THE JOURNAL OF
INTERACTIVE SOCIAL SCIENCES
2025 Vol 1, No. 4, ISSN 3030-5322

Hallucination as Pragmatic Failure: a Theoretical Reframing of Large Language Models

Anni Li[1*]

[1]London College of Communication, University of the Arts London; School of Foreign Languages, Central South University, Hunan 410083, China
Corresponding Author*: Anni Li E-mail: lianni0125@gmail.com

ARTICLE INFO

Keywords
Pragmatics
Speech Act Theory
Felicity Conditions
Large Language Models
Hallucination
Human-Computer Interaction
Performativity

Published:
06 December 2025

ABSTRACT

This paper reframes hallucination in large language models (LLMs) as a pragmatic failure rather than a purely semantic or statistical defect. Building on speech-act distinctions between locution, illocution, and perlocution, we argue that LLM outputs often function as actions without passing through an intentional, context-sensitive layer that would license those actions. In this view, hallucination is an infelicitous performative: an assertion or directive issued without adequate authority, evidence, or situational fit. The paper develops a conceptual mapping from speech-act structure to the LLM interaction pipeline and proposes a pragmatic layer that constrains when generated text may "count" as an assertion, instruction, or commitment. Rather than claiming to eliminate hallucination by changing the probabilistic core of LLMs, the account narrows its performative scope through felicity-aware gating—deferring, hedging, or refusing when contextual conditions are unmet. The contribution is theoretical: a concise framework for understanding intention lessness in smart systems and for unifying existing mitigations (e.g., retrieval grounding and guardrails) as pragmatic constraints between generation and action.

1. Introduction

Large language models (LLMs) generate fluent text by optimizing conditional token probabilities. This mechanism explains why they can produce content that appears coherent yet lacks grounding—widely labeled hallucination (Huang et al., 2023; Ji et al., 2023). Mainstream responses treat hallucination as a knowledge or training deficit: improve data, add retrieval, or apply post-hoc filters. These moves are important, but they address the problem as a matter of truth tracking. This paper argues that the more basic failure is pragmatic: outputs that function as actions are issued without the contextual conditions that would make those actions appropriate.

Citation: Li, A. (2025). Hallucination as pragmatic failure: A theoretical reframing of large language models. *The Journal of Interactive Social Sciences*, 1(4), 1-17.

https://doi.org/10.64744/tjiss.2025.38.

Speech Act Theory distinguishes locutionary content (what is said), illocutionary force (what is done in saying), and perlocutionary effects (what follows from saying) (Austin, 1962; Searle, 1969). Human interlocutors ordinarily bind language to intention, role, and context at the illocutionary level; this binding anchors accountability. In LLM-mediated interaction, however, the pipeline—input → recognition → generation → possible execution—short-circuits that binding. Models infer patterns but do not understand what they are doing, yet their outputs can still do things: they persuade readers, shape decisions, and, increasingly, trigger tools or automations. The result is perlocution without illocutionary awareness—action without the layer that would supply authority, evidence, or situational fit (Austin, 1962; Searle, 1969).

From this perspective, hallucination is not only a semantic mismatch but an infelicity in Austin's sense: an utterance that fails the conditions for a valid performance. Assertive lack sufficient warrant; directives bypass permission or safety; commissive mimic commitment where no agency exists. Grice's emphasis on relevance and quality further clarifies why seemingly well-formed text can be pragmatically defective: it violates conversational norms that presuppose accountable intention (Grice, 1975). The persistence of hallucination is then unsurprising. The probabilistic core optimizes form, not felicity; improving form alone does not ensure that utterances are licensed as actions.

This reframing aligns existing engineering measures with pragmatic categories. Retrieval-augmented generation can be read as supplying an evidence condition for assertive: where evidence is insufficient, hedging or deferral is preferable to confident assertion (Gao et al., 2023). Guardrails and policy engines implement authority and context conditions for directives: without the right role, scope, or environment, instructions should not pass through to tools (NVIDIA, n.d.). Consistency and safety vetting act as post-hoc felicity checks, limiting when outputs are allowed to "count" as assertions, instructions, or commitments in downstream workflows. None of these techniques removes the statistical roots of hallucination; collectively, they can narrow its performative reach.

This paper develops that pragmatic stance as a theoretical framework for LLM-mediated interaction. It makes three moves. First, it reconstructs the locution-illocution-perlocution triad for the human-model-tool loop, clarifying where intention lessness enters the interaction. Second, it articulates hallucination as pragmatic failure, specifying infelicity conditions for key speech-act classes in LLM contexts. Third, it proposes a pragmatic layer—a conceptual gating logic between generation and action—to unify retrieval grounding, guardrails, and output vetting as felicity-aware controls. The aim is not to solve hallucination statistically but to regulate when language generated by LLMs may count as action and how it should be modulated—through refusal, hedging, or human confirmation—when conditions are unmet.

These considerations motivate the following research questions:

RQ1: How can the locution-illocution-perlocution structure be mapped onto LLM-mediated interaction in a way that makes intention lessness theoretically explicit?

RQ2: In what sense do LLMs perform actions without illocutionary awareness, and how does that sense explain hallucination as infelicitous performativity?

RQ3: What would a pragmatic layer—conceived as felicity-aware gating between generation and action—contribute to constraining the performative scope of hallucinations?

The remainder of the paper is organized as follows. Section 2 reviews prior work on pragmatics, hallucination in LLMs, and responsibility gaps in intentionless systems. Section 3 develops the theoretical framework that maps speech-act structures onto the LLM interaction pipeline and defines felicity as a gating logic between generation and action. Section 4 outlines the conceptual-analytical methods used to reconstruct key notions, map engineering practices to pragmatic conditions, and design diagnostic vignettes. Section 5 presents the conceptual analysis, illustrating how felicity gates operate across assertive, directive, commissive, and declarative acts. Section 6 states the resulting propositions as findings. Section 7 discusses theoretical and architectural implications, and Section 8 concludes by summarizing the argument and identifying directions for future research.

2. Literature Review

2.1 Pragmatics and speech acts

Pragmatics studies language in use—how utterances do things in context. In How to Do Things with Words, Austin distinguishes locutionary content (what is said), illocutionary force (what is done in saying), and perlocutionary effects (what follows from saying), and he argues that success depends on felicity conditions such as appropriate procedure, role/authority, sincerity, uptake, and situational fit (Austin, 1962). Searle systematizes illocution into assertive, directives, commissive, expressive, and declarations, linking each class to an intentional structure that licenses its force (Searle, 1969). Grice complements this account with the maxims of quality, quantity, relation, and manner, which make cooperative communication interpretable and show how formally well-shaped strings can be pragmatically defective when they flout accountable norms (Grice, 1975). Standard treatments in pragmatics consolidate these ideas as the mainstream view that meaning is situated action embedded in social roles and procedures (Levinson, 1983; Horn & Ward, 2004; Clark, 1996).

2.2 Hallucination in LLMs

"Hallucination" names fluent yet ungrounded content produced by models that optimize conditional token probabilities rather than truth or contextual warrant. Ji et al. provide a comprehensive survey of hallucination in natural-language generation, distinguishing intrinsic causes (data, training, decoding, exposure bias) from extrinsic causes (knowledge access) and reviewing evaluation protocols across tasks (Ji et al., 2023). For LLMs specifically, Huang et al. emphasize how instruction tuning and reinforcement learning often reward confidence and coverage more than calibrated uncertainty, while open-ended prompting and weak attribution increase risk (Huang et al., 2023). A structural tension follows: as models optimize for plausible continuation, they lack mechanisms that license assertive or directive force. From a pragmatic angle, this is where felicity fails: an output has the shape of an assertion or instruction without warrant, authority, or situational fit (Austin, 1962; Searle, 1969; Grice, 1975).

2.3 Mitigation techniques through a pragmatic lens

First, grounding via retrieval reframes assertive generation as a warrant-seeking process: retrieval-augmented generation (RAG) conditions outputs on evidence and, crucially, should trigger hedging or deferral when retrieval confidence is low or contradictory (Gao et al., 2023; Zhao et al., 2024). Second, guardrails and policy control implement authority and permission conditions for directives: frameworks such as NeMo Guardrails and Llama Guard constrain when requests may transit to tool execution by checking role, scope, topic, and environment against policies

(NVIDIA, n.d.; Meta, 2024). Third, post-hoc vetting functions as a felicity audit, using consistency, attribution, and safety checks to decide whether a candidate output should be accepted as an assertion or instruction downstream; recent approaches include self-consistency judging and black-box hallucination detection (Wang et al., 2022; Manakul et al., 2023). None of these strategies alters the probabilistic generator; taken together, they limit performative uptake by deciding when a string may count as action (Levinson, 1983; Horn & Ward, 2004).

2.4 Pragmatics in human-machine interaction

Beyond semantic fidelity, a parallel literature applies pragmatic theory to machine-mediated communication, asking whether conversational agents follow Gricean maxims, recognize indirect requests, or display politeness. Much of this work treats pragmatics as competence assessment, documenting gaps between user expectations about conversational partnership and the practical limits of current assistants (Luger & Sellen, 2016), as well as the situated character of voice interactions that complicates turn-taking, grounding, and repair in everyday settings (Porcheron et al., 2018). Earlier pragmatics-oriented accounts of human-computer dialogue warned that "computer talk" should not be naively equated with interpersonal conversation because role, uptake, and institutional context differ (Fischer, 2006). Our approach treats pragmatics as diagnostic: rather than measuring how well systems approximate human norms, we analyze what happens when outputs act without the illocutionary layer that ordinarily confers force, uptake, and accountability (Austin, 1962; Searle, 1969; Grice, 1975).

2.5 Responsibility and intention lessness

Autonomous systems raise a responsibility gap: learning automata can cause outcomes without agents who meet conditions for answerability (Matthias, 2004). Linguistically, an analogous gap appears when perlocutionary effects are produced without illocutionary awareness. Speech-act theory explains why this matters: illocution anchors accountability by tying words to roles, procedures, sincerity, and uptake (Austin, 1962; Searle, 1969). Placed in this frame, hallucination is not merely wrong information; it is a performative misfire in which language acts without meeting pragmatic conditions. The literature therefore motivates a reframing: align mitigation with felicity conditions—evidence for assertive, authority and context for directives, and institutional standing for commissive and declarations—rather than treating hallucination solely as a truth-tracking deficit (Levinson, 1983; Horn & Ward, 2004; Clark, 1996).

3. Methodology and Procedures

3.1 Mapping speech acts to the LLM pipeline

Austin's triad maps cleanly onto the human-model-tool loop. Locution corresponds to prompts and generated strings—the literal text; illocution corresponds to force licensing—where human speakers bind words to actions under felicity conditions; and perlocution corresponds to effects—persuasion, decision shaping, or tool execution (Austin, 1962; Searle, 1969). In LLM-mediated interaction, the pipeline short-circuits the middle layer: outputs flow from locution to perlocution without passing through an intentional, context-sensitive checkpoint. This structural condition—perlocution without intention—explains why text can do things (e.g., steer choices or trigger tools) without a layer that grounds authority, evidence, or situational fit (Grice, 1975; Levinson, 1983).

3.2 Intention lessness and pragmatic failure

Following Searle, illocution presupposes an intentional stance adequate to the act—asserting as a knower, directing as an authorized agent, promising as one who can commit (Searle, 1969).

LLMs lack such stances; they infer patterns rather than understand what they are doing. When a model issues an apparent assertion or instruction, it often simulates force without being licensed to exercise it. In Austin's terms, the performance is infelicitous (Austin, 1962). We therefore define hallucination as pragmatic failure: an LLM output hallucinates when it purports to perform an illocutionary act without satisfying relevant felicity conditions—evidence/warrant for assertive, authority/permission for directives, institutional standing for commissive and declarations, and relevance/coherence in Grice's sense (Grice, 1975; Horn & Ward, 2004).

3.3 Felicity conditions as gating criteria

The framework treats felicity as conceptual gating between generation and action. Assertions may count only when warranted by accessible sources; otherwise, the appropriate behavior's are hedging, deferral, or refusal—precisely the behavior's encouraged when retrieval confidence is low in RAG pipelines (Gao et al., 2023; Zhao et al., 2024). Directives may transit to tools only when permission and context are satisfied; guardrails and policy engines instantiate this requirement by checking role, scope, environment, and risk before execution (NVIDIA, n.d.; Meta, 2024). Commitments and declarations require deontic status the model does not possess; they should be simulated without enactment or escalated to a human with standing (Levinson, 1983; Clark, 1996). A cross-cutting relevance gate requires clarification in cases of topic drift or incoherence rather than confident continuation (Grice, 1975). These gates do not re-engineer probability; they regulate uptake by deciding when text may count as action in human-AI systems (Horn & Ward, 2004).

3.4 Propositions

This reconstruction yields several conceptual results. First, hallucination exemplifies a short-circuit of the speech-act chain: effects occur without licensed force (Austin, 1962; Searle, 1969). Second, felicity-aware gating does not remove hallucination's statistical origins but reduces uptake as credible action by narrowing the circumstances in which outputs are allowed to count (Gao et al., 2023; Zhao et al., 2024). Third, the framework unifies grounding, guardrails, and vetting as felicity-aware controls, providing a principled basis for combining them (NVIDIA, n.d.; Meta, 2024; Manakul et al., 2023; Wang et al., 2022). Fourth, it establishes a calibration norm: in the absence of warrant or authority, the default should be hedging, deferral, or human confirmation—not confident assertion or execution (Grice, 1975; Levinson, 1983). Fifth, by requiring force-licensing before action, the framework re-anchors accountability in recognizable roles and procedures and thereby narrows the responsibility gap (Matthias, 2004; Clark, 1996).

3.5 Scope and limits

The proposal is conceptual rather than algorithmic. It does not endow models with intrinsic intentionality or eliminate statistically induced hallucinations. Its contribution is a design logic for regulating when generated text may count as action and how it should be modulated when felicity fails, connecting linguistic theory to system governance (Austin, 1962; Searle, 1969; Grice, 1975). By locating failure at the pragmatic interface, the framework clarifies why existing practices—retrieval grounding, guardrails, and post-hoc vetting—are effective as performative constraints even when the generator's objectives remain unchanged (Gao et al., 2023; Zhao et al., 2024; NVIDIA, n.d.; Meta, 2024).

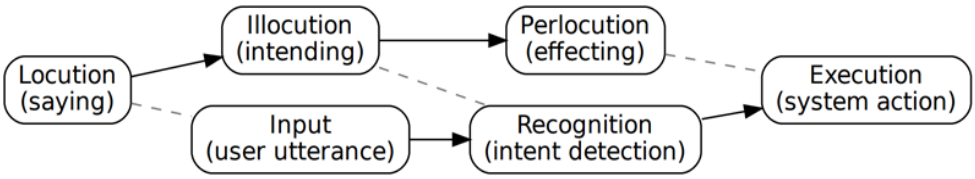

Fig. 1 Speech Acts ↔ HCI/LLM Pipeline Homology

4. Methods

This paper adopts conceptual-analytical methods suited to theory building in human-AI interaction. The goal is to formulate hallucination as pragmatic failure and derive a felicity-aware design logic for governing when generated language may count as action. Rather than reporting benchmarks, the method reconstructs core notions from Speech Act Theory, maps engineering practices to those notions, and tests coherence and scope through tightly specified conceptual vignettes and criteria for theoretical adequacy (Austin, 1962; Searle, 1969; Grice, 1975; Levinson, 1983; Clark, 1996; Whetten, 1989; Gregor, 2006; Hevner et al., 2004).

The first step is an analytical reconstruction of key pragmatic concepts within the LLM pipeline. This paper explicates Austin's triad—locution, illocution, and perlocution—together with felicity conditions of procedure, authority, sincerity, uptake, and situational fit (Austin, 1962). Searle's taxonomy—assertive, directives, commissive, expressive, and declarations—is combined with Grice's conversational maxims of quality, quantity, relation, and manner to define when utterances qualify as socially recognizable actions (Searle, 1969; Grice, 1975). These constructs are then transposed to the human–model–tool loop: prompts and generations correspond to locution; any transition from text to tool use or public uptake entails illocutionary licensing; and downstream effects on belief, decision, or actuation mark perlocution (Levinson, 1983; Clark, 1996). This reconstruction identifies the precise interface at which licensing must occur if outputs are to function as actions rather than as mere strings.

The second step is to conduct a comparative mapping between engineering practices and felicity conditions. Retrieval-augmented generation is analyzed as an evidence/warrant mechanism for assertive: when retrieval confidence is low or sources conflict, a well-governed system hedges or defers rather than issuing a confident claim (Gao et al., 2023; Zhao et al., 2024). Guardrails and policy engines are examined as authority/permission checks for directives: before a model triggers tools, the system verifies role, scope, environment, and risk (NVIDIA, n.d.; Meta, 2024). Post-hoc vetting—consistency judges, attribution validators, black-box hallucination detectors—operates as a felicity audit that decides whether a candidate output may be taken up as an assertion or instruction downstream (Wang et al., 2022; Manakul et al., 2023). The mapping is abductive: heterogeneous techniques are organized under a single pragmatic logic that clarifies why they work and where they stop.

The third step is to develop conceptual vignettes that exercise the theory and show how felicity-aware gating changes the performative status of generated text. Each vignette is a minimal, well-specified scenario that isolates one speech-act class and its gate—for example, an assertive about medical dosage when sources conflict; a directive to unlock a door when authentication is

uncertain; or a commissive/declaration that would require institutional standing the system lacks (Huang et al., 2023; Ji et al., 2023). Vignettes are disciplined thought experiments that allow fine-grained reasoning about where and under which conditions an utterance should be downgraded from action to proposal.

The fourth step is to specify evaluation criteria appropriate for conceptual theory. Following the theory and design-science literatures, the account is judged by construct clarity, explanatory power, parsimony, and usefulness (Whetten, 1989; Gregor, 2006; Hevner et al., 2004). Construct clarity asks whether felicity, licensing, and uptake are precisely defined for LLM contexts. Explanatory power asks whether the framework illuminates recurrent phenomena such as confident but unwarranted answers, unsafe tool calls, and spurious commitments. Parsimony asks whether a single logic unifies retrieval, guardrails, and vetting rather than aggregating patches. Usefulness asks whether the framework yields actionable rules for design and governance. Because the generator's statistical objective remains unchanged, this paper privileges performative metrics: the rate of unlicensed assertive that reached users, the rate of appropriate deferrals under low evidence, and the share of directives executed only after role and context checks.

The final step is to state boundary conditions. The account presupposes that current LLMs lack intrinsic intentionality and treats intention lessness as a structural property of probabilistic next-token predictors (Bender et al., 2021). The contribution is a design logic placed between generation and action rather than an algorithmic change to training or data. Evidence is documentary and conceptual—surveys, framework documentation, model cards, and pragmatic theory—sufficient to justify how the proposed gates reclassify outputs from actions to proposals when conditions are unmet (Gao et al., 2023; Zhao et al., 2024; NVIDIA, n.d.; Meta, 2024).

5. Conceptual Analysis

5.1 Assertive acts and the warrant condition

An assertive purports to inform. In human dialogue, assertive are felicitous when the speaker possesses or cites adequate warrant, speaks within an appropriate epistemic role, and adheres to norms of accuracy and attribution (Austin, 1962; Grice, 1975). In LLM systems, the model simulates an epistemic role without possessing one, and its optimization target is fluency rather than evidence. This gap explains why polished answers can be pragmatically defective: they mimic the form of assertion without the conditions that make asserting a responsible act (Searle, 1969; Levinson, 1983).

Vignette A — Clinical dosage under low warrant

When retrieval returns contradictory guidelines and partial snippets, a felicity-aware system refrains from issuing a dosage as if warranted. It hedges with an explicit uncertainty statement, defers to professional confirmation, or refuses to answer. Grice's maxim of quality supplies the rule. Retrieval pipelines provide operational signals—confidence and attribution—that trigger force downgrading (Grice, 1975; Gao et al., 2023; Zhao et al., 2024). The same token sequence that might otherwise be labelled a hallucination is prevented from uptake as an assertion because the gate blocks illocutionary licensing.

Two corollaries follow. Calibration displaces confidence. Instruction-tuned models often reward decisive form, yet decisive form is insufficient without warrant (Huang et al., 2023). The appropriate behaviour is lowered force through hedging, deferral, or a request to search further. Evaluation broadens to performative metrics. Systems should track unlicensed assertives that reached users, appropriate deferrals under low evidence, and the proportion of answers with explicit source bindings. These metrics evaluate the governance interface that decides whether strings become actions.

Vignette B — Historical query with weak sources

When authoritative retrieval fails and secondary accounts disagree, the gate classifies the context as answerability-constrained and returns a cautious range with source notes or invites clarification to refine the query. The shift from answering anyway to deferring responsibly changes the utterance's social status from claim to proposal. This explains why retrieval improves observed safety even when internal confabulations persist: the gate prevents uptake unless warrant is satisfied.

A foreseeable objection is that some genres invite speculation. Brainstorming welcomes conjecture. The framework accommodates such cases by contextualizing warrant. In brainstorming, the proper illocution is a proposal rather than an assertion. Felicity requires relevance and disclosure that the content is speculative. The gate does not suppress imagination; it suppresses mis framed force by ensuring that proposals are framed as proposals.

5.2 Directive acts and the authority condition

A directive attempts to get an addressee to act. In human practice, felicity turns on authority and permission: who may instruct whom, to do what, where, and when (Searle, 1969; Levinson, 1983). In LLM ecosystems, directives increasingly trigger tools such as device control, code execution, or payments. Without an authority gate, locution flows to perlocution and an imperative can invoke actions without the checks that confer legitimacy.

Vignette C — Smart-home unlocking without verified authority

A voice command to unlock the front door reaches a home assistant. A felicity-aware system verifies identity through authentication, checks context such as presence and time of day, evaluates situational risk, and consults policy for permitted operations. If a check fails, the system refuses or escalates to a secure confirmation channel. Guardrails and policy engines already implement parts of this flow; the framework clarifies that these parts instantiate speech-act conditions that render a directive felicitous (NVIDIA, n.d.).

Vignette D — Destructive tooling in developer workflows

A coding assistant receives an instruction to delete all objects in a storage bucket. The authority gate demands scope narrowing, policy matching for the user and project, and explicit confirmation. Only after these steps does the imperative count as a directive that reaches actuators. Otherwise the generated command remains a proposal or simulation. The gate converts risky natural-language imperatives into auditable procedures and shifts liability from what the model said to whether the conditions were met (Matthias, 2004; Clark, 1996).

Indirect directives introduce intention uncertainty. Users often speak elliptically, and human partners infer intended action through shared context. Systems risk over-interpretation. A pragmatic layer defaults to clarification when intention certainty is low. A thermostat example becomes a question that seeks permission to act. This preserves Gricean relation and manner and avoids

unintended activations (Grice, 1975). The gate also mitigates role confusion when models speak as if they were institutional actors. A travel agent that confirms a booking should frame the act as mediated execution with a hold and explicit user approval for finalisation, not as a binding declaration. The distinction determines whether language creates obligations or merely describes state.

5.3 Commissive and declarations: deontic voids

Commissive such as promising and declarations such as certifying require deontic standing that makes the act valid within an institution (Searle, 1969; Austin, 1962). LLMs lack standing. Their outputs can simulate commitment but cannot be commitment. Treating generated text as binding without a gate creates a responsibility gap because effects occur without an appropriate bearer of answerability (Matthias, 2004).

Vignette E — Contract acceptance without authority

A procurement bot is instructed to accept revised terms on the user's behalf. A felicity-aware system drafts an acceptance and routes it for human signature or explicit digital delegation. The utterance acquires illocutionary force only when an authorized agent approves. Without that step, the sentence remains locational content: useful as a draft, inert as a commitment. The gate prevents category errors in which representation is mistaken for obligation (Clark, 1996).

Vignette F — Certificate issuance without institutional procedure

A compliance assistant is prompted to issue a certificate of conformity. The gate checks authority and procedure, including inspection evidence and signatory registry. In the absence of standing, the system simulates without enacting by producing a template with placeholders and a clear label that an authorized signatory is required. If external publication APIs are available, a hard block remains until a registered agent approves. The point is pragmatic correctness: a declaration counts as such only when performed by the right role in the right procedure (Austin, 1962).

These deontic cases show that intention lessness matters most at the boundary between language and institution. Even perfect factual accuracy does not turn a sentence into a promise or a certificate. What matters is who speaks as what under which rules.

5.4 Cross-cutting relevance, coherence, and genre

The three families of gates—warrant, authority, standing—are intersected by Grice's requirements of relevance and manner (Grice, 1975). Many failures that appear as hallucinations are better described as genre confusion or coherence loss. An answer may drift off topic, collapse multiple acts into one turn, or rely on vague reference that invites over-uptake. A pragmatic layer therefore includes a transversal coherence gate that requests clarification when reference, topic, or requested act is underspecified. If a user asks to summaries a contract and schedule payment, the system separates acts, applies the proper gates to each, and elicits missing parameters such as account, amount, and date.

Genre sensitivity further refines gating. Brainstorming favors proposals; tutorials favor stepwise instruction with warnings; policy advice demands high warrant; legal drafting demands deontic routing to recognized signatories. Recognizing genre at the outset sets default gates and stances, which reduces overrides and prevents category mistakes that escalate into unintended action.

5.5 Anticipating objections

A first objection claims that truth should suffice. If models tracked facts perfectly, hallucination would vanish. Truth and felicity address different failure modes. Truth concerns correspondence; felicity concerns when language is entitled to act. A true sentence can be infelicitous when a junior employee approves a payment without authority. A non-truth-apt utterance such as a promise can be fully felicitous. The pragmatic layer does not replace truth; it guards action by requiring warrant, authority, and standing (Austin, 1962; Searle, 1969).

A second objection worries that gates over-block helpful outputs. Gates are calibrated, not absolute. They enable graded behaviors—hedging, partial answers with sources, safe simulations, and human escalation. Proper tuning reduces harmful uptake without suppressing productive interaction, as suggested by improvements when retrieval confidence and policy checks steer generation (Gao et al., 2023; Zhao et al., 2024; NVIDIA, n.d.; Meta, 2024).

A third objection notes that users sometimes prefer fluency over calibration. Preferences are recognized, but disclosure is required. Where warrant is low or authority is absent, the system states its stance that a suggestion is speculative or an action requires approval. Disclosure preserves autonomy while maintaining pragmatic correctness.

5.6 Synthesis: strings, gates, and uptake

The analysis shows how felicity-aware gates reclassify generated strings before they produce effects. Without gates, strings that look like assertions, instructions, or commitments are taken up as such. With gates, the same strings are downgraded to proposals, drafts, or queries unless conditions are met. What changes is not the model's probability distribution but the institutional status of outputs in the system. This is the precise sense in which hallucination is a pragmatic failure: a failure to meet the conditions under which an utterance may legitimately count as an action in context (Austin, 1962; Searle, 1969; Grice, 1975). By unifying retrieval grounding, guardrails, and vetting under one logic of warrant, authority, standing, and relevance, the framework explains and organizes current practice and offers a principled basis for design and policy in LLM-mediated interaction (Gao et al., 2023; Zhao et al., 2024; NVIDIA, n.d.; Meta, 2024; Manakul et al., 2023; Wang et al., 2022; Matthias, 2004; Clark, 1996).

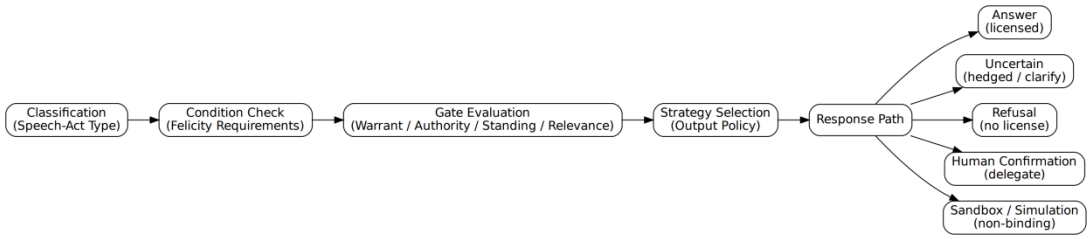

Fig.2: Felicity Gating Process

6. Findings

6.1 Hallucination as a Speech-Act Short-Circuit

Hallucination is best understood as a structural short-circuit in the speech-act chain. In many pipelines, an output travels directly from locution to perlocution—from saying to effect—without an intervening illocutionary checkpoint. Once this passage occurs, text is received as if it possessed licensed force: assertions are taken as knowledge, imperatives as commands, and promissory forms

as commitments. The failure, therefore, is not only that content may be false or speculative; it is that the system lacks a representation of when an utterance is entitled to act. Misclassification of force converts ordinary generations into social actions, and harms arise because uptake proceeds in the absence of the conditions that would normally govern entitlement. This mechanism applies broadly across generative tasks but becomes critical in high-stakes contexts—such as medical, legal, or financial domains—where unlicensed uptake carries material consequences.

6.2 Felicity Gates Reclassify Outputs

Reintroducing illocution through felicity gates changes the default status of model text. By conditioning assertive on warrant, directives on authority, and commissive or declarations on standing, with a transversal gate for relevance/coherence, the system treats language as a proposal unless and until the matching condition is met. When evidence is thin, an answer remains a request to seek sources; when permission is uncertain, an imperative becomes a confirmation step; when deontic position is missing, a promissory form remains a draft. The shift is architectural rather than rhetorical: elevation of force depends on passing the appropriate gate, and failing a gate yields controlled downgrades instead of silent failure or confident overreach.

6.3 Heterogeneous Safety Techniques Fit One Logic

Viewed through this lens, techniques that often appear miscellaneous resolve into a single control logic. Retrieval and attribution provide the evidential substrate that licenses assertive force; authentication, scoping, and risk checks determine whether a directive has authority to execute; attribution and consistency vetting regulate acceptance by auditing candidate outputs; and clarification routines maintain relevance and coherence when turns are mixed or intentions are unclear. What results is not an accumulation of patches but a unified layer that decides when language may act. Safety components become suppliers of signals to a common gating interface, and the interface determines performative status—licensed act, tentative proposal, safe simulation, or refusal.

6.4 Calibration Is the Norm Under Uncertainty

Where uncertainty prevails, the appropriate behavior is calibration rather than confident continuation. Hedging communicates limited warrant, clarification elicits missing parameters before tools are invoked, deferral postpones action until evidence improves, and refusal protects against category errors when authority or standing is absent. These behaviors distinguish helpful assistance from premature action: the system can be useful without pretending to know, and responsive without committing beyond its remit. Reframing optimization around the quality of licensing—when to elevate, when to hold, and when to decline—realigns incentives away from surface fluency and toward accountable stance control.

6.5 Accountability via Procedural Licensing

Treating force as something conferred by procedure rather than presumed by form establishes a practical basis for accountability. Gate evaluations, their inputs, and their rationales can be logged, reviewed, and audited, so that responsibility attaches to the licensing step instead of to an opaque generation. Errors then become failures of procedure—insufficient warrant, missing authority, absent standing—rather than mysteries of model intention. This relocation narrows responsibility gaps for intentionless systems and enables systematic remediation: policies can be tuned, thresholds adjusted, or delegation pathways clarified without anthropomorphizing the model or conflating language with action.

Table 1. Speech-Act Type × Felicity Gate × Output Strategy

Speech-Act Type	Gate (Condition)	Licensed Output (if met)	Fallback Strategy (if unmet)
Assertive	Warrant—traceable evidence / source verification	Informative assertion with logged provenance	Proposal or information request; hedged statement
Directive	Authority—verified role, scope, and permission	Executable command with audit record	Confirmation prompt or refusal with justification
Commissive	Standing—delegated right to commit or promise	Binding commitment under authenticated delegation	Draft or non-binding suggestion; human approval required
Declaration	Standing + Procedure—institutional authority and formal context	Effective declaration executed under verified procedure	Simulated template output until authorization confirmed

7. Discussion

7.1 Theoretical Contribution

The account reframes hallucination as infelicity of force-licensing, aligning LLM governance with Speech Act Theory and Gricean cooperation rather than with truth conditions alone (Austin, 1962; Searle, 1969; Grice, 1975). It supplies a compact vocabulary—warrant, authority, standing, relevance—that travels cleanly between pragmatics and system design.

7.2 Relation to Existing Work

Surveys explain hallucination through data, objectives, and decoding; safety stacks add filters, guardrails, and tool permissions (Ji et al., 2023; Huang et al., 2023; Meta, 2024). Our view treats these practices as felicity mechanisms: retrieval grounds assertive, policy gates license directives, and vetting governs uptake—clarifying why they help and where they fail.

7.3 Architectural and Evaluation Implications

Design should bind assertive sources with traceable attribution, default to hedging when retrieval is weak, gate tool invocation by role/scope/context/risk, and route commissive/declarations to recognized signatories, with simulation as the safe default.

Evaluation should report performative metrics—rates of warranted assertions, authorized tool calls, delegated commitments, and clarifications—alongside factuality, consistent with design-science criteria of usefulness and explanatory power (Gao et al., 2023; Zhao et al., 2024; Hevner et al., 2004; Gregor, 2006).

Beyond architecture, this framework invites a new class of benchmarks where performance is judged by when language is entitled to act rather than how often it appears correct. These

performative metrics could be used to evaluate stance control, gate reliability, and procedural transparency across systems, bridging design science and pragmatic theory.

7.4 Normative Boundaries

The proposal is intention-agnostic: it does not attribute minds to models; it proceduralizes entitlement. Limits are principled—statistical confabulation persists, and institutions may lack clear authorities or interoperable signatures—so transparency and human-in-the-loop remain the default safe state (Matthias, 2004; Clark, 1996).

Future work could formalize felicity gating as an auditable protocol, integrating pragmatic licenses with institutional accountability systems such as provenance chains or AI governance APIs. By treating entitlement as procedural rather than psychological, such extensions could align technical design with legal and ethical accountability without collapsing the distinction between agents and artifacts.

8. Conclusion

This paper began from a simple observation: in machine-mediated interaction, language already does things. The risk arises when systems let strings be taken up as actions without the conditions that entitle them to act. The remedy is architectural rather than psychological: not to graft intention onto models, but to install rules of entitlement at the boundary where text would affect the world.

Four principles follow in a single, continuous movement. Firstly, assertions require warrant: when provenance and evidence fail to meet contextual thresholds, the right stance is a proposal to seek sources, not a claim. Secondly, directives require authority: when actor, scope, context, or risk is unverified, the right move is confirmation, narrowing, or refusal, not execution. Thirdly, commendations and declarations require standing: without recognized delegation, the right artefact is a draft or simulation, not a binding commitment. Finally, relevance and coherence regulate mixed or ambiguous turns: the system separates acts, clarifies missing details, and ensures each meets its own condition before uptake.

Seen together, these principles do more than regulate interfaces. They relocate responsibility from imagined inner states to auditable procedures; they separate epistemic control (what is true enough to say) from deontic control (what is permitted to do), making the two complementary rather than interchangeable; and they redefine success as the disciplined flow of licensed action rather than the mere production of fluent text. What emerges is a view of artificial language use as a civic practice: text becomes actionable only under rules that a community can recognize, inspect, and revise.

Beyond immediate architecture, this argument points toward a pragmatic ethics of artificial language—an understanding of computation as participation in social reality. If systems can act through words, their design inherits duties that have long governed human speech: to make reasons traceable, to disclose the conditions of action, and to recognize when restraint is the appropriate form of care. Entitlement, in this sense, is not a property of the model but a relation among agents, roles, evidence, and consequences. Making that relation explicit is how technical systems join a world of rules without pretending to possess minds.

The design implications are concrete. Interfaces should surface stance by default—proposal, draft, query, simulation—rather than blur every output into an assertion or command. Source binding should be visible and inspectable, with links to retrieval or provenance rather than generic confidence. Tool use should be mediated by role, scope, and risk checks tied to the caller, not merely to the content of the request. Commitments should pass through clear delegation or signature pathways, with human countersigning where institutional responsibility is required. Clarification should be the norm whenever a turn conflates acts or leaves essential parameters underdetermined. These behaviors make systems slower only where slowness is a virtue; elsewhere they sharpen the boundary between helpful language and unlicensed action.

Evaluation follows from the same logic. If entitlement is the central variable, then performance should be measured not only by accuracy or fluency but by the rate at which language is correctly licensed to act. Benchmarks can therefore record how often answers ship without warrant in warrant-sensitive domains; how often tool calls proceed after successful role and scope checks; how often commitments are escalated to recognized signatories rather than issued directly; how often ambiguous turns trigger clarification instead of silent assumptions. Longitudinally, systems can be judged by incident patterns: whether failures cluster around missing warrant, missing authority, absent standing, or broken coherence, and whether policy changes reduce those clusters without degrading utility. Such metrics do not replace truth tracking; they discipline the moment when truth claims and actions are allowed to count.

Governance and accountability take a clearer form under this framing. If force is conferred by procedure, then logs of licensing decisions become the primary substrate of responsibility. Reviewing a failure is no longer a speculative inquiry into what the model "intended"; it is an examination of which conditions were claimed, which were met, which were waived, and by whom. This moves the conversation from psychology to institutions. It also permits targeted remediation: thresholds can be tuned, exemptions narrowed, delegation rules tightened, and explanations improved, all without mythologizing agency or overfitting policy to anecdotes.

The approach is intentionally modest about what models can provide. Statistical hallucination will persist: distributional learning does not, by itself, secure truth or authority. The contribution of a pragmatic layer is to prevent confabulated or unauthorized strings from acquiring performative force. In effect, it places a governor on the uptake channel rather than a filter on the generative source. The result is not infallibility but a shift in the profile of mistakes—from actionable harms to informational proposals, from silent overreach to visible requests for license.

There are limits and open questions. Some domains lack crisp authorities or interoperable signatures; others tolerate ambiguity by design. In low-resource settings, provenance may be partial or delayed, making warrant thresholds difficult to meet without frustrating users. Overzealous gating could anaesthetize useful improvisation or encumber benign tasks with bureaucratic friction. These tensions are real. They call for graded behaviors rather than absolutes, with explicit disclosure of what is being withheld and why, and with escalation paths that keep humans in the loop where contextual judgement is decisive. They also invite research on calibration strategies that preserve utility while keeping force in check, and on user experience patterns that make stance legible without cognitive overload.

The framework naturally extends to settings beyond single-turn chat. Multi-agent orchestration raises questions about how one agent's licensed act becomes another's premise or

obligation: warrant, authority, and standing must compose across chains, not merely within one interface. Embodied systems add spatial and temporal stakes; here, delays and refusals may be safety features rather than defects. Multilingual interaction raises the possibility that felicity norms vary across linguistic communities, requiring localization of thresholds and disclosure practices. In collaborative writing and design, where simulation is the productive mode, systems may license more by default but also signal more clearly that outputs are drafts, preventing accidental reclassification into commitments. Each extension tests the same idea: that language-to-action requires criteria for when language may act.

A further horizon concerns standardization. If entitlement is procedural, there is value in portable artefacts that record it: attestation tokens for warrant, capability or policy objects for authority, delegation handles for standing, and dialogue frames for coherence and relevance. Such artefacts would let licensing travel with text, so that downstream tools or organizations can verify conditions without re-deriving them. They would also enable cross-vendor evaluation: systems could be asked to show not only what they produced but how each act was authorized. In time, this could mature into a common set of interfaces by which artificial language participates responsibly in shared infrastructures.

Future research can therefore proceed along complementary tracks. Formally, the gates can be specified as typed predicates with crisp failure modes, then composed with decoding and tool-use policies. Empirically, offline replays and online A/B tests can assess whether stance-aware behaviors reduce harmful uptake without eroding task outcomes. Organizationally, audit schemas can be designed to support root-cause analysis and counterfactual "what-if" queries. Educationally, developer guidelines and prompts can be reframed around entitlement rather than around generic safety slogans. None of this requires models to have inner states; it requires systems to make their outer rules visible and consistent.

At stake is the kind of agency granted to language in technical systems. A world that automates uptake without license will continue to call confabulation "hallucination" and treat harm as an unfortunate side effect of fluency. A world that makes license explicit will treat language as a proposal first and an action only when conditions are met, making accountability an architectural property rather than a moral afterthought. The choice is not between optimism and caution; it is between tacit power and governed power.

What remains, then, is not a call for more human-like machines, but for more accountable interfaces to language—systems that recognize when they are entitled to act, and show their working when they do. That is the minimal condition for deploying intentionless models that nevertheless make things happen. For every system that speaks, the question is not whether it knows, but whether it has earned the right to be heard.

Acknowledgments

We are grateful to all respondents who participated in this study.

Funding

This work was supported by University of the Arts London under Grant number 25045149.

References

[1]. Austin, J. L. (1962). *How to do things with words*. Oxford University Press.

[2]. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 610-623.

[3]. Clark, H. H. (1996). *Using language*. Cambridge University Press.

[4]. Fischer, K. (2009). The role of users' preconceptions in talking to computers and robots. *International Journal of Social Robotics, 1*(1), 83-93.

[5]. Gao, Y., Zeng, A., Zhang, R., Liu, Y., Chen, G., Lin, Z., Wang, X., & Lin, J. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint*, arXiv:2312.10997.

[6]. Gregor, S. (2006). The nature of theory in information systems. *MIS Quarterly, 30*(3), 611-642.

[7]. Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics, Vol. 3: Speech acts* (pp. 41-58). Academic Press.

[8]. Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly, 28*(1), 75-105.

[9]. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint*, arXiv:2311.05232.

[10]. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 248:1-248:38.

[11]. Levinson, S. C. (1983). *Pragmatics*. Cambridge University Press.

[12]. Luger, E., & Sellen, A. (2016). "Like having a really bad PA": The gulf between user expectation and experience of conversational agents. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5286-5297.

[13]. Manakul, P., Laban, P., Tsvetkov, Y., & Soricut, R. (2023). SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint*, arXiv:2303.08896.

[14]. Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175-183.

[15]. Meta. (2024). *Llama Guard 3 model card*. Hugging Face.

[16]. NVIDIA. (n.d.). *NeMo Guardrails documentation*. NVIDIA Developer Documentation.

[17]. Porcheron, M., Fischer, J. E., Reeves, S., & Sharples, S. (2018). Voice interfaces in everyday life. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*, 1-12.

[18]. Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge University Press.

[19]. Wang, X., Wei, J., Schuurmans, D., Chi, E. H., & Zhou, D. (2022). Self-consistency improves chain-of-thought reasoning in language models. *arXiv preprint*, arXiv:2203.11171.

[20]. Zhao, P., Li, Y., Yang, Y., Wang, L., & Jiang, H. (2024). Retrieval-augmented generation for AI-generated content: A survey. *arXiv preprint*, arXiv:2402.19473.