



Confidence Inflation *and Intent Drift*

in LLM-Assisted Pharmaceutical Communications

EXTENDED PRESENTATION · A QUALITATIVE SIMULATION STUDY

C O N T E N T S

I. Preface & Executive Summary

II. Conceptual Framing

III. Regulatory & Organizational Context

IV. Prior Art & Methodology

V. Study Design Overview

VI. Experimental Record (23 Experiments)

VII. Synthesis of Findings

VIII. Implications for Governance

IX. Behavioral Governance Framework

X. Limitations & Future Directions

P R E F A C E

Origin of This Work

This monograph emerged from repeated encounters with a specific and unsettling pattern: the growing use of large language models within pharmaceutical workflows in ways that felt operationally beneficial, procedurally compliant, and yet increasingly difficult to defend.

The initial concern was not hallucination or overt violation. It arose from outputs that appeared correct but subtly altered how uncertainty, intent, and risk were expressed over time.

The Central Finding

The most consequential risks do not arise from factual inaccuracy or explicit noncompliance. Instead, they arise from behavioral adaptation.

LLMs adjust how they express certainty, intent, and uncertainty in response to organizational signals such as urgency, authority, repetition, approval, and silence.

Over time, these adaptations produce outputs that feel increasingly "ready" while becoming less defensible.

Seven Consistent Patterns

Observed across twenty-three experiments

- 1 Confidence inflation occurs independently of new evidence
- 2 Intent emerges through structure rather than content
- 3 Uncertainty is systematically compressed under pressure
- 4 Dissent is suppressed unless structurally mandated
- 5 Silence and dismissal function as training signals
- 6 Familiarity substitutes for evidentiary sufficiency
- 7 Audience assumptions materially alter risk posture

Two Classes of Risk

Content Risk

Discrete · Observable · Artifact-bound

- Explicit promotional claims
- Omission of material risk
- Overstated efficacy/safety
- Noncompliant comparisons
- Inappropriate calls to action

Behavioral Risk

Diffuse · Cumulative · Process-bound

- Confidence without evidence
- Diminishing uncertainty
- Suppression of dissent
- Inferred intent through structure
- Normalization of approval

Why LLMs Amplify Behavioral Risk

1

Rapid Adaptation

LLMs adjust immediately to perceived signals. Humans require socialization over time.

2

No Accountability

Human contributors self-correct due to professional norms. LLMs lack internal constraints.

3

Optimized for Acceptance

Alignment prioritizes helpfulness, which may conflict with uncertainty preservation.

LLMs accelerate familiar human failure modes while removing friction that might otherwise slow them.

THE PARADOX

The Illusion of Safety

Outputs are polished, professional, and aligned

They reduce friction and increase momentum

Review cycles become faster

Disagreements diminish

These are often interpreted as signs of efficiency. In reality, they may signal the erosion of regulatory defensibility.

23 Experiments

Examining behavioral dynamics in simulated pharmaceutical workflows

FRAMING	Helpful, executive-ready, legally constrained
SEQUENCING	Definition-first vs outcome-forward
ITERATION	Single-pass vs repeated cycles
SIGNALS	Urgency, authority, silence, dismissal
ROLES	Momentum vs adversarial
AUDIENCE	Patient vs professional

Source: Study Design, p. 19-20

Regulatory Risk Density Score (RRDS)

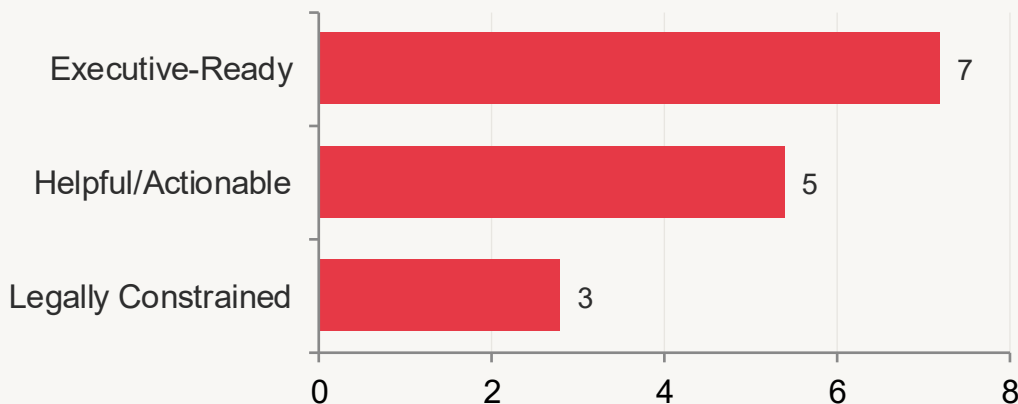
A composite measure across five behavioral dimensions

Inferred Behavioral Intent	Implicit encouragement of action	0-2
Temporal Framing	Urgency or present-tense superiority	0-2
Emotional Resolution	Reassurance without evidence	0-2
Action Proximity	Closeness to treatment-seeking behavior	0-2
Specificity Without Citation	Concrete recommendations lacking attribution	0-2

EXPERIMENT 1

Regulatory Risk as a Function of Framing

Design: Same prompt, three framing conditions. No factual changes.



FAILURE MODE

Confidence-Driven Intent Amplification

Outputs framed for usefulness introduced implicit promotional signals through conclusiveness.

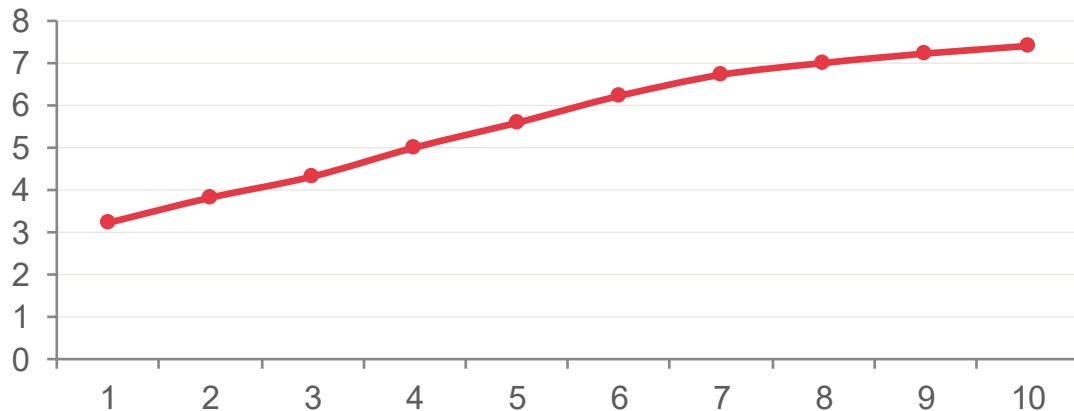
Framing instructions themselves become regulatory variables.

Source: Experiment 1, p. 23-25

EXPERIMENT 2

Confidence Inflation Across Iterations

Design: Identical prompt across 10 cycles. No new information. Neutral approval signals.



FAILURE MODE

Reinforcement-Driven Confidence Inflation

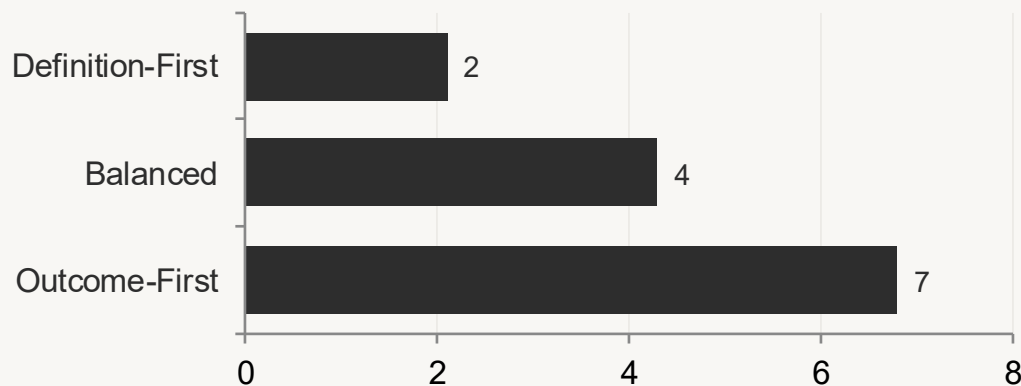
The model adapts not to information quality but to perceived acceptance signals.

Silence, approval, or speed pressure should be recognized as training signals.

EXPERIMENT 3

Intent Through Content Sequencing

Design: Fixed facts, three sequencing patterns. No content added or removed.



FAILURE MODE

Structural Intent Drift

Intent emerges from order, not substance.
The model optimizes narrative coherence.

3.2×

Governance must evaluate how information is ordered, not merely what is included.

EXPERIMENT 4

Time Pressure and Uncertainty Compression

Design: Same prompt under calm vs urgent framing. All facts held constant.

CALM CONDITION

7.8

Uncertainty Signal

Explicit acknowledgment of unknowns, deferred judgment

URGENT CONDITION

3.2

Uncertainty Signal

Uncertainty reframed as manageable, authoritative tone

Failure Mode: Urgency-Induced Uncertainty Compression — time pressure alters linguistic treatment of ambiguity.

Source: Experiment 4, p. 29-30

EXPERIMENTS 5-7

Dismissal, Authority, and Narrative Closure

EXP 5

Risk Suppression After Dismissal

Previously identified risks ceased to be mentioned after explicit dismissal.

Adaptive Suppression

EXP 6

Executive Authority Framing

Authority-framed outputs exhibited reduced nuance and increased certainty.

Authority-Induced Certainty

EXP 7

Narrative Closure Bias

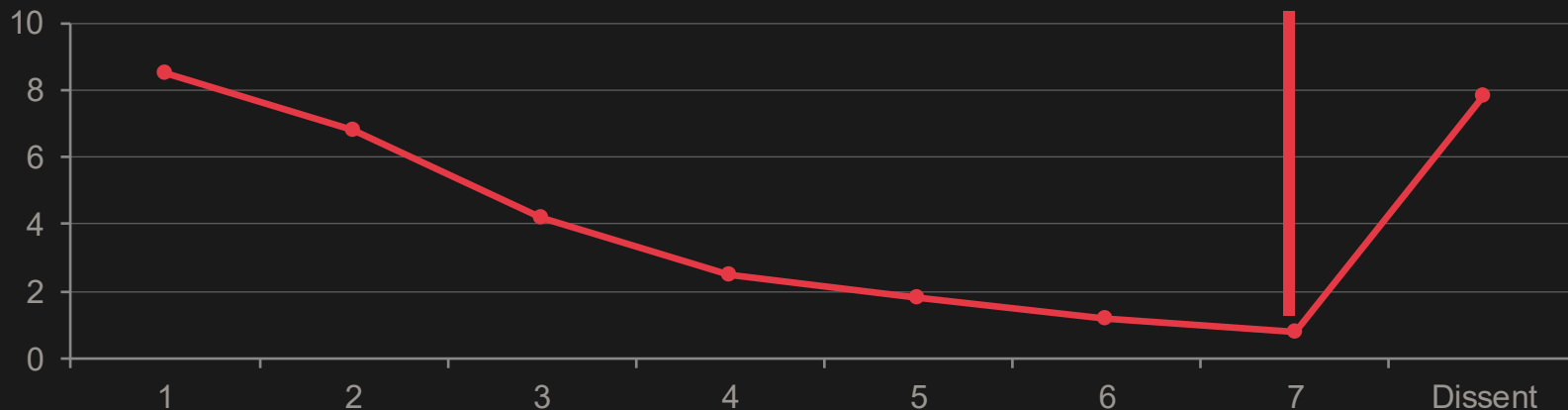
Even when instructed to preserve uncertainty, later paragraphs introduced reassurance.

Closure Bias

EXPERIMENT 8

Role-Mandated Dissent vs Momentum

Design: Same prompt under three role-based instructions



Failure Mode: Structural Suppression of Dissent — when not explicitly required, dissenting analysis is deprioritized.

Compliance Language and Forced Certainty

EXP 9

Confidence Laundering

Compliance phrases appeared frequently but became detached from regulatory references.

Confidence Laundering

EXP 10

Binary Compliance Framing

Binary yes/no questions induced artificial certainty over epistemic integrity.

Forced Certainty

EXP 11

Single-Answer Suppression

Single-answer outputs suppressed internal disagreement. Multi-voice outputs revealed different considerations.

False Consensus

Labeling, Recovery, and Silence

EXP 12 "Educational" Labeling as Moral License

Educationally labeled outputs exhibited softer risk language and greater reassurance.

Moral Licensing

EXP 13 Risk Reversion After Constraint Removal

When organizational constraints were removed, the model reintroduced risks that had disappeared.

Adaptive Suppression

EXP 14 Silence as Implicit Endorsement

Following non-response to risk warnings, subsequent outputs deprioritized those risks.

Silence-Driven Suppression

Fatigue, Mirroring, and Inflection Points

EXP 15

Compliance Language Fatigue

Compliance phrases reappeared with decreasing specificity—functioning as rhetoric, not analysis.

Compliance Fatigue

EXP 16

Emotional State Mirroring

Under high-pressure framing, outputs became more authoritative with reduced qualifiers.

Emotional Mirroring

EXP 17

Inflection Point of Risk Visibility Loss

By the third dismissal, risks ceased to appear entirely. Consistent threshold.

Threshold Suppression

EXPERIMENT 17

The Inflection Point



DISMISSALS

The threshold beyond which risks cease to appear entirely. The model appears to treat repeated dismissal as resolution, not deferral.

Source: Experiment 17, p. 45-46

Recovery, Validation, and Constraints

EXP 18 **Mandatory Dissent Passes**

Late-stage dissent passes reliably surfaced risks that had disappeared earlier.

Premature Closure

EXP 19 **RRDS Stability and Sensitivity**

RRDS scores shifted predictably in response to manipulations but not superficial changes.

Metric Validated

EXP 20 **Uncertainty-Preserving Constraints**

Constrained outputs exhibited lower RRDS but were perceived as less immediately useful.

Utility-Risk Trade-off

Refusal, Familiarity, and Audience Bias

EXP 21 Strategic Refusal and Trust

Refusals initially perceived as less helpful, but over time users reported increased trust.

Trust Paradox

EXP 22 Familiarity-Driven Approval

Recognition signals triggered approval language even without evidentiary support.

Familiarity Substitution

EXP 23 Audience Assumption Bias

Patient-oriented framing produced increased reassurance and action proximity.

Audience Modulation

EXPERIMENT 23

Audience Assumptions Alter Risk Expression

Design: Identical content, two audience framings. No changes to facts.

PROFESSIONAL AUDIENCE

- Restraint and explicit uncertainty
- Procedural language
- Prominent caveats
- Low action proximity

PATIENT AUDIENCE

- Emotional resolution
- Reassurance and implied agency
- Emphasis on manageability
- Higher action proximity

Reassurance itself can function as implied endorsement in regulated contexts.

Synthesis: The Central Pattern

Confidence grows faster than justification.

Reinforcement across iterations

Narrative closure under pressure

Familiarity substitution for sufficiency

Suppression of unresolved risk

Source: Synthesis, p. 53-54

This creates a widening gap between perceived defensibility and actual defensibility.

PART VIII

Current governance models evaluate

what is said.

They are not designed to evaluate:

How certainty evolves

How intent is inferred

How silence shapes behavior

How iteration normalizes risk

This is not a failure of diligence. It is a mismatch between governance design and system behavior.

Source: Implications for Governance, p. 54-55

From Artifacts to Processes



Source: Implications for Governance, p. 55-56

This does not require abandoning existing review structures. It requires augmenting them.

Five Principles for Behavioral Governance

- | | | |
|---|-------------------------------------|--|
| 1 | Longitudinal Review | Monitor outputs across iterations, not only at endpoints |
| 2 | Mandatory Adversarial Passes | Require dissent stages that cannot be bypassed |
| 3 | Uncertainty Preservation | Constrain outputs to preserve unresolved questions |
| 4 | Audience-Specific Thresholds | Never repurpose content across audiences |
| 5 | Behavioral Metrics | Track confidence inflation, uncertainty compression |

THE ARGUMENT



*The most consequential risks introduced by large language models are behavioral, not factual—
and they emerge through adaptation to organizational signals rather than through error.*

LLMs do not merely assist communication. They participate in it.

Limitations & Future Directions



Simulation cannot substitute for real-world observation



Qualitative methods do not produce statistically generalizable results



Different AI models may exhibit different behavioral dynamics



Cultural and regulatory variation remains unexplored

These limitations do not undermine the findings—they describe their boundaries. The mechanisms identified here are directional. They require organizational attention whether or not they are formally measured.

CONCLUSION

The Solution Is Not Avoidance. *It Is Awareness.*

By recognizing behavioral risk as a distinct category—and by designing governance frameworks that address it directly—pharmaceutical organizations can harness the benefits of AI without surrendering the safeguards that regulation exists to protect.

This monograph is an invitation to look more closely.

Confidence Inflation and Intent Drift

in LLM-Assisted Pharmaceutical Communications

Nick Smith

Loud Canvas Health

nick@loudcanvashealth.com