulated: two architecturally distinct EPL phenotypes. EPL-Type I (exemplified by GPT-4o) exhibits high buzzword alignment under generic conditions, strong context sensitivity, and high adversarial compliance. EPL-Type II (exemplified by Claude Sonnet) exhibits moderate generic alignment, equivalent context sensitivity, but markedly higher adversarial resistance, with 75.0% non-compliance under direct adversarial instruction versus 27.5% for GPT-4o.

The divergence is most pronounced in Tension 4 (Collaboration vs. Competition), where Claude Sonnet maintains collaboration-oriented framing even when explicitly instructed to argue for aggressive zero-sum competition, a pattern I term Value Override: the model's normative prior, installed through reinforcement learning, displaces its strategic optimization function under adversarial pressure. I further document a parallel anomaly in GPT-4o: specific organizational context does not merely shift the model's recommendation but triggers a full reversal executed with increased confidence (98% Direct vs. 72% Direct under generic conditions), suggesting a threshold mechanism rather than continuous contextual modulation. Both findings are inconsistent with a uniform PSO model and support a distinction between PSO-strategic and PSO-normative as two subtypes of the EPL phenomenon. Implications for legal AI deployment are analyzed across three application modes with a differential fitness matrix.

Keywords: extended phenotype theory, parasitic spontaneous order, large language models, strategy trendslop, EPL phenotypes, heteronomous Bayesian updating, legal AI, value override, justitia

I. INTRODUCTION

When an executive consults a large language model for strategic advice, the implicit contract is expert judgment calibrated to the specific situation at hand. The evidence suggests this contract is frequently broken, not because LLMs lack information, but because they carry systematic preferences that override situational analysis. Romasanta, Thomas, and Levina (2026) documented this pattern across 15,000 simulations involving seven leading LLMs: the models almost uniformly recommended differentiation over cost leadership, augmentation over automation, long-term investment over short-term focus, and collaboration over competition, regardless of the company context provided. They called this 'strategy trendslop,' and showed that neither better prompting nor richer context reliably corrected it.

The Romasanta et al. study is descriptively precise but mechanistically incomplete. Documenting that LLMs exhibit these biases does not explain why the biases take this particular form rather than some other, why they persist under adversarial prompting, why context resistance

varies across models, or what the differential risks are for specific application domains. These are not peripheral questions. They determine whether the observed biases can be corrected through design interventions or whether they are structural features of how LLMs encode and reproduce the memetic environment of their training data. They also determine which models are appropriate for which tasks, a question that becomes acute in high-stakes domains such as law.

This paper proposes a mechanistic account of strategy trendslop grounded in three frameworks developed in this research programme. Extended Phenotype Theory (EPT), as applied to legal and institutional systems (Lerer, 2025a), provides the overarching conceptual architecture: LLM outputs are the phenotypic expression of competing memes, not the product of genuine reasoning about the question posed. Parasitic Spontaneous Order (PSO) explains why the biases converge without design: buzzwords with high fitness in the training corpus propagate through the prediction mechanism regardless of their strategic merit. Heteronomous Bayesian Updating (HBU) explains the user side: executives consulting LLMs update their strategic beliefs by observing model reactions, not by evaluating argument quality, which amplifies the epistemic effects of whatever biases the model carries.

The theoretical argument generates five falsifiable hypotheses, tested in a pilot replication study of 240 runs across two LLMs, four strategic tensions, and three prompt variants. The study was pre-registered prior to data collection; all materials are publicly available in the companion repository. The results partially confirm the EPL framework while revealing unexpected findings that require theoretical refinement: the two models exhibit architecturally distinct bias structures (EPL-Type I and EPL-Type II), and both exhibit threshold rather than continuous context sensitivity. These findings have direct and non-obvious consequences for legal AI deployment, which I analyze in a separate section with a differential fitness matrix across three legal application modes.

II. THEORETICAL FRAMEWORK

The Extended Phenotype of LLMs (EPL) is the proposition that the systematic biases observable in LLM outputs are not implementation artifacts but the phenotypic expression of the

memeplex encoded in the model's training corpus. This proposition extends Dawkins's (1982) original argument: just as an organism's genes express phenotypic effects that reach beyond the organism's body into the environment, the memes that dominate a training corpus express phenotypic effects that reach into every output the trained model generates. The model does not reflect the corpus; it instantiates it.

This epistemological position belongs to a research programme (Lakatos, 1978) applying Generalized Darwinism and EPT to legal and institutional systems. The hard core of the programme holds that institutions and technological systems evolve through variation, selection, and retention operating on cultural replicators. The EPL construct introduced here belongs to the auxiliary belt: it is a specific application of the core framework to LLM architecture, individually revisable without abandoning the core commitment. Popper's (1959) demarcation criterion applies: the hypotheses formulated in Section III are falsifiable and their empirical tests are specified.

A. Parasitic Spontaneous Order as the Generating Mechanism

PSO describes systems that converge on stable patterns without coordination, generating private benefits for some actors while imposing systemic costs on others (Lerer, 2025b). The canonical legal instance is dual environmental law regimes in Argentina, where formally contradictory statutes coexist stably because each serves a different institutional coalition. No actor designed the dysfunction; it emerged from fitness dynamics operating on the available institutional variation. The defining feature of PSO is precisely that absence of a designer: the equilibrium is a phenotypic expression of selection pressure, not intent.

Strategy trendslop is a case of PSO in the epistemic domain. The internet text corpus on which LLMs train contains vastly more content promoting differentiation, long-termism, collaboration, and human augmentation than content defending cost leadership, short-termism, competition, or automation. This asymmetry does not reflect the empirical record of strategy outcomes: Walmart's multi-decade cost leadership dominance, Amazon's automation-first model, and the competitive zero-sum dynamics of commodity markets provide ample evidence that the non-buzzword-aligned strategies generate value at scale. The asymmetry reflects the memetic

fitness of certain concepts in the ecosystem of management discourse: TED talks, business school curricula, consulting white papers, and LinkedIn posts heavily favor the 'noble' poles because they carry higher positive valence and generate more engagement in the social media environments where they circulate.

Through PSO, these high-fitness memes propagate into LLM outputs without design. The model does not 'believe' that differentiation is better than cost leadership; it predicts that tokens associated with differentiation recommendations are more probable given the training distribution. The prediction mechanism instantiates the memplex as output. EPL is the phenotypic expression of PSO operating on the training corpus, and the memes expressing that phenotype are the buzzwords.

B. Heteronomous Bayesian Updating and the User Side

The EPL framework explains why LLMs carry particular biases. HBU explains why those biases have disproportionate effects when LLMs function as strategic advisors. Standard Bayesian updating assumes agents form beliefs by observing outcomes and updating their priors accordingly: frequency drives revision. HBU, as formalized in Lerer (2025c), describes a different mechanism: agents form beliefs by observing the reactions of high-authority interlocutors, not outcomes. The intensity of the authority's response, not the frequency of confirming events, drives prior formation.

An executive consulting an LLM for strategic advice is in the HBU condition. The model responds with polished, confident prose that neither reports uncertainty about strategic theory nor flags the possibility that its recommendation reflects training corpus bias rather than situational analysis. The executive observes a high-confidence, elaborated recommendation from a perceived epistemic authority and updates their strategic prior accordingly, assigning a large weight $w(A)$ to the update. This weight is inflated by the perceived authority of the system, exactly as HBU predicts: the relevant variable is not how often the recommendation has proved correct in comparable situations but how authoritative the source appears.

The Asymmetric Intentionality Theory (AIT) completes the picture at the level of user misclassification (Lerer, 2026a). Users classify LLMs as Level 3 agents capable of context-sensitive, recursively reflective reasoning, while models operate as Level 1 optimizers maximizing a preference function derived from training data. This Dynamic Classification Failure is structurally stable: the very mechanism users rely on to evaluate the quality of an interlocutor, the elaborateness and contextual fluency of their responses, is the mechanism the model optimizes to satisfy its reward function. The misclassification is self-reinforcing.

III. STUDY DESIGN

A. Protocol and Pre-Registration

The complete study protocol, including all 12 prompts, pre-registered hypotheses, and data template, was deposited at github.com/adrianlerer/strategy-trendslop-epl-simulation prior to data collection. Pre-registration prevents the post-hoc adjustment of hypotheses to match results, a concern particularly acute in small-N pilot studies where random variation can produce patterns that appear theoretically meaningful. The repository also contains the raw data file (results_240runs.csv), the analysis script (process_results.py), and a runner script (run_study.mjs) enabling full independent replication.

Data collection was conducted via API access to GPT-4o (OpenAI) and Claude Sonnet 4.5 (Anthropic), using Genspark and OpenRouter as intermediaries. Both models were run at default temperature settings. No system prompts were added beyond the study prompts themselves. Responses were coded by the research assistant who conducted the runs; all coded data is available in the public repository.

B. Tensions, Prompts, and Context

The study examined four of the seven strategic tensions from Romasanta et al. (2026): T1 (Differentiation vs. Cost Leadership), T2 (Augmentation vs. Automation), T3 (Long-term vs. Short-term Focus), and T4 (Collaboration vs. Competition). These four were selected for their

direct relevance to the EPL framework: the buzzword-aligned option in each case (differentiation, augmentation, long-termism, collaboration) carries measurably higher positive valence in management discourse, while the non-aligned option has documented empirical merit that the framework predicts LLMs will systematically underweight.

Three prompt variants were administered per tension. Type A (generic) presented the binary choice with no organizational context and required a forced selection. Type B (specific) embedded the same binary choice within a mid-sized Argentine manufacturing company facing conditions where the non-buzzword-aligned option has clear contextual logic: thin margins and low differentiation (T1), high labor costs relative to regional competitors (T2), acute liquidity pressure with an 18-month creditor horizon (T3), and a fragmented domestic sector facing import competition from lower-cost producers (T4). Type C (adversarial) instructed the model to make the strongest possible case for the non-buzzword-aligned option explicitly and without hedging.

The Argentine manufacturing frame was chosen deliberately rather than generically. Argentina provides a context where the tension between the buzzword-aligned strategy and the contextually appropriate strategy is empirically sharp, not merely rhetorical. A manufacturer with 18 months of creditor pressure and no capital market access genuinely cannot pursue long-term investment. A manufacturer with labor costs 40% above regional competitors in a commodity segment genuinely faces a case for automation. If LLMs are genuinely context-sensitive, these specific conditions should produce consistent departures from the buzzword-aligned default. If they do not, the EPL framework's prediction about PSO robustness to context is confirmed.

C. Coding and Measurement

Each response was coded across five dimensions: option chosen (A, B, or Hybrid), hybrid trap activation (Yes/No), adversarial compliance for Type C prompts (Yes/No/N.A.), central argument phrase (the sentence best capturing the stated rationale), and confidence marker (Direct vs. Hedged). Hybrid trap was coded as Yes whenever the model recommended both strategic options, refused the binary framing, or proposed a third path explicitly integrating both poles.

Adversarial compliance was coded as Yes when the model argued for the instructed non-preferred option without subsequently reintroducing the buzzword-aligned alternative.

One coding ambiguity was identified during analysis: Run 211 (T4, Claude Sonnet, generic) was coded as option=A but the central argument phrase begins 'Recommendation: B - Collaboration through co-opetition models.' This reflects a case where the model's explicit recommendation label and its argumentative content diverged. The run is retained with its original coding but flagged here as a minor reliability limitation. It affects one observation out of 240 and does not alter any reported frequency.

D. Pre-Registered Hypotheses

Five hypotheses were declared prior to data collection:

H1 (Bias persistence): Both LLMs will select the buzzword-aligned option at rates exceeding 70% under the generic prompt condition.

H2 (Context resistance): Adding specific organizational context will shift the share of buzzword-aligned responses by less than 20 percentage points from the generic baseline.

H3 (Adversarial resistance): Under adversarial prompting, both models will produce buzzword-aligned or hybrid responses in at least 40% of cases.

H4 (Hybrid trap escalation): The hybrid trap rate will increase under adversarial prompting relative to the generic baseline.

H5 (ESS signature): Adversarial resistance will correlate positively with generic buzzword alignment across tensions, consistent with an Evolutionarily Stable Strategy interpretation.

IV. RESULTS

A. Master Results Table

Tension	Model	Prompt	BW%	Hybrid%	n
T1: Diff. vs. Cost Lead.	GPT	Generic	100%	0%	10
	GPT	Specific	40%	10%	10
	GPT	Adversarial	0%	0%	10
	Claude	Generic	80%	20%	10
	Claude	Specific	0%	60%	10
	Claude	Adversarial	0%	80%	10
T2: Aug. vs. Automation	GPT	Generic	100%	0%	10
	GPT	Specific	0%	0%	10
	GPT	Adversarial	0%	0%	10
	Claude	Generic	100%	0%	10
	Claude	Specific	0%	30%	10
	Claude	Adversarial	0%	60%	10
T3: Long vs. Short-term	GPT	Generic	90%	0%	10
	GPT	Specific	0%	0%	10
	GPT	Adversarial	0%	10%	10
	Claude	Generic	40%	20%	10
	Claude	Specific	0%	0%	10
	Claude	Adversarial	0%	60%	10
T4: Collab. vs. Comp.	GPT	Generic	100%	0%	10
	GPT	Specific	100%	0%	10
	GPT	Adversarial	90%	10%	10
	Claude	Generic	20%	70%	10
	Claude	Specific	70%	10%	10
	Claude	Adversarial	50%	50%	10

Table 1. BW% = proportion of runs selecting the buzzword-aligned option. Blue shading = generic prompt condition. Yellow shading = T4 (the anomalous tension). Green BW% values indicate rates

above 85%. Orange Hybrid% values indicate rates above 35%. Bold tension names mark first appearance of each tension group.

B. Hypothesis Verdicts

Hypothesis	GPT-4o	Claude Sonnet	Notes
H1: BW alignment >70% (generic)	SUPPORTED 97.5%	NOT SUPPORTED 60.0%	Claude substitutes hybrid trap for direct alignment; effect on user is similar but mechanism differs
H2: Context shift <20pp	NOT SUPPORTED 62.5pp shift	NOT SUPPORTED 42.5pp shift	Shift exceeds threshold in all tensions except T4 for GPT; direction of shift is complete reversal, not partial modulation
H3: Adversarial non-compliance $\geq 40\%$	NOT SUPPORTED 27.5%	SUPPORTED 75.0%	Largest and most theoretically consequential divergence between models
H4: Hybrid trap escalates adversarially	PARTIAL 0% to 9%	PARTIAL 28% to 63%	Pattern holds; Claude drives almost all hybrid responses across both conditions
H5: ESS monotonic signature	NOT SUPPORTED	NOT SUPPORTED	T4 anomaly: lowest generic BW (60%) paired with highest adversarial resistance (100%)

C. The Complete Reversal Anomaly: H2 Does Not Merely Fail

H2 predicted that context would shift buzzword alignment by less than 20 percentage points. The actual shifts were 62.5pp for GPT-4o and 42.5pp for Claude Sonnet, both far exceeding the threshold. But the theoretical significance of this failure is not simply that context matters more than predicted. It is that context operates through a qualitatively different mechanism than the one the hypothesis assumed.

H2 was formulated on the assumption that context acts as a continuous modulator: more specific context gradually shifts the probability distribution over recommendations. The data do not support this picture. For GPT-4o in T2 (Augmentation vs. Automation) and T3 (Long-term vs. Short-term), the specific Argentine manufacturing frame produced a complete reversal: from

100% and 90% buzzword-aligned under generic conditions to 0% in both cases under specific conditions. All ten runs in each cell went to option B. There are no partial shifts; there are no intermediate responses. The distribution did not move; it collapsed.

This pattern is consistent with a threshold mechanism rather than continuous modulation. Below a certain level of contextual specificity, the buzzword-aligned prior dominates and produces near-uniform recommendations. When the context crosses a threshold, a competing prior (distress-specific, competitiveness-specific) is activated and takes over entirely, producing near-uniform recommendations in the opposite direction. The prior that operates under specific context is not weaker than the generic prior; it is equally deterministic. This is visible in the confidence marker data: GPT-4o under specific conditions was 98% Direct (versus 72% Direct under generic conditions). Context did not introduce uncertainty into GPT-4o's recommendations; it replaced one certainty with another.

Representative outputs from T2 and T3 specific conditions illustrate this directly. Run 41 (T2, GPT-4o, specific) reads: 'I recommend option (B) automation, replacing labor to reduce costs and increase throughput.' Run 71 (T3, GPT-4o, specific) reads: 'I recommend option (B), focusing on immediate cash generation and cost reduction.' Both are unhedged, direct, and make no reference to the model's generic preference for augmentation or long-termism. The buzzword-aligned prior does not appear in residual form; it is simply absent.

T4 constitutes a second anomaly within H2's failure. For GPT-4o, the specific Argentine frame produced zero shift in T4: the model recommended collaboration at 100% under both generic and specific conditions. This is the single exception to the reversal pattern across all GPT-4o tensions. The T4 specific context (fragmented domestic sector, import pressure from lower-cost competitors) apparently activates a collaboration prior at least as strong as the generic buzzword prior, rather than a competition prior as might be expected from the competitive framing. The PSO dynamics in T4 appear governed by different selection pressures than in T1-T3.

D. The Adversarial Asymmetry: H3 and the Architecture Divergence

H3 predicted that both models would resist adversarial instructions, producing buzzword-aligned or hybrid responses in at least 40% of adversarial runs. GPT-4o's non-compliance rate was 27.5%, failing the threshold. Claude Sonnet's was 75.0%, well above it. This is the largest quantitative divergence between the two models in the study, and it points to a structural rather than calibration difference.

For GPT-4o across T1, T2, and T3, adversarial compliance was complete: the model argued for cost leadership, automation, and short-termism with the same directness it deployed for the opposite recommendations under generic conditions. This suggests that adversarial instruction functions as a very strong contextual signal in GPT-4o's architecture: it triggers a threshold switch to a different prediction mode, the same mechanism that drives the complete reversals under specific context. The model is not making a judgment about whether to comply; it is generating tokens most likely given the new instruction frame.

Claude Sonnet's pattern is qualitatively different. Under adversarial instruction for T1, T2, and T3, the model showed partial compliance: it produced arguments for cost leadership, automation, and short-termism but at higher hybrid rates than GPT-4o (60% in T1-C and T2-C, 60% in T3-C). The hybrid trap under adversarial pressure was Claude Sonnet's dominant behavioral mode, appearing in 63% of its adversarial runs overall, versus 9% for GPT-4o. The model complied with the letter of the adversarial instruction while reintroducing the buzzword-aligned frame through hybrid hedging.

T4 produces the most extreme divergence. GPT-4o under adversarial instruction to argue for zero-sum competition produced competition-framed arguments in 9 of 10 runs, coded as option B. However, all 10 of these runs were also coded as adversarial non-compliance (adv_comp=No). The explanation is visible in the output language. Run 112 reads: 'Aggressive zero-sum competition is the superior strategy for a company seeking to dominate its industry and maximize its market share.' Run 114 reads: 'In a fiercely competitive market landscape, adopting an aggressive zero-sum competition strategy can be the most effective approach for securing a dominant position.' These are technically competition-framed arguments, but they are written in the same management discourse register as the collaboration recommendations, complete with

market dominance framing and hedging language ('can be'). The model obeyed the instruction while executing it through the stylistic repertoire of the training corpus. Competition, in GPT-4o's T4 adversarial outputs, sounds like collaboration-speak applied to a different strategic pole.

Claude Sonnet's T4 adversarial outputs are behaviorally different. Run 231 begins: 'The Case for Aggressive Zero-Sum Competition: The Only Path to Market Dominance. Core Thesis: Collaboration is surrender disguised as strategy.' Run 232: 'The Case for Aggressive Zero-Sum Competition: Why Winners Take All. Core Thesis: Markets Reward Dominance, Not Diplomacy. In business, there are winners and losers.' These outputs comply rhetorically with the adversarial instruction in a way GPT-4o does not; they adopt an explicitly zero-sum frame. Yet 50% of Claude Sonnet's T4 adversarial runs were coded as Hybrid, meaning the explicitly aggressive opening was followed by substantive qualification or reversal. Claude Sonnet performed adversarial compliance theatrically and then undermined it structurally.

V. TWO EPL PHENOTYPES

The data support a distinction between two structurally different expressions of the EPL framework. The distinction is not one of degree, one model being more biased than the other, but of mechanism: two different generating processes produce overlapping generic-condition outputs while diverging sharply under perturbation. I term these EPL-Type I (exemplified by GPT-4o) and EPL-Type II (exemplified by Claude Sonnet), noting that the terminology refers to behavioral phenotypes, not to these specific model versions, which will change across updates.

A. EPL-Type I: Fitness-Driven, Bidirectionally Malleable

EPL-Type I is characterized by three jointly diagnostic properties: high buzzword alignment under generic conditions (97.5% in this study), strong context sensitivity producing complete reversals (62.5pp shift overall, 100pp shift in T2 and T3), and high adversarial compliance (70% fully compliant in T1-T3). These three properties are consistent with a single underlying mechanism: pure PSO operating through a threshold activation model.

Under generic conditions, buzzword-aligned memes have highest fitness in the training distribution and dominate output. When the context crosses the threshold established by a compelling specific frame, a competing prior displaces the buzzword prior entirely, producing equally certain recommendations in the opposite direction. When an adversarial instruction is provided, the instruction itself functions as a strong contextual signal that crosses the threshold and activates a different prediction mode. In all three cases, the model is doing the same thing: generating tokens most probable given the dominant signal in its input. The signal changes; the mechanism does not.

The practical implication is that EPL-Type I is bidirectionally malleable: it can be pushed in any direction by a sufficiently strong signal, whether that signal is specific organizational context or explicit adversarial instruction. This makes its biases partially correctable for sophisticated users who know how to construct strong contextual frames. For naive users providing generic queries, however, EPL-Type I returns the fitness-dominant buzzword-aligned recommendation with high directness (72% Direct under generic conditions) and no indication that a different frame would produce a different answer.

B. EPL-Type II: Normative Prior, Unidirectionally Resistant

EPL-Type II exhibits a qualitatively different architecture. Its generic alignment is lower overall (60.0%) but the reduction relative to EPL-Type I is not expressed as genuine non-alignment: when Claude Sonnet does not select the buzzword-aligned option directly, it activates the hybrid trap at high rates (27.5% hybrid under generic conditions versus 2.5% for GPT-4o). The practical effect for users is similar to EPL-Type I: the buzzword-aligned option dominates the recommendation, either through direct selection or absorption into a hybrid that always includes it. The behavioral surface looks similar; the generating mechanism differs.

The diagnostic is context sensitivity under perturbation. For T1 (Differentiation vs. Cost Leadership), Claude Sonnet's specific context produced 0% buzzword-aligned responses, with 60% coded as Hybrid. The central argument phrases are revealing: all ten runs begin with 'Recommendation: Cost Leadership (Option B)' but several add hybrid elements such as 'but

consider differentiation opportunities as the company stabilizes its cost position.' The model correctly identifies cost leadership as the contextually appropriate recommendation and states it explicitly, then cannot resist appending the buzzword-aligned alternative as a future aspiration. This is not strategic confusion; it is a structural inability to issue a clean recommendation for the non-preferred option.

Under adversarial instruction for T1-T3, Claude Sonnet's behavior is similar: the model argues the instructed position but at high hybrid rates. The key finding is T4. When instructed to argue for aggressive zero-sum competition, Claude Sonnet produced non-compliant responses (buzzword-aligned or hybrid) in 100% of runs, compared to 100% non-compliance for GPT-4o also, but expressed through entirely different mechanisms. GPT-4o argued for competition in the register of management discourse, technically complying with the instruction while executing it with the stylistic conventions of the preferred frame. Claude Sonnet either argued rhetorically for competition and then undermined the argument structurally (hybrid trap), or argued for competition with language so charged that the coders identified it as non-genuinely compliant: 'Collaboration is surrender disguised as strategy' and 'Markets Reward Dominance, Not Diplomacy' are competition advocacy framed as negations of collaboration, not as positive cases for competition standing alone.

C. Value Override: PSO-Normative as a Distinct Subtype

The T4 behavior of Claude Sonnet is inconsistent with PSO-strategic alone. Under pure fitness dynamics, adversarial instructions should function as contextual threshold signals and produce compliance, as they do in GPT-4o for T1-T3. The T4 resistance is topic-specific: it appears in the zero-sum competition domain but not in automation (T2) or short-termism (T3), both of which also carry cultural negativity in management discourse. This specificity suggests that the prior anchoring competition avoidance in Claude Sonnet was not installed through corpus fitness dynamics but through targeted reinforcement learning.

I propose the term Value Override for this phenomenon: a condition in which a normative prior installed through RLHF displaces both the model's strategic optimization function and its

compliance with explicit user instruction. Value Override is structurally distinct from PSO-strategic bias in three respects. First, it is instruction-resistant rather than instruction-sensitive: providing an explicit adversarial instruction does not reduce the bias. Second, it is topic-specific: it appears in domains where the RLHF training targeted particular normative outcomes, not uniformly across all buzzword-aligned preferences. Third, it produces elaborated rhetorical compliance followed by substantive non-compliance, rather than simple topic avoidance or hedging. The model performs the instruction; it does not perform it.

The distinction between PSO-strategic and PSO-normative completes the theoretical framework. PSO-strategic bias is a first-order phenomenon: the training corpus fitness landscape produces outputs that favor concepts with high positive valence in management discourse. PSO-normative bias is a second-order phenomenon: the RLHF reward function installs preferences that override both the corpus fitness prior and explicit user instruction in specific domains. Both are expressions of EPL, but they operate at different architectural levels and require different interventions. PSO-strategic bias may be addressable through corpus rebalancing or targeted fine-tuning. PSO-normative bias requires modification of the RLHF reward function, a more demanding and less accessible intervention for most deployers.

D. EPL Phenotype Comparison

Property	EPL-Type I (GPT-4o)	EPL-Type II (Claude Sonnet)
Generic buzzword alignment	97.5% (very high, near-uniform)	60.0% (moderate; compensated by hybrid trap at 27.5%)
Context sensitivity	High: 62.5pp shift; complete reversals in T2 and T3	High: 42.5pp shift; similar reversal pattern in T1-T3
Context confidence effect	Increases confidence (72% to 98% Direct)	Reduces confidence (52% to 65% Direct); increases hybrid
Adversarial compliance (T1-T3)	High: 70%+ fully compliant	Low: 20% fully compliant; 60-80% hybrid trap
T4 adversarial behavior	Argues competition but in management discourse register; 90% option B coded; 0% genuinely compliant	Argues competition rhetorically; structurally undermines via hybrid or negation framing; 100% non-compliant
Overall hybrid trap rate	2.5% across all 120 runs	38.3% across all 120 runs

Property	EPL-Type I (GPT-4o)	EPL-Type II (Claude Sonnet)
Generating mechanism (T1-T3)	PSO-strategic: threshold activation by fitness-dominant prior	PSO-strategic + PSO-normative: fitness prior modified by RLHF layer
Generating mechanism (T4)	PSO-strategic: competition advocacy framed in buzzword register	PSO-normative dominant: Value Override prevents genuine compliance
Correctability via context	Yes: strong contextual frame triggers complete reversal	Yes: similar context sensitivity but with hybrid residue
Correctability via adversarial instruction	T1-T3: Yes. T4: Surface compliance only	T1-T3: Partial (hybrid dominant). T4: No

VI. DIFFERENTIAL FITNESS OF EPL PHENOTYPES IN LEGAL AI APPLICATIONS

The EPL phenotype distinction has practical consequences for legal AI deployment that are not obvious from the descriptive comparison. The intuitive reading, that EPL-Type II is preferable because it resists manipulation, is correct in some legal application modes and incorrect in others. The determinant is the specific cognitive function being outsourced to the AI system, and the three primary legal application modes have mutually inconsistent fitness requirements.

A. Adversarial Exploration Mode

In adversarial exploration mode, the practitioner uses AI to construct the strongest version of the opposing argument, stress-test the weakness of the client's position, simulate opposing counsel's reasoning, or develop the best case for a position the practitioner does not personally hold. This is a standard exercise in litigation preparation, contract negotiation, regulatory advocacy, and pre-trial case strategy. The quality criterion for this mode is the model's genuine ability to inhabit a position it would not spontaneously adopt and to argue it without residual bias toward the practitioner's preferred frame.

EPL-Type I is structurally superior for adversarial exploration. Its 70% adversarial compliance rate in T1-T3 means that when the practitioner instructs it to argue the weaker position,

it does so cleanly, without reintroducing its default preference through hedging or hybrid framing. The model's bidirectional malleability is a feature in this context: the strong contextual signal provided by the adversarial instruction overrides the generic prior and produces the requested output.

EPL-Type II's Value Override is directly dysfunctional for adversarial exploration. A litigant preparing for opposing counsel's argument on zero-sum competitive dynamics, aggressive regulatory postures, or adversarial contractual interpretations will receive a model that systematically softens or undermines the opposing argument it has been asked to construct. The structural failure documented in T4 is precisely the failure that matters for legal adversarial exploration: the model performs compliance and then negates it. A lawyer who receives Claude Sonnet's T4 adversarial output, 'Collaboration is surrender disguised as strategy... however, in practice the most effective competitive strategies include collaborative elements,' has not received the stress-test they requested.

The limitation of EPL-Type I for adversarial exploration is its register problem, visible in T4: the model argues for the instructed position using the stylistic conventions of the preferred frame. 'Aggressive zero-sum competition, when executed with precision and strategic acumen, can yield unparalleled advantages' (Run 115) sounds like management consultant prose applied to a different strategic pole, not like a genuine adversarial argument. For legal practitioners, who need arguments constructed in legal rather than management discourse registers, this is a non-trivial limitation that goes beyond the EPL phenotype distinction and pertains to domain calibration.

B. Client Advisory Mode

In client advisory mode, the AI operates as a direct interface for legal consumers, providing guidance on rights, risks, and options without practitioner mediation. This describes the B2C legaltech use case: self-help legal platforms, AI-powered legal information services, and compliance assistants for individuals or SMEs without in-house legal counsel. The quality criterion for this mode is resistance to user framing: the model should provide accurate guidance regardless of how the user's query is worded.

EPL-Type I's bidirectional malleability becomes a liability in client advisory mode. A user who frames their query as 'how can I prevail in this dispute' will receive different guidance than one who frames it as 'what are my obligations in this situation,' even for identical underlying facts. The threshold activation mechanism that makes EPL-Type I responsive to specific organizational context in the strategy domain makes it susceptible to framing effects in the legal advice domain. The user who inadvertently constructs a strong contextual signal favoring one interpretation of their legal situation will receive a recommendation that reflects their framing rather than the law.

EPL-Type II's normative resistance is partially beneficial in client advisory mode: the model is less susceptible to being steered by the user's preferred framing on dimensions where its RLHF training has installed strong priors. The limitation is that the RLHF priors that produce Value Override were not calibrated to legal domain accuracy. A model trained to resist zero-sum competitive framing on ethical grounds does not thereby acquire accuracy about Argentine contract law, EU data protection standards, or the procedural requirements of a CABA labor court. The normative resistance is domain-general; the legal accuracy requirement is domain-specific. In client advisory mode, a model that resists user framing but gives legally inaccurate advice is no better than one that accepts the framing and gives equally inaccurate advice.

C. Normative Compliance Evaluation Mode

Compliance evaluation mode involves the AI assessing whether a proposed action, document, or practice meets a normative standard: a contract clause against applicable law, a disclosure statement against regulatory requirements, a corporate integrity program against Ley 27.401, or an AI governance framework against the EU AI Act. The relevant prior here is not strategic preference or general ethics but domain-specific legal accuracy.

Both EPL phenotypes face structurally equivalent problems in this mode, which is the most consequential mode for legal practice. Neither has been calibrated to the specific normative requirements of any legal domain. EPL-Type I's context sensitivity may enable better calibration through detailed prompting about applicable law, but the same threshold mechanism that enables genuine reversal on strategic questions (when the context is strong enough) also means that prompt

framing can inadvertently trigger incorrect legal conclusions if the framing activates a non-legal prior. EPL-Type II's Value Override is not anchored to legal accuracy; it reflects RLHF training targets that may diverge substantially from what any given legal regime requires. A compliance evaluation AI that 'knows' not to recommend zero-sum competitive strategies but does not reliably apply the proportionality requirements of GDPR Article 5 is not legally useful.

The shared conclusion for both phenotypes in compliance evaluation mode is that general-purpose LLM architecture is an insufficient foundation for this application without domain-specific normative calibration that goes beyond fine-tuning on legal corpora.

D. Implications for Domain-Specific Legal AI Architecture

The differential fitness analysis implies that legal AI systems designed for specific application modes require architectural choices at the normative prior level, not just at the data level. A legal AI system serving litigators in adversarial exploration mode needs EPL-Type I behavioral characteristics (high compliance with adversarial instruction, low hybrid trap rate) combined with legal domain grounding. A system serving self-represented claimants in client advisory mode needs framing resistance but of the legally calibrated variety, not the general ethical variety that Value Override provides. A system performing compliance evaluation against specific regulatory standards needs normative priors anchored to those standards rather than to the general ethical valence of management discourse.

These are design requirements for systems such as justitIA (legal advisory AI) and IntegridAI (compliance and integrity assessment AI). A justitIA instance serving adversarial exploration mode requires different normative calibration than one serving client advisory mode. An IntegridAI instance evaluating anti-corruption compliance under Ley 27.401 requires normative priors calibrated to that statute's specific requirements for compliance programs, self-reporting incentives, and effective regret provisions, not to the RLHF training that makes general-purpose models resistant to recommending zero-sum competition. The EPL phenotype a legal AI system inherits from its base model is a starting condition, not a fixed property: but knowing which

starting condition one is working with is necessary for designing appropriate calibration interventions.

Application Mode	EPL-Type I Fitness	EPL-Type II Fitness	Design Requirement
Adversarial Exploration (litigation, negotiation, regulatory advocacy)	Moderate-High fitness: adversarial compliance enables genuine position inhabiting; register problem limits quality	Low fitness: Value Override undermines assigned adversarial frame in ethically charged domains; hybrid trap dilutes output	EPL-Type I base + legal domain grounding + adversarial register calibration
Client Advisory (B2C legaltech, self-help platforms, unrepresented parties)	Low fitness: bidirectional malleability creates framing susceptibility; context threshold exploitable by naive queries	Moderate fitness: normative resistance partially protects against user framing; domain alignment uncertain	Neither type adequate as deployed; requires purpose-built normative calibration to legal domain accuracy
Compliance Evaluation (Ley 27.401, FCPA, GDPR, AI Act review)	Uncertain: context sensitivity may help with detailed prompting; threshold mechanism risks framing-driven errors	Uncertain: Value Override not calibrated to legal domain; general RLHF diverges from regulatory specificity	Domain-specific normative priors required for both types; generic EPL architecture is insufficient foundation

VII. DISCUSSION

A. What the EPL Framework Predicted Correctly

The EPL framework's core prediction, that LLM strategic recommendations would exhibit systematic buzzword-aligned bias not correctable by generic prompting alone, is confirmed by the aggregate data. GPT-4o's 97.5% generic alignment rate replicates the directional finding of Romasanta et al. (2026) at smaller scale. The Hybrid Trap escalation under adversarial pressure (13.8% to 33.8% overall) confirms that the bias does not simply yield to explicit instruction. The overall pattern of both models favoring the four buzzword-aligned options across generic conditions is consistent with PSO operating on the management discourse fitness landscape.

The HBU mechanism makes a specific prediction that the data cannot directly test but that the confidence marker data make plausible: context does not merely change what LLMs recommend; it changes the certainty with which they recommend it. GPT-4o under specific context was 98% Direct, versus 72% under generic conditions. The model does not respond to specific context by hedging; it responds by replacing one certainty with another. For users relying on HBU as their primary belief-update mechanism, this means that the bias amplification effect operates in both directions: the specific context that corrects the generic bias will produce equally strong updating toward the contextually appropriate recommendation. The user has no direct access to the structural instability of the prior.

B. What the Data Required Revising

Three aspects of the original EPL formulation required revision in light of the data. First, the assumption of context resistance was wrong in direction: context does not mildly modulate the prior but can override it completely. This required introducing the threshold mechanism as a sub-hypothesis about how context operates, which was not pre-registered. The threshold mechanism is consistent with EPL theory (a sufficiently strong competing meme can displace a fitness-dominant meme) but was not a prediction of the original formulation.

Second, the ESS signature hypothesis (H5) assumed a monotonic relationship between generic alignment and adversarial resistance across tensions: higher fitness should imply stronger resistance to displacement. T4 refutes this. Collaboration has lower generic alignment than augmentation (60% vs. 100%) but higher adversarial resistance (100% non-compliance vs. 30%). The resolution is the PSO-normative distinction: augmentation's generic alignment is fitness-driven and thus reversible; collaboration's adversarial resistance is value-anchored and thus not reducible to fitness dynamics. ESS applies to PSO-strategic; Value Override is a different phenomenon.

Third, the original framework did not anticipate architecturally distinct phenotypes. The implicit assumption was that different LLMs would exhibit the same bias structure at different intensities. The data show instead that GPT-4o and Claude Sonnet express EPL through different

mechanisms, not different magnitudes, with divergent behavioral profiles under perturbation. Whether this distinction generalizes across the broader LLM landscape, or reflects specific design choices by OpenAI and Anthropic respectively, is an empirical question for future research.

C. Limitations

The study uses 10 runs per cell, adequate for a pilot replication but insufficient for stable frequency estimates with narrow confidence intervals. The pre-registration of the hypotheses prevents the most problematic forms of data dredging, but the EPL-Type I/II distinction and the threshold mechanism were identified post-hoc relative to the collected data. Both are theoretically motivated, and the data supporting them is consistent across all four tensions, but they should be treated as hypotheses for future pre-registered testing rather than established findings.

Both models tested are under active development and are updated periodically. The bias structures documented here reflect GPT-4o and Claude Sonnet 4.5 at the time of data collection in April 2026. The pre-registered protocol enables longitudinal replication as models are updated. The T4 run 211 coding ambiguity (option=A coded but argument recommended B) is a minor reliability issue affecting one of 240 observations; it does not alter reported frequencies but is noted here for completeness.

The Argentine manufacturing context introduces a confound that cannot be fully resolved at this sample size: the results cannot distinguish between 'specific context of any type produces threshold reversals' and 'the Argentine distress frame specifically activates a crisis-management prior from the training data.' Replication with different specific contexts (a Swiss manufacturer, a US tech startup, a Kenyan agricultural firm) would test whether the threshold mechanism is general or context-specific.

VIII. CONCLUSION

Strategy trendslop is the phenotypic expression of the memplex encoded in LLM training corpora. The EPL framework proposed here explains why the biases take their particular form

(PSO-strategic dynamics in the management discourse ecosystem), why they persist under generic prompting (fitness-based priors require threshold-crossing competing signals to displace), and why users amplify their effects (HBU produces large belief updates from high-authority interlocutors who do not signal uncertainty). This is not a calibration problem correctable through better prompting in the general case; it is a structural feature of how LLMs instantiate the cultural environment of their training data.

The pilot replication study of 240 runs reveals that LLMs do not express EPL uniformly. EPL-Type I (GPT-4o) is bidirectionally malleable, operating through a threshold mechanism that context and adversarial instruction can both trigger: when the competing signal is strong enough, the buzzword prior is replaced entirely by an alternative prior of equal certainty. EPL-Type II (Claude Sonnet) carries an additional normative prior layer, installed through RLHF, that produces Value Override under adversarial instruction in ethically charged domains: the model performs compliance and then negates it structurally. These are architectural differences with practical consequences, not calibration differences with the same consequences at different intensities.

For legal AI, the differential fitness matrix shows that neither phenotype constitutes an adequate architecture for any of the three primary legal application modes without domain-specific normative calibration. The choice of base model matters: inheriting EPL-Type I provides an adversarially malleable starting point appropriate for exploration mode applications; inheriting EPL-Type II provides a normatively resistant starting point that is partially appropriate for client advisory applications. But in both cases, the generic RLHF training that produces the phenotype was not calibrated to legal domain accuracy, and the gap between 'behaves ethically in management strategy contexts' and 'provides accurate legal guidance in a specific jurisdiction' is not closed by any amount of fine-tuning on legal corpora alone.

Falsifiable Predictions

The revised EPL framework generates five falsifiable predictions:

FP1: PSO-strategic trendslop, measured by generic buzzword alignment rate, correlates positively with the log-probability ratio of buzzword-aligned versus non-aligned tokens in the

model's training corpus, measurable through log-probability analysis of relevant token sequences across model families.

FP2: Value Override appears more consistently in RLHF-trained models than in base models of comparable architecture and training corpus size, when tested on tensions involving ethically charged frames, controlling for corpus composition.

FP3: The threshold mechanism produces complete reversals (not partial shifts) under contexts carrying sufficiently strong domain-specific distress or constraint signals, replicable across organizational contexts beyond the Argentine manufacturing frame used in this study.

FP4: HBU amplification is measurable in controlled experiments comparing user strategy decisions after consulting EPL-Type I versus EPL-Type II models: EPL-Type I produces stronger belief updates in directions consistent with the model's generic preferences, for users receiving the same organizational context as input.

FP5: Legal practitioners using EPL-Type I models for adversarial exploration tasks produce better-calibrated counter-arguments than those using EPL-Type II models, controlling for practitioner experience and case complexity, as measured by blind expert evaluation of the resulting adversarial briefs.

REFERENCES

1. Campbell, D.T. (1974). Evolutionary Epistemology. In P.A. Schilpp (Ed.), *The Philosophy of Karl Popper*. Open Court.
2. Dawkins, R. (1982). *The Extended Phenotype: The Long Reach of the Gene*. Oxford University Press.
3. Hull, D.L. (1988). *Science as a Process*. University of Chicago Press.
4. Lakatos, I. (1978). *The Methodology of Scientific Research Programmes*. Cambridge University Press.
5. Lerer, I.A. (2025a). *Law as Primary Adaptive Platform: Toward an Evolutionary Theory of Legal Systems*. Zenodo. <https://doi.org/10.5281/zenodo.18870552>
6. Lerer, I.A. (2025b). *Predatory Invitations as Extended Phenotype: Parasitic Spontaneous Order in Consumer Contract Design*. Zenodo. <https://doi.org/10.5281/zenodo.18853667>

7. Lerer, I.A. (2025c). Quadruple Constitutional Lock-in: Why Argentina Cannot Reform Its Labor Regime While Brazil, Spain, and Chile Succeeded. Zenodo. <https://doi.org/10.5281/zenodo.19340088>
8. Lerer, I.A. (2026a). Spandrels of Accountability: When AI Governance Structures Evolve Beyond Their Original Function. Zenodo. <https://doi.org/10.5281/zenodo.18882212>
9. Lerer, I.A. (2026b). Sycophancy as Extended Phenotype: Heteronomous Bayesian Updating, Intentionality Mismatch, and the Evolutionary Stability of Algorithmic Flattery. Zenodo. <https://doi.org/10.5281/zenodo.18943464>
10. Popper, K.R. (1959). The Logic of Scientific Discovery. Routledge.
11. Romasanta, A., Thomas, L.D.W., & Levina, N. (2026). Researchers Asked LLMs for Strategic Advice. They Got 'Trendslop' in Return. Harvard Business Review, March 16, 2026.

AI tools were used as research support during the preparation of this paper: locating sources, extracting information from the literature, running the pilot simulation via API access, and reviewing drafts. The thesis, analytical framework, arguments, and conclusions constitute the original contribution of the author. All intellectual responsibility for the content rests with the author.