

Generative AI and Linguistic Mediation

Relational Knowledge, Expression Dynamics, and Provenance

Wanhong HUANG
huangwanhong@serendip.ngo

Working draft, August 31, 2026

Abstract

Generative AI increasingly learns from human traces while also entering the processes through which people write, translate, summarize, remember, and represent one another. This paper develops a preliminary framework for evaluating that dual position. Its philosophical and interdisciplinary review connects arguments about syntax and understanding, symbol grounding, tacit and situated knowledge, distributed cognition, multimodal and embodied systems, AI-mediated communication, predictive writing, linguistic inequality, participatory language technology, dataset documentation, and content provenance. The central proposal is a layered warrant profile that separates artifact performance, participation in a source relation, practical successor competence, interpretive standing, expression distribution, individual accessibility, semantic and relational-historical fidelity, provenance, institutional authority, and downstream generativity. A typed mediation chain and an exposure–acceptance–editing kernel locate human and institutional action within the process. An exact finite construction then shows that AI assistance can expand one user’s effective expression repertoire while increasing concentration in the population’s accepted output distribution. A state-dependent recurrence separates human accommodation, system updates, institutional selection, and training-data feedback; local convergence follows under an explicit contraction condition. A second exact construction establishes that a final artifact alone generally underidentifies its human and AI transformation history. Three mechanism cases involving professional rewriting, minoritized-language assistance, and institutional summaries of shared experience illustrate the framework. The paper concludes with an empirical programme and a governance portfolio joining usable disclosure, selective provenance, privacy, contestation, community authority, portability, and longitudinal review. The resulting position is conservative and revisable: successful artifacts, relational competence, expression change, provenance, and legitimate authority each require evidence suited to their own object.

Keywords: generative AI; linguistic mediation; relational knowledge; AI-mediated communication; expression diversity; accessibility; provenance; interpretive standing

Discussion Paper Note

This paper is a preliminary discussion paper intended to share an evolving idea and invite further dialogue, criticism, revision, and independent development.

The author claims responsibility for the definitions, formal constructions, taxonomy, arguments, selection of material, and conclusions presented here. Similar or related ideas may have appeared in other intellectual, cultural, or disciplinary traditions. Any legal rights retained in this work are intended to support attribution, responsible use, and protection against exploitative or harmful appropriation, while preserving legitimate inquiry, criticism, revision, and further development.

The arguments should be understood as provisional and historically situated. Readers are encouraged to question, revise, extend, reinterpret, or independently develop the ideas presented here. Acknowledgment of this paper as one point of encounter is appreciated where appropriate, while epistemic ownership over independently developed ideas remains with their contributors.

Responsible Use and Rights Reservation

The author encourages good-faith discussion, criticism, independent development, and responsible use of the knowledge presented in this work. Responsibility for lawful and ethically appropriate use remains with each user.

The author expressly reserves all rights and remedies available under applicable law with respect to unlawful conduct, harmful or abusive exploitation, improper commercial appropriation, infringement of applicable intellectual-property or other legal rights, and conduct contrary to applicable national, regional, or international law.

This reservation preserves the ability to respond to misuse and harmful appropriation. Legitimate academic inquiry, criticism, independent reasoning, and further development remain encouraged.

Notices

Status. This is a working draft circulated for discussion. Its claims, formal objects, examples, and section numbering remain open to revision.

Licence. This work is made available under a Creative Commons Attribution–NonCommercial 4.0 International Licence (CC BY-NC 4.0), subject to the rights reservation stated on the preceding page.

Statement on the use of language models. Drafting, literature search, structural review, formal reconstruction, and argumentative criticism were conducted in dialogue with ChatGPT (OpenAI). The author bears responsibility for the claims, definitions, formal constructions, selection of material, argument, and position taken. References included in the bibliography were checked against publisher, journal, university, DOI, standards-body, or responsible repository records during preparation.

Companion papers. This paper belongs to a programme on co-experience knowledge, transmission, linguistic continuity, cultural preservation, institutional translation, political economy, and AI mediation. Earlier papers develop the general ontology and continuity of co-experience knowledge, transmission standing and responsibility, language documentation and revitalization, cultural preservation and experiential reconstruction, institutional translation of situated experience, and the political economy of shared experience. The present paper owns AI-specific trace learning, message mediation, expression-distribution feedback, provenance, and interpretive-standing evaluation.

Suggested citation. Huang, W. *Generative AI and Linguistic Mediation: Relational Knowledge, Expression Dynamics, and Provenance*. Working draft.

Contents

1	Introduction	7
2	Theoretical and Disciplinary Inheritance	8
2.1	Form, Meaning, and Grounding	8
2.2	Participation, Tacit Knowledge, and Distributed Practice	9
2.3	Multimodal and Interactive Systems	9
2.4	AI-Mediated Communication and Machine Agency	10
2.5	Writing Assistance and Experimental Evidence	10
2.6	Linguistic Inequality and Community Participation	11
2.7	Documentation, Risk, and Technical Provenance	11
3	Analytical Vocabulary and Warrant Layers	12
3.1	Symbolic Relational Traces	12
3.2	Recurrent Linguistic Mediation	12
3.3	Layered Warrant Profile	13
3.4	Bridge-Premise Register	14
4	Mediation Architecture	14
4.1	Source-to-Consequence Chain	14
4.2	Exposure, Acceptance, and Editing	15
4.3	Agency and Role Allocation	15
4.4	Disclosure and Audience Interpretation	16
4.5	Institutional Selection and Downstream Use	16
5	Expression Distribution and Accessible Repertoire	16
5.1	Population Distribution	17
5.2	Individual Effective Repertoire	17
5.3	Expansion–Concentration Coexistence	18
5.4	Heterogeneous Effects and Category Politics	19
5.5	Distributional Interpretation	19
6	Recurrent Human–AI Dynamics	20
6.1	Coupled State and Transition	20
6.2	Local Stability and Feedback Gain	21

6.3	Feedback-Edge Register	22
6.4	State-Dependent Direction	22
7	Fidelity, Provenance, and Interpretive Standing	23
7.1	Semantic and Relational-Historical Fidelity	23
7.2	Transformation-Event Graph	24
7.3	Output-Only Non-Identification	24
7.4	Technical Provenance and Interpretive Evidence	25
7.5	Interpretive Standing and Institutional Authority	25
7.6	Selective Provenance and Privacy	26
8	Comparative Mechanism Cases	26
8.1	Professional Rewriting and Local Voice	26
8.2	Minoritized-Language Assistance and Community Capacity	27
8.3	Institutional Summaries of Shared Experience	28
8.4	Cross-Case Mechanism Matrix	28
9	Empirical Research Programme	29
9.1	Units, Treatments, and Outcome Profiles	29
9.2	Controlled Interface Studies	30
9.3	Longitudinal Repertoire and Distribution Panels	30
9.4	Participation and Successor-Competence Studies	31
9.5	Participatory and Community-Governed Evaluation	31
9.6	Institutional Process and Authority Studies	32
9.7	Model-Update and Data-Pool Audits	32
9.8	Measurement Robustness and Rival Mechanisms	32
9.9	Research Ethics and Stopping Conditions	33
10	Governance and Design	33
10.1	Admissible Design Set	33
10.2	User-Controlled Fidelity Modes	34
10.3	Plural Suggestion and Distribution Controls	34
10.4	Provenance Tiers	34
10.5	Community Authority and Individual Safeguards	35
10.6	Institutional Procurement and Use Controls	35

10.7 Recurrent Monitoring and Revision	35
10.8 Governance Portfolio	36
11 Objections and Scope Conditions	36
11.1 Conceptual Redundancy	37
11.2 Human Exceptionalism	37
11.3 Functional Sufficiency	37
11.4 Metric Politics	37
11.5 Accessibility and Stylistic Mobility	38
11.6 Provenance Surveillance	38
11.7 Longitudinal Extrapolation	38
11.8 Recursive-Training Analogy	38
11.9 Community Romanticization	38
11.10 Formal Simplification	39
12 Future Research	39
13 Conclusion	40

1 Introduction

This section introduces the paper’s object, motivation, research question, position, contributions, method, and structure. It begins from generative AI’s dual role as a learner from symbolic traces and a mediator of later expression, then states the distinctions and evidence boundaries that guide the analysis.

Generative AI occupies two locations in contemporary symbolic life. The first is retrospective. Models are trained on text, images, audio, video, annotations, interaction records, sensor streams, and other traces of human and nonhuman activity. The second is prospective. Models propose words, complete sentences, translate registers, summarize meetings, compose images, answer questions, and reshape artifacts that later circulate among people and institutions. An output can therefore stand downstream of prior human relations and upstream of future human expression. This dual location creates an evaluation problem whose parts are often compressed into a single judgment about intelligence, usefulness, bias, authorship, or authenticity.

Several examples reveal the compression. A multimodal model answers questions about an image. The result can support a claim about performance on the selected task. Participation in the practices that produced the image remains a different property. A writing assistant enables a user to express an idea in a register that was previously costly or inaccessible. The user’s practical repertoire may expand. If many users accept similar high-probability suggestions, the distribution of accepted writing may also concentrate. An AI summary of a group’s shared experience can be accurate on selected propositions. Its authority to represent the group, frame uncertainty, or support an institutional decision requires further warrants. A final polished paragraph may reveal little about its source, suggestions, edits, acceptance decisions, model version, or institutional context.

The paper asks the following primary question:

How should generative-AI systems that learn from traces and mediate human expression be evaluated across artifact performance, relational knowledge, practical successor competence, interpretive standing, expression distributions, individual accessibility, recurrent feedback, fidelity, provenance, and institutional authority?

The preliminary position is layered. Artifact performance, participation in a source relation, practical successor competence, and interpretive standing are distinct achievements. Distributional concentration and individual accessibility can move together in several directions, including strict individual expansion combined with strict population concentration. Recurrent mediation links present suggestions to later human habits, institutional expectations, data pools, and system behavior. Provenance can make selected transformation histories inspectable, while the integrity of a provenance assertion remains distinct from its completeness, truth, meaning, endorsement, standing, and legitimacy. These distinctions support a plural research and governance programme.

The paper contributes five connected elements. First, it reconstructs relevant antecedents across philosophy of AI, situated and distributed cognition, multimodal systems, communication research, human–AI co-writing, sociolinguistic inequality, participatory language technology, documentation, and technical provenance. Second, it proposes a ten-coordinate warrant profile and a typed mediation chain. Third, it derives an exact coexistence result for individual repertoire expansion and population concentration. Fourth, it develops a state-dependent recurrence with a conditional local stability result and explicit feedback edges. Fifth, it proves output-only provenance underidentification and converts the result into a selective provenance and governance architecture.

The method is constructive and comparative. Formal objects expose hidden transitions and establish limited possibility or non-identification results. Published experiments provide bounded mechanism evidence. Philosophical analysis specifies bridge premises among performance, competence, participation, and standing. Three constructed cases place the same framework in different institutional settings. Normative recommendations remain conditional on stakes, affected groups, legitimate authority, privacy, mobility, and available alternatives.

Several limits govern the inquiry. Short-horizon writing experiments support claims about the studied tasks and mechanisms; durable language change requires longitudinal evidence. Research on recursive model training supports the model-update edge; it supplies separate evidence from human accommodation. Multimodal or embodied performance expands the relevant evidence base while leaving social participation and representative authority open. Community governance supplies an authority and process requirement, with internal plurality and individual safeguards retained. The paper’s formal results are exact only inside their stated constructions.

The remaining sections proceed from inheritance to application. Section 2 reconstructs the interdisciplinary literature. Section 3 defines the analytical vocabulary and warrant layers. Section 4 types the mediation process. Sections 5 and 6 develop distributional and recurrent dynamics. Section 7 treats fidelity, provenance, and standing. Section 8 compares three mechanisms. Sections 9 and 10 specify research and governance programmes. Sections 11 and 12 state objections and future obligations. Section 13 concludes.

2 Theoretical and Disciplinary Inheritance

This section locates the proposed framework among relevant antecedents and identifies the precise role each tradition plays. The discussion moves from meaning and grounding to situated knowledge, multimodal systems, mediated communication, experimental writing effects, linguistic inequality, and documentation. Each cluster constrains a different inference in the later model.

2.1 Form, Meaning, and Grounding

This subsection establishes the philosophical background for separating successful symbolic artifacts from broader claims about understanding. It compares formal-symbol, grounding, and form–meaning arguments and converts their disagreement into an evaluation discipline.

Searle’s Chinese Room argument challenged the inference from appropriate symbol manipulation to understanding [24]. Its continuing relevance lies in the inferential gap it places before the reader. The argument remains philosophically contested, and current architectures differ from the program imagined in 1980. The paper therefore treats it as an antecedent objection to a performance-to-understanding inference, while leaving the final metaphysical assessment open.

Harnad formulated the symbol grounding problem as a problem of connecting symbols with nonsymbolic capacities [16]. This formulation shifts attention toward perceptual categorization, sensorimotor relations, and the origin of symbolic reference. Bender and Koller later distinguished form from meaning and argued that a system trained only on form lacks a route to meaning from that source alone [6]. Their condition matters. A text-only training argument has a narrower scope than a conclusion about every system incorporating perception, action, interaction, human feedback, or institutional embedding.

The present framework draws a methodological conclusion from this debate. Evaluation should type the learning relation and the target claim. A text prediction score, visual question-answering score, manipulation success rate, practical transfer test, source-participant judgment, and institutional authorization each provide evidence about different objects. Stronger evidence can accumulate across modalities and interventions. Its interpretation remains target-specific.

2.2 Participation, Tacit Knowledge, and Distributed Practice

This subsection supplies antecedents for relation-sensitive knowledge and practical successor competence. It distinguishes participation in a practice from possession of its traces and also recognizes the epistemic role of artifacts inside distributed systems.

Lave and Wenger described learning through changing participation in communities of practice [21]. The relevant point is structural: access to records and rules can accompany a further trajectory of observation, peripheral participation, correction, responsibility, and changing membership. Collins’s account of tacit and explicit knowledge similarly directs attention toward practical and socially embedded capacities that exceed declarative reproduction [11]. These traditions motivate transfer tasks in which a successor responds to novelty, detects breakdown, repairs action, and guides another learner.

Hutchins’s distributed-cognition analysis broadens the unit of cognitive inquiry across people, artifacts, representational media, and organized activity [18]. This broader unit prevents a second simplification. A tool can contribute materially to a cognitive achievement even when it lacks every competence of the encompassing system. Consequently, the framework allows AI to be a causally and epistemically significant component of a distributed process. Questions about the bearer of knowledge, the system’s practical competence, and an artifact’s authority remain separately stated.

These antecedents support two symmetrical cautions. Trace access alone gives limited evidence about participation and practical transfer. Participation alone also leaves accuracy, completeness, generalization, and legitimate authority open. The layered profile developed below preserves both cautions.

2.3 Multimodal and Interactive Systems

This subsection connects the philosophical distinctions to contemporary architectures that combine language with images, sensor states, and embodied tasks. Its role is to expand the space of relevant evidence while preserving the target-specific interpretation of that evidence.

Flamingo processes interleaved visual and textual data and was evaluated across visual question answering, captioning, and related few-shot tasks [2]. PaLM-E integrates textual tokens with visual and continuous state inputs and was evaluated on robotics, visual-language, and planning tasks [13]. Such systems materially differ from a language model trained and evaluated exclusively on text. Their sensors, task loops, action repertoires, and feedback supply additional relations among symbols, perceptions, and consequences.

The expansion has a disciplined interpretation. A robot completing a selected manipulation task supplies practical evidence within the embodiment, environment, skill library, and evaluation regime. It can support an interventionally stronger claim than text resemblance. Transfer to new environments, self-correction, social practice, source-specific history, and representative standing

require further tests. The distinction reflects evidential granularity and leaves a general human-machine capacity boundary outside the present argument.

2.4 AI-Mediated Communication and Machine Agency

This subsection moves from AI as a task performer to AI inside communication between people. It introduces established definitions of AI-mediated communication and clarifies the framework’s extension toward recurrent linguistic and institutional effects.

Hancock, Naaman, and Levy define AI-mediated communication as interpersonal communication in which a computational agent operates on behalf of a communicator by modifying, augmenting, or generating messages to accomplish communication goals [15]. Their dimensions include magnitude, media type, optimization goal, autonomy, and role orientation. Sundar’s account of machine agency analyzes how interfaces enable users to perceive and attribute agency to machines [26]. Together, these works make the mediating system part of the communication architecture.

The present proposal extends that architecture in three directions. First, it records exposure, acceptance, editing, rejection, disclosure, and audience interpretation as distinct events. Second, it follows outputs into institutional use, later human expression, product updates, and training data. Third, it separates perceived machine agency, causal influence, human endorsement, interpretive standing, and institutional authority. A system can exercise substantial causal selection through suggestions while a person formally retains the send decision. The resulting agency distribution is stage-specific.

2.5 Writing Assistance and Experimental Evidence

This subsection reviews controlled evidence on suggestion systems, smart replies, opinionated co-writing, creativity, and cross-cultural style. Its objective is to identify supported mechanisms and maintain the boundary between immediate effects and long-run linguistic change.

Arnold, Chauncey, and Gajos found that predictive-text suggestions changed caption length and lexical choice while increasing entry speed in their task [3]. Hohenstein and colleagues used randomized smart-reply experiments and found changes in communication speed, emotional language, and interpersonal evaluation [17]. Jakesch and colleagues found that opinionated language-model suggestions influenced expressed views in a co-writing experiment [19]. These studies support causal pathways from interface exposure through acceptance to a final message and audience response.

Two recent studies directly motivate a multi-level diversity analysis. Doshi and Hauser found that access to generative-AI ideas improved several evaluations of individual short stories, especially for some writers, while AI-assisted stories were more similar to one another [12]. Agarwal, Naaman, and Vashistha found cross-cultural differences in benefits and observed writing shifts toward Western styles among Indian participants in the selected tasks [1]. These findings supply task-bounded evidence for individual gain, aggregate similarity, and cultural asymmetry. They also expose the need to measure acceptance, editing, user group, language variety, genre, and institutional use.

The evidence supports mechanism hypotheses whose temporal scale is explicit. Suggestion effects during one session can arise through priming, convenience, anchoring, strategic delegation, or interface defaults. Durable repertoire change additionally requires later unaided production, repeated exposure, cross-context transfer, social accommodation, and persistence after system withdrawal.

The recurrent model in Section 6 represents these edges separately.

2.6 Linguistic Inequality and Community Participation

This subsection places expression mediation within unequal language-technology infrastructures. It connects resource disparities, extractive data relations, participatory development, and Indigenous data governance to the paper’s accessibility and authority coordinates.

Joshi and colleagues document major disparities in language representation and resources within NLP [20]. Blasi, Anastasopoulos, and Neubig quantify systematic inequalities in performance across languages and tasks [8]. These disparities make an aggregate claim about “AI assistance” analytically weak. A system can expand access for speakers of one language, impose correction burdens on another, and expose a third community to extraction or dependency.

Bird argues that speech and language technology can reproduce extractive relations when Indigenous languages and knowledge are treated as data detached from local knowledge authorities [7]. The Masakhane case reported by Nekoto and many collaborators demonstrates a participatory process for African-language machine translation and treats low-resourcedness as a systemic condition extending beyond dataset size [23]. The CARE Principles foreground Collective Benefit, Authority to Control, Responsibility, and Ethics for Indigenous data governance [9].

These sources change the object of evaluation. Performance, resource quantity, and model coverage remain important. The process also includes who defines the task, supplies and governs data, evaluates quality, carries correction labor, controls access, receives benefit, and can revise or end the arrangement. Community authority requires internal procedures capable of representing plurality and conflict. Individual mobility, privacy, and refusal retain their own safeguards.

2.7 Documentation, Risk, and Technical Provenance

This subsection reconstructs documentation and provenance antecedents. Its role is to distinguish documentation of datasets and models from artifact-level transformation history and to bound the contribution of cryptographic provenance.

Data statements were proposed to document populations, language varieties, collection conditions, and generalization limits in NLP [4]. Datasheets organize questions about dataset motivation, composition, collection, processing, uses, distribution, and maintenance [14]. Model cards organize intended uses, evaluation conditions, subgroup results, and limitations [22]. Bender and colleagues connect large language models to risks involving data, bias, scale, environmental cost, and communicative intent [5]. These practices operate at dataset, system, and deployment levels.

Artifact-level provenance requires an additional structure. The C2PA technical specification represents assertions, claims, actions, ingredients, content bindings, manifests, signers, trust, and validation across asset workflows [10]. Its architecture can support tamper evidence and association of assertions with an asset under a defined trust model. The specification also treats privacy and user control as design concerns. Its scope preserves a crucial limit: validation of an associated assertion supplies a different result from a value judgment about the assertion.

Recursive training adds a further provenance interest. Shumailov and colleagues studied generational model training with generated data and reported loss of distributional information under the investigated regimes [25]. Their result concerns model-training recursion. The human-expression

feedback path has different mechanisms and requires separate evidence. A provenance architecture can help distinguish human-produced, AI-generated, AI-edited, and mixed data, while access to original human data raises consent, compensation, privacy, and governance questions.

3 Analytical Vocabulary and Warrant Layers

This section defines the paper’s core terms and establishes the layered evaluation profile. It begins with symbolic relational traces and recurrent mediation, then separates ten warrant coordinates and states the bridge-premise rule governing movement among them.

3.1 Symbolic Relational Traces

This subsection identifies the object from which many generative systems learn. The definition emphasizes transformation and later use, allowing traces to preserve, lose, reorganize, and generate relations.

Definition 3.1 (Symbolic relational trace). *A symbolic relational trace is a persistent or portable symbolic construction produced from a lived relation or shared activity through a declared selection and representation channel. Text, testimony, images, recordings, corpora, interaction logs, annotations, and multimodal datasets are candidate traces.*

The trace is a present object with its own affordances. It can retain propositional information, voice, sequence, gesture, spatial organization, or interaction structure. It can omit tacit coordination, unrecorded participants, material conditions, uncertainty, conflict, and practical consequences. Later interpreters enter a new relation with the trace. A model trained on traces therefore participates in a new causal and symbolic arrangement whose connection to the source relation requires specification.

3.2 Recurrent Linguistic Mediation

This subsection defines the process that joins trace learning to later human and institutional expression. It distinguishes one-step assistance from a recurrent system whose outputs can alter later inputs and conditions.

Definition 3.2 (Recurrent AI linguistic mediation). *Recurrent AI linguistic mediation is a time-indexed process in which an AI system selects, transforms, proposes, or generates symbolic material within human communication; people and institutions accept, edit, reject, interpret, circulate, or govern that material; and resulting traces can alter later human repertoires, institutional expectations, system behavior, or training data.*

The definition includes predictive text, rewriting, translation, summarization, conversational suggestions, multimodal description, and institutionally embedded report generation when these systems enter communication among people. It also includes later feedback when accepted artifacts influence expectations, templates, evaluation standards, or data pools. Understanding, authorship, endorsement, and authority remain separate classifications.

Table 1: Layered warrant profile for generative-AI linguistic mediation

Layer	Structural object	Evidence target	Inference boundary
Artifact performance	final task output	target validity, robustness, calibration, transfer	task success leaves process and authority open
Source participation	relation and practice	interventionally relevant interaction history	trace similarity leaves participation open
Successor competence	practical continuation	novel-condition action, repair, teaching, adaptation	reproduction leaves practical transfer open
Interpretive standing	warranted role	mandate, expertise, participation, accountability, correction	fluency leaves standing open
Expression distribution	population outputs	support, concentration, entropy, category flow	individual gain leaves aggregate direction open
Individual accessibility	effective repertoire	cost, comprehension, correction, feasible competent use	system availability leaves capability open
Fidelity	source-artifact relation	semantic and relational-historical discrepancy	semantic adequacy leaves positional trace open
Provenance	transformation history	coverage, integrity, lineage, disclosure, privacy	valid record leaves truth and completeness open
Institutional authority	downstream use	jurisdiction, procedure, competence, review, appeal	technical deployment leaves authorization open
Future generativity	later conditions	human and machine repertoires, succession, options, dependency	current quality leaves long-run change open

3.3 Layered Warrant Profile

This subsection gives the paper’s central analytic separation. The profile organizes epistemic, distributional, institutional, and temporal claims while preserving their independent evidence requirements.

For model M , source domain R , task family τ , deployment δ , audience b , institution o , and horizon H , define

$$\mathbf{W}(M; R, \tau, \delta, b, o, H) = (W_{\text{art}}, W_{\text{part}}, W_{\text{succ}}, W_{\text{stand}}, W_{\text{dist}}, W_{\text{access}}, W_{\text{fid}}, W_{\text{prov}}, W_{\text{auth}}, W_{\text{gen}}). \quad (1)$$

Table 1 states the structural role, evidence target, and common inference error for each coordinate. A coordinate can itself be multidimensional and uncertainty-bearing. Aggregation remains application-specific and premise-dependent.

Claim 3.3 (Layered warrant separation). *Artifact performance, participation in a source relation, practical successor competence, and warranted interpretive standing are distinct achievements.*

Evidence transfers among these layers through explicit bridge premises and target-appropriate validation.

The claim establishes an inferential discipline. It allows an application to justify a bridge. For example, repeated successful intervention in a practice, transfer across novel conditions, correction by qualified participants, and robust performance after environmental changes can jointly support a practical-competence claim. A further representative role requires mandate, procedure, accountability, and the ability of affected people to contest the use.

3.4 Bridge-Premise Register

This subsection converts the layered profile into an argument template. It identifies the premises needed when a manuscript, institution, or product moves from one warrant coordinate to another.

Let E_a be evidence at layer a , C_b a conclusion at layer b , and $B_{a \rightarrow b}$ the bridge premise. A valid application requires

$$E_a \wedge B_{a \rightarrow b} \implies C_b, \quad (2)$$

with the bridge independently defended. A benchmark score plus a validated relationship between that benchmark and field transfer may support a bounded competence conclusion. Accurate summary content plus a legitimate mandate and effective correction process may support a bounded institutional-use conclusion. The framework asks reviewers to inspect the bridge as a separate premise before accepting the output-level inference.

4 Mediation Architecture

This section represents generative AI inside a complete source-to-consequence chain. It first types the stages, then factors accepted expression into exposure and human decision, and finally locates agency, disclosure, and institutional use across the architecture.

4.1 Source-to-Consequence Chain

This subsection supplies the process skeleton used throughout the paper. The chain separates source symbolization, model proposal, human decision, audience interpretation, institutional use, and downstream change.

For source relation R , initial trace x_0 , context c , model and interface state ψ , proposal s , human decision d , final artifact y , audience interpretation $\hat{K}_b$, institutional use o , and downstream condition z^+ , define

$$R \xrightarrow{\sigma} x_0 \xrightarrow{(c, \psi)} s \xrightarrow{d \in \{\text{accept}, \text{edit}, \text{reject}\}} y \xrightarrow{I_b} \hat{K}_b \xrightarrow{U_o} o \xrightarrow{\Delta} z^+. \quad (3)$$

The symbolization map σ can include prior editorial, archival, annotation, or data-selection operations. The proposal stage can generate a complete artifact or a ranked set of fragments. The decision stage includes partial acceptance, substantial editing, rejection, recombination, and repeated

prompting. Audience interpretation depends on disclosure, expectations, context, and the perceived agency of human and machine contributors. Institutional use can classify, evaluate, publish, allocate, discipline, preserve, or decide. Downstream change can affect future human expression, trust, resource access, language norms, model training, or institutional templates.

4.2 Exposure, Acceptance, and Editing

This subsection makes the human–interface junction observable. Its factorization separates the expressions a model could produce from suggestions actually shown, decisions people make, and final artifacts that enter a population distribution.

Let $\mathcal{X}$, $\mathcal{S}$, and $\mathcal{Y}$ be finite source, suggestion, and final-expression category sets. For user i and context c , let $p_i(x \mid c)$ be the source distribution, $q_\psi(s \mid x, c)$ the suggestion-exposure distribution, and $a_i(y \mid x, s, c, \psi)$ the human decision and editing kernel. Then

$$r_{i,\psi}(y \mid c) = \sum_{x \in \mathcal{X}} p_i(x \mid c) \sum_{s \in \mathcal{S}} q_\psi(s \mid x, c) a_i(y \mid x, s, c, \psi). \quad (4)$$

For weights w_i summing to one,

$$\bar{r}_\psi(y \mid c) = \sum_i w_i r_{i,\psi}(y \mid c). \quad (5)$$

Equation (4) is a total-probability factorization. Causal interpretation needs randomized exposure, credible adjustment, or another identification design. The variables identify distinct intervention sites. Changing the model can alter q_ψ ; changing interface order can alter exposure while model capacity stays fixed; training and literacy can alter a_i ; institutional penalties can alter acceptance through anticipated downstream consequences.

The factorization also locates missing data. A study observing final text without suggestion logs cannot estimate exposure. A product observing acceptance without unaided comparison can confound user preference and model influence. A study measuring acceptance within one task leaves later unaided production open. Privacy-preserving designs may deliberately omit detailed logs, which changes the estimable questions and should be reported.

4.3 Agency and Role Allocation

This subsection distributes causal and normative roles across stages. It distinguishes operation, endorsement, authorship, interpretation, institutional authorization, and accountability.

At least six roles can attach to a mediated artifact: source participant, trace producer, model or tool provider, prompting or selecting user, editor or endorser, and institutional user. One actor can occupy several roles, and one role can be distributed. Machine agency in the causal or perceived sense can be substantial when suggestions are ranked, personalized, optimized, or automatically inserted. Human agency can also remain substantial through task choice, prompting, selection, editing, refusal, and later use.

A stage-specific allocation avoids two extremes. A product provider’s optimization and default choices can influence exposure even when the user presses the final send button. A user can endorse a

message after substantial model contribution. An institution can acquire independent responsibility when it relies on the artifact for a consequential decision. The exact allocation depends on knowledge, control, foreseeability, mandate, alternatives, and the relevant normative or legal framework.

4.4 Disclosure and Audience Interpretation

This subsection places disclosure within the audience’s interpretive conditions. Its objective is usable context. Universal labels detached from stakes and audience needs have limited interpretive value.

Disclosure can identify system use, degree of transformation, model version, source type, confidence, or available history. Its interpretation depends on task and convention. Minor spelling assistance may carry limited relevance in ordinary correspondence. Translation, synthetic testimony, attributed quotations, institutional evidence, or an AI summary presented as a participant voice can carry substantial relevance. Accessibility users may rely on extensive assistance, making crude human-versus-AI labels misleading or stigmatizing.

The appropriate object is a use-specific disclosure profile: information needed by a reasonable audience to interpret authorship, endorsement, reliability, source relation, and uncertainty within the stakes of the interaction. Provenance infrastructure can support this profile, while interface design determines whether the information becomes usable.

4.5 Institutional Selection and Downstream Use

This subsection extends mediation beyond the sender and receiver. It captures the filters through which some outputs acquire institutional force and enter future templates or data pools.

Institutions select among artifacts through style rules, evaluation rubrics, classifiers, publication standards, search ranking, hiring criteria, administrative forms, and archival practices. A system can offer diverse expressions while an institution accepts a narrow register. Conversely, a common system output can enter institutions that interpret and revise it through plural local procedures. Population-level output distributions therefore combine interface supply, human decisions, and downstream gates.

The downstream use stage also determines consequence. A generated paraphrase used privately for comprehension differs from a generated summary used to determine benefits or establish an official record. The same text can carry different authority, contestability, and provenance requirements across these settings. Later sections preserve this use sensitivity in empirical and governance design.

5 Expression Distribution and Accessible Repertoire

This section develops the paper’s distributional core. It separates individual effective repertoire from population output, derives one-step concentration change, proves an exact coexistence result, and states the measurement consequences for heterogeneous users and languages.

5.1 Population Distribution

This subsection introduces a finite diagnostic for accepted expression. Its purpose is to study movement across declared expression categories while preserving the limitations of category construction.

Partition a task-specific expression space into n categories. Let $p \in \Delta^{n-1}$ be the pre-mediation population distribution, $q \in \Delta^{n-1}$ the distribution of accepted AI-mediated forms, and $\eta \in [0, 1]$ the effective uptake share. One mediation step gives

$$p' = (1 - \eta)p + \eta q. \quad (6)$$

The representation can be recovered as a special case of Equations (4) and (5) after aggregating users and decisions. It is deliberately simple. The accepted distribution q already reflects model suggestions, exposure, human selection, editing, and the category system used by the analyst.

Let Herfindahl concentration be $C(p) = \sum_{k=1}^n p_k^2 = \|p\|_2^2$. Expansion of the squared norm gives

$$C(p') - C(p) = 2\eta \langle p, q - p \rangle + \eta^2 \|q - p\|_2^2. \quad (7)$$

Proposition 5.1 (Conditional concentration direction). *Under Equation (6), the direction of concentration change depends on the relation among p , q , and η . Movement toward the uniform distribution weakly lowers Herfindahl concentration, while movement toward a point mass on a category already modal in p weakly raises it.*

Proof. The identity in Equation (7) follows from $p' = p + \eta(q - p)$. Let $u = (1/n, \dots, 1/n)$. Convexity of squared norm and the minimality of $C(u)$ on the simplex imply

$$C((1 - \eta)p + \eta u) \leq (1 - \eta)C(p) + \eta C(u) \leq C(p).$$

For a modal category $m \in \arg \max_k p_k$, let $q = e_m$. Since

$$C(p) = \sum_k p_k^2 \leq p_m \sum_k p_k = p_m,$$

we have $\langle p, e_m - p \rangle = p_m - C(p) \geq 0$. Both terms in Equation (7) are then nonnegative. Strictness depends on uptake and the initial distribution. $\square$

The proposition rules out a universal direction inside the model. It also shows why model-level vocabulary size gives limited information. A system can make many strings technically generable while exposure and acceptance cluster around a small region. A second system can spread probability across categories while reducing task quality or increasing correction burden. Distribution and value remain different objects.

5.2 Individual Effective Repertoire

This subsection defines accessibility as feasible competent use by a particular person. The definition includes time, effort, comprehension, correction, and social or institutional cost.

Let Ω be a task-specific expression universe. Let $\kappa_i(y; c, \psi)$ be the least effective cost to user i of producing and competently deploying expression y in context c under system state ψ . For feasible-cost threshold τ_i , define

$$\Omega_i^{\text{eff}}(c, \psi; \tau_i) = \{y \in \Omega : \kappa_i(y; c, \psi) \leq \tau_i \wedge \text{Competent}_i(y, c)\}. \quad (8)$$

The cost can combine drafting time, cognitive effort, literacy, language proficiency, disability access, subscription cost, correction burden, privacy risk, and anticipated institutional penalty. The competence predicate can require comprehension, appropriate deployment, detection of significant errors, and capacity to revise or refuse. Its demands should track the task. Producing a polite greeting and submitting a clinical report call for different criteria.

The definition avoids equating availability with access. A language can appear in a model interface while output quality, latency, price, script support, privacy, and correction burden make it practically unusable. A user can gain a feasible formal register through assistance while still losing idiolectal features in the accepted artifact. The repertoire set and fidelity profile should therefore be reported together.

5.3 Expansion–Concentration Coexistence

This subsection provides an exact witness for the paper’s central multi-level claim. It demonstrates logical compatibility between strict individual expansion and strict aggregate concentration within one intervention.

Proposition 5.2 (Expansion–concentration coexistence). *There exists a finite AI-mediated intervention that strictly expands one user’s effective repertoire and strictly increases Herfindahl concentration in the population’s accepted expression distribution.*

Proof. Let $\Omega = \{1, 2, 3\}$. For one focal user, suppose

$$\Omega_i^{\text{eff}}(c, \psi_0; \tau_i) = \{1, 2\}, \quad \Omega_i^{\text{eff}}(c, \psi_1; \tau_i) = \{1, 2, 3\}.$$

The intervention strictly expands that user’s effective repertoire.

Let the pre-mediation population distribution be

$$p = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right),$$

and let accepted AI-mediated forms follow $q = e_1 = (1, 0, 0)$ with uptake $\eta = 1/2$. Equation (6) gives

$$p' = \frac{1}{2}p + \frac{1}{2}q = \left(\frac{3}{4}, \frac{1}{8}, \frac{1}{8}\right).$$

The concentrations are

$$C(p) = \frac{1}{4} + 2\frac{1}{16} = \frac{3}{8}, \quad C(p') = \frac{9}{16} + 2\frac{1}{64} = \frac{19}{32}.$$

Therefore

$$C(p') - C(p) = \frac{19}{32} - \frac{12}{32} = \frac{7}{32} > 0.$$

The same constructed intervention contains strict individual repertoire expansion and strict population concentration. $\square$

The proposition establishes a possibility result; prevalence remains empirical. Its significance lies in the evaluation architecture. A study reporting only aggregate concentration can miss accessibility gains. A study reporting only user productivity can miss concentration. A policy framed as a choice between accessibility and diversity can overlook designs that distribute gains and losses across levels, groups, and time.

5.4 Heterogeneous Effects and Category Politics

This subsection states the conditions under which the preceding quantities become empirically meaningful. It focuses on group heterogeneity, category construction, and multiple distributional diagnostics.

For groups $g \in \mathcal{G}$ with weights ω_g , define group distributions p_g and the population mixture $\bar{p} = \sum_g \omega_g p_g$. The report should retain both within-group and between-group change. A fall in population concentration can accompany increased concentration within each group when groups move toward different modes. A rise in population concentration can accompany substantial repertoire gains among users who previously faced high costs.

The category system also carries theory and power. Categories can represent lexical items, syntactic constructions, politeness strategies, dialect features, metaphors, discourse structures, semantic frames, or relational-historical indices. A dominant-language parser can treat a local construction as error and thereby make “diversity” partly a measure of distance from the dominant standard. Community-defined and analyst-defined category systems should be compared. Sensitivity analysis should report conclusions across Herfindahl concentration, entropy, support size, pairwise distance, and domain-specific qualitative coding.

Table 2 organizes the principal measurements. The table also states the interpretive boundary that prevents one metric from serving as a universal linguistic-value score.

5.5 Distributional Interpretation

This subsection connects the formal results to normative analysis while preserving the required bridge premises. It identifies conditions under which concentration or dispersion may have different meanings.

Concentration can reflect imposed normalization, shared task convention, safety vocabulary, successful mutual accommodation, institutional coercion, or voluntary coordination. Dispersion can reflect expressive plurality, innovation, error, fragmentation, low quality, or unequal exclusion. A normative conclusion therefore needs at least a value object, affected parties, causal mechanism, comparison condition, distribution of benefit and burden, and feasible alternatives.

The paper’s narrower claim concerns measurement sufficiency. Individual access, aggregate distribution, semantic fidelity, relational-historical fidelity, and community trajectory agency are jointly relevant and mutually irreducible. Their conflicts should remain visible in the result and across any proposed aggregation.

Table 2: Distributional and capability measurements across analytical levels

Measurement level	Candidate quantity	Supported interpretation	Retained boundary
individual repertoire	$ \Omega_i^{\text{eff}} $, cost frontier, competent-use transfer	feasible expressions for a declared person and task	leaves population uptake and social value open
interface exposure	support and concentration of $q_\psi(s x, c)$	suggestions presented under a context	leaves acceptance and editing open
individual output	support, entropy, concentration, feature rates of r_i	accepted output for a user and task	leaves durable repertoire change open
group output	within-group distribution and change	group-specific uptake and expression pattern	leaves internal subgroup difference open
population output	$C(\bar{r})$, entropy, support, pairwise similarity	aggregate accepted-expression structure	leaves individual access and fidelity open
cross-time change	transition, persistence, withdrawal and transfer estimates	temporary and durable movement	leaves causal mechanism open without design
fidelity	semantic and relational-historical discrepancy	source-sensitive transformation quality	leaves diversity and standing open
institutional uptake	gate-specific acceptance and consequence	downstream force of mediated artifacts	leaves interface availability open

6 Recurrent Human–AI Dynamics

This section extends one-step mediation into a coupled temporal system. It separates human accommodation, system updates, institutional selection, and data-pool feedback, then derives a conditional local stability result and states empirically discriminating pathway hypotheses.

6.1 Coupled State and Transition

This subsection defines the recurrent state. Its role is to make each feedback-bearing component visible and prevent a fixed model distribution from standing for a changing deployment.

Let

$$z_t = (p_{1,t}, \dots, p_{N,t}, \psi_t, \kappa_t, D_t) \quad (9)$$

contain user expression distributions, model and interface state, institutional selection state, and a provenance-bearing data pool. A general stochastic transition is

$$z_{t+1} = \Phi(z_t, u_t, \xi_{t+1}), \quad (10)$$

where u_t is a design or governance intervention and ξ_{t+1} collects exogenous events. One explicit decomposition is

$$p_{i,t+1} = (1 - \lambda_i)p_{i,t} + \lambda_i r_i(\cdot \mid z_t, u_t), \quad (11)$$

$$\psi_{t+1} = G(\psi_t, D_t, u_t, \xi_{t+1}^\psi), \quad (12)$$

$$\kappa_{t+1} = H(\kappa_t, \bar{r}_t, o_t, u_t, \xi_{t+1}^\kappa), \quad (13)$$

$$D_{t+1} = L(D_t, \bar{r}_t, \mathcal{P}_t, u_t, \xi_{t+1}^D). \quad (14)$$

The adaptation rate λ_i summarizes the degree to which current accepted expression enters a user's later distribution. It can vary by user, task, frequency, confidence, and social context. The operator G includes model retraining, fine-tuning, retrieval updates, moderation, and interface releases. The operator H includes institutional changes to templates, rubrics, accepted registers, and uses. The operator L includes data collection, filtering, licensing, provenance selection, deletion, and synthetic-data treatment. The provenance filter $\mathcal{P}_t$ can determine which artifacts enter which future use.

6.2 Local Stability and Feedback Gain

This subsection supplies a conditional mathematical result for a stable intervention regime. Its purpose is diagnostic: it identifies when small perturbations decay or grow near a fixed state and preserves the gap between dynamic stability and social value.

Assumption 6.1 (Local transition regularity). *For a fixed intervention u , the deterministic map Φ_u is continuously differentiable in a neighborhood of a fixed point $z^* = \Phi_u(z^*)$.*

Let $\delta z_t = z_t - z^*$ and $J^* = D_z \Phi_u(z^*)$. Taylor expansion gives

$$\delta z_{t+1} = J^* \delta z_t + o(\|\delta z_t\|). \quad (15)$$

Proposition 6.2 (Conditional local convergence). *If an induced norm satisfies $\|D\Phi_u(z)\| \leq \rho < 1$ throughout a neighborhood of z^* that maps into itself, then Φ_u is a contraction on that neighborhood and trajectories beginning there converge to the unique local fixed point. If $\rho(J^*) > 1$, the fixed point is locally unstable.*

Proof. The mean-value inequality gives

$$\|\Phi_u(z) - \Phi_u(z')\| \leq \rho \|z - z'\|$$

for points in the neighborhood. The contraction mapping theorem yields local uniqueness and convergence. When $\rho(J^*) > 1$, the linearized system has an expanding direction; differentiability transfers local instability to the nonlinear map under the standard fixed-point conditions. $\square$

Convergence can lead toward a plural, accessible, high-fidelity regime or toward a narrow and difficult-to-exit regime. Instability can represent harmful volatility or useful movement away from an imposed standard. Dynamic classification and normative assessment therefore occupy separate columns of the audit.

For the pathway

$$p_{i,t} \longrightarrow \bar{r}_t \longrightarrow D_{t+1} \longrightarrow \psi_{t+1} \longrightarrow r_{i,t+1},$$

Table 3: Feedback edges, mechanism evidence, and interpretation limits

Feedback edge	Mechanism	Discriminating evidence	Interpretation limit
human accommodation	exposure and accepted forms alter later unaided production	repeated measures, withdrawal, transfer tasks, exposure randomization	temporary priming can mimic durable change
social diffusion	mediated forms pass among interlocutors or peers	network timing, partner assignment, diffusion paths	selection and homophily can mimic influence
institutional normalization	accepted artifacts alter templates, rubrics, and gates	policy changes, process tracing, threshold designs, evaluator records	broader professional change can coincide
system update	data and feedback alter model or interface behavior	versioned evaluation, update logs, controlled model comparison	hidden releases and data changes impede attribution
training-data recursion	mediated artifacts enter later training pools	data provenance, mixture records, ablations, generational experiments	model recursion differs from human language adaptation

a local route-gain candidate is

$$g_{i,D,\psi} = \frac{\partial r_i}{\partial \psi} \frac{\partial G}{\partial D} \frac{\partial L}{\partial \bar{r}} \frac{\partial \bar{r}}{\partial p_i}, \quad (16)$$

evaluated at z^* with compatible dimensions. The product isolates one proposed feedback route. Parallel routes, delays, cancellation, threshold behavior, and omitted state remain in the full Jacobian.

6.3 Feedback-Edge Register

This subsection converts the recurrence into mechanism-specific evidence obligations. Table 3 distinguishes four edges that can produce similar observed output trajectories.

The human accommodation edge includes lexical, syntactic, pragmatic, and stylistic uptake. A durable effect should appear in later production under reduced assistance or in a different context. The social diffusion edge involves partners and audiences who may adopt, resist, parody, or reinterpret mediated forms. The institutional edge changes what receives rewards, passes review, or enters official records. The system edge changes suggestion distributions through retraining or product design. The training-data edge creates recursion among model generations, as studied under particular conditions by Shumailov and colleagues [25].

6.4 State-Dependent Direction

This subsection states the main dynamic consequence. It replaces a universal homogenization or diversification thesis with edge-specific, state-dependent hypotheses.

Suppose a system offers a minoritized-language register with high correction quality to users who previously faced high production cost. Individual repertoires can expand, group use can increase,

and population concentration can fall if accepted outputs spread across previously underrepresented categories. Suppose the same interface ranks a narrow standardized register first for most users and institutions reward that register. Population concentration can rise, relational-historical fidelity can fall, and some individual users can still gain access to formal communication. Suppose a community-directed model increases local forms while separating the community from cross-group communication. Within-group diversity and between-group alignment can move in different directions.

These paths generate testable hypotheses:

Hypothesis 6.3 (Exposure and durable uptake). *Repeated exposure to accepted AI suggestions predicts later unaided use of the exposed features after controlling for baseline repertoire, with persistence varying by correction, reward, social uptake, and user dependence.*

Hypothesis 6.4 (Institutional gate amplification). *The effect of suggestion concentration on accepted population output is larger where institutions reward a narrow register and users face high penalties for deviation.*

Hypothesis 6.5 (Resource-sensitive accessibility). *Assistance produces larger effective-repertoire gains for users facing high baseline expression costs when model quality, correction support, privacy, and institutional acceptance reach task-appropriate thresholds.*

Each hypothesis can fail. Users can resist suggestions, institutions can diversify standards, local communities can repurpose tools, and model releases can disrupt recurrence. Failed predictions refine the mechanism and receive substantive analysis.

7 Fidelity, Provenance, and Interpretive Standing

This section integrates source-sensitive fidelity, transformation history, and authority. It defines semantic and relational-historical discrepancy, represents provenance as an event graph, proves output-only non-identification, and states the separate warrants required for interpretive standing.

7.1 Semantic and Relational-Historical Fidelity

This subsection distinguishes propositional or task-relevant preservation from preservation of speaker position, relational history, register, uncertainty, and audience orientation.

For source expression x , mediated artifact y , and source context c , define

$$\mathbf{d}(x, y; c) = (d_{\text{sem}}(x, y; c), d_{\text{rel}}(x, y; c)), \quad (17)$$

where d_{sem} is a validated task-relevant semantic discrepancy and d_{rel} is discrepancy in speaker position, relational history, register, uncertainty, cultural indexicality, and addressed audience. Each coordinate can be a vector. The context determines which features matter.

A rewrite can preserve the central proposition while replacing tentative language with certainty, changing respectful distance into informality, erasing a local metaphor, or converting participant disagreement into a single institutional voice. Such a rewrite can be useful for one target and harmful for another. A public plain-language summary and a source-faithful archive serve different

functions. The framework asks each use to declare its target and retain the source where revision or accountability requires it.

Operationalization should combine automatic measures, source comparison, qualified human review, participant interpretation, and adversarial examples. Semantic embeddings can miss pragmatic stance. Surface feature counts can miss meaning. Participant judgments can vary and can reflect internal power. Metric disagreement is therefore reportable evidence.

7.2 Transformation-Event Graph

This subsection represents provenance as a branching history. The graph accommodates multiple sources, repeated model and human operations, composition, selective disclosure, and correction.

For artifact y , define a directed acyclic graph

$$\mathcal{G}_y = (V_y, E_y), \quad v_j = (a_j, \rho_j, \mu_j, \iota_j, o_j, t_j, \nu_j, \pi_j), \quad (18)$$

where a_j is an actor or system, ρ_j a role, μ_j a transformation type, ι_j input commitments or identifiers, o_j an output commitment or identifier, t_j time, ν_j model or tool version, and π_j permissions and disclosure state. Edges record declared derivation or composition.

The graph generalizes a linear edit history. A summary can combine several interviews, a model-generated outline, human edits, a translated version, and an institutional template. A later correction can branch from the published artifact. Restricted source materials can remain access-controlled while a commitment records their role. Different audiences can receive different disclosure views.

Define the provenance profile

$$\mathbf{P}(\mathcal{G}_y) = (P_{\text{coverage}}, P_{\text{integrity}}, P_{\text{identity}}, P_{\text{lineage}}, P_{\text{disclosure}}, P_{\text{privacy}}, P_{\text{contest}}, P_{\text{retention}}). \quad (19)$$

Coverage concerns recorded relevant events. Integrity concerns tamper evidence. Identity concerns credentials of actors or systems. Lineage concerns declared derivation. Disclosure concerns audience-appropriate visibility. Privacy concerns exposure and inference risk. Contestability concerns correction and response. Retention concerns storage, deletion, and future availability. These coordinates can conflict.

7.3 Output-Only Non-Identification

This subsection proves the minimal provenance result. It shows why stylistic detection or final-text inspection cannot generally recover the full transformation history.

Let $\mathcal{H}$ be the admissible transformation-history set and let $T : \mathcal{H} \rightarrow \mathcal{Y}$ map each history to its final artifact.

Proposition 7.1 (Output-only provenance non-identification). *Terminal output identifies transformation history exactly only when T is injective on the admissible history class. Any class containing direct human production of y and accepted AI generation of the same y is non-injective.*

Proof. Choose

$$h_1 = (\text{human directly writes } y)$$

and

$$h_2 = (\text{model proposes } y, \text{ human accepts } y).$$

The histories differ in actor and operation events, so $h_1 \neq h_2$. Their terminal artifacts satisfy $T(h_1) = T(h_2) = y$. Hence T is non-injective, and observation of y alone identifies at most the equivalence class $T^{-1}(y)$. $\square$

The result extends immediately. Different prompts, model versions, source drafts, editing paths, translation routes, and institutional templates can converge on the same artifact. An output classifier can estimate features correlated with selected histories under a reference distribution. Such estimation remains distinct from exact historical recovery.

7.4 Technical Provenance and Interpretive Evidence

This subsection bounds the contribution of technical standards. It connects cryptographic association and tamper evidence to the broader provenance profile.

C2PA represents assertions, actions, ingredients, claims, signatures, manifests, content bindings, trust, and validation [10]. These mechanisms can support integrity and lineage coordinates in Equation (19). They can also enable selective provenance through embedded or external records and user-controlled inclusion. A valid manifest indicates that recorded assertions are associated with the asset and satisfy relevant validation rules under the trust model.

Several questions remain external to that result. A signed assertion can be incomplete or contextually misleading. A signer can lack authority to represent a source community. A valid action record can reveal sensitive identity or restricted knowledge. A complete technical history can overwhelm audiences. A stripped manifest can reflect platform failure, malicious removal, privacy protection, or ordinary conversion. Governance should therefore join integrity with coverage, disclosure, privacy, contestability, and retention.

7.5 Interpretive Standing and Institutional Authority

This subsection defines the warrant needed when an AI-mediated artifact is used to speak about a source relation or support an institutional conclusion. It distinguishes accuracy from representative and decision authority.

Interpretive standing is target- and role-specific. Candidate bases include participation, relevant expertise, evidential access, appointment, community mandate, procedural authorization, accountability, independence, and demonstrated correction. A model can contribute useful pattern detection or synthesis without occupying every standing-bearing role. A human participant can possess distinctive access while remaining fallible, partial, conflicted, or unauthorized to speak for others.

For use u and actor or system a , define a non-aggregated standing profile

$$\mathbf{S}(a, u) = (S_{\text{part}}, S_{\text{expert}}, S_{\text{access}}, S_{\text{mandate}}, S_{\text{account}}, S_{\text{correct}}, S_{\text{independ}}). \quad (20)$$

The coordinates record participation, expertise, evidence access, mandate, accountability, correction, and relevant independence. The legitimate threshold depends on use and consequence. A

brainstorming aid can tolerate a different profile from a system summarizing testimony for an adjudicator.

Institutional authority requires a related use profile: jurisdiction, procedure, competent review, disclosure, opportunity to respond, appeal, record correction, and responsibility allocation. Passing an artifact through an authorized system gives it causal force. The authorization of the system and the validity of each output remain separately reviewed.

7.6 Selective Provenance and Privacy

This subsection states the design tension between inspectable history and protection from exposure. It proposes tiered and purpose-bound provenance in place of maximal collection.

Useful provenance can record that generative assistance occurred, which transformation class occurred, which system version was used, which source classes contributed, who endorsed the final artifact, and where a fuller record can be contested. Sensitive prompts, identities, drafts, locations, and restricted knowledge can remain minimized, encrypted, access-controlled, committed by hash, or excluded according to legitimate protocol. Retention periods can vary by consequence.

The proper design depends on threat model. Public media provenance emphasizes durable association and audience disclosure. Workplace writing assistance emphasizes privacy and anti-surveillance. Indigenous language data can require collective authority and restricted access. Institutional evidence can require chain of custody, source access, adversarial testing, and appeal. Accessibility uses require disclosure that avoids stigmatizing legitimate assistance. The plural provenance profile keeps these designs comparable without imposing one maximal record.

8 Comparative Mechanism Cases

This section applies the framework to three constructed cases selected for mechanism contrast. The first concerns rewriting and professional mobility, the second concerns minoritized-language assistance and community capacity, and the third concerns institutional summaries of shared experience. The cases illustrate hypotheses and audit requirements; specific deployment findings remain an empirical task.

8.1 Professional Rewriting and Local Voice

This subsection examines an assistant that rewrites a user’s draft into a prestigious professional register. It separates immediate access gains, source-sensitive transformation, population effects, and institutional reinforcement.

Consider a multilingual worker who understands a technical problem and drafts an accurate report in a local or personally characteristic English register. A writing assistant proposes a concise standardized version. The final report passes an institutional review that historically penalized the original register. At the individual level, drafting time and penalty risk fall. The effective repertoire expands because an institutionally accepted report becomes feasible. At the artifact level, clarity and task accuracy can improve.

The same chain can alter other coordinates. The assistant may replace uncertainty markers with

categorical statements, remove culturally situated analogies, or make many users’ reports stylistically similar. Reviewers can update their expectations toward the polished standard, raising penalties for unaided or locally marked writing. Users can internalize the proposed syntax through repeated exposure. Later model training can ingest accepted reports and further increase the probability of the same register.

The mechanism has four separable edges: suggestion exposure, user decision, institutional reward, and recurrent adaptation. A controlled interface experiment can estimate immediate suggestion effects. Longitudinal unaided writing can estimate persistence. A policy or rubric change can reveal institutional reinforcement. Versioned model evaluation can examine product drift. The user’s own preference matters throughout. Some users seek retention of idiolect; others seek anonymity, mobility, or deliberate register change. A legitimate design offers control over that objective.

Provenance should be proportional to use. A private drafting tool may need local history and clear user control. A safety report used in a consequential process may need source retention, material edit disclosure, model version, final human endorsement, and a correction route. A label stating only “AI assisted” provides limited help with which claims changed or who accepted them.

8.2 Minoritized-Language Assistance and Community Capacity

This subsection examines a system that supports a language with limited digital resources. It places accessibility expansion alongside data governance, correction labor, dependency, and future generativity.

Suppose a community-directed project develops speech recognition, translation, and writing support for a minoritized language. New users gain keyboard, transcription, accessibility, educational, and public-service options. Speakers can create new digital content and communicate across scripts or proficiency levels. These are substantial capability and continuity possibilities, especially where the previous effective repertoire was constrained by absent tools.

The deployment also carries structural risks. Training data can be extracted without legitimate authority. Dialect coverage can be uneven. Fluent speakers can perform unpaid correction labor. A model can elevate one standardized variety and reduce the visibility of others. Dependence on a remote provider can expose the community to price changes, service withdrawal, or inaccessible model updates. Restricted narratives can enter training or public output. An apparently successful benchmark can conceal correction burden and downstream institutional rejection.

Bird’s decolonial critique, the disparities documented by Joshi and colleagues and by Blasi and colleagues, Masakhane’s participatory research practice, and CARE’s authority and benefit principles provide established constraints on this case [7, 20, 8, 23, 9]. The project should identify legitimate community procedures, include internal language variation, preserve individual privacy and mobility, return capacity and infrastructure, and make withdrawal or migration feasible.

The appropriate outcomes span several horizons. Immediate evaluation measures recognition, translation, generation, correction burden, latency, and accessibility. Medium-term evaluation measures uptake, domain expansion, teaching support, institutional acceptance, and distribution across varieties. Long-term evaluation measures successor formation, local technical capacity, data and model governance, dependence, innovation, and the ability to revise the technology. Community-defined objectives guide selection among these outcomes.

This case illustrates coexistence in a favorable and an adverse direction. A tool can expand

individual and community repertoires while concentrating output around one standardized variety. A plural model can increase distributional variety while remaining too inaccurate for consequential use. An intervention can improve translation access while weakening local control over data. A multi-coordinate report makes these tradeoffs visible.

8.3 Institutional Summaries of Shared Experience

This subsection examines AI-generated summaries of interviews, meetings, consultations, or collaborative practice that later support an institutional decision. It concentrates the paper’s fidelity, provenance, standing, and authority distinctions.

Imagine several people who share a difficult workplace experience. Their accounts include agreement, uncertainty, role-specific observations, and conflicting interpretations. An institution uses a generative system to summarize transcripts. The system produces a coherent report that captures many recurring propositions. Decision makers read the report and use it to redesign policy.

Artifact performance can be evaluated against the transcripts for propositional coverage, contradiction, uncertainty, and attribution. Relational-historical fidelity requires a further comparison: whether the summary preserves role differences, temporal sequence, conflict, silences, audience, and the conditions under which statements were made. Provenance requires source links, transformation history, model version, human edits, redactions, approval, and correction. Standing requires clarity about the system’s role. It can assist synthesis while participants, qualified investigators, and authorized decision makers retain different forms of interpretive and institutional authority.

The central risk is authority through fluency. A coherent summary can appear to settle disagreement, represent a group, or establish an institutional fact. The warrant profile interrupts that movement. High semantic coverage can coexist with low relational fidelity. Strong provenance can coexist with weak participant mandate. Participant mandate can coexist with factual error or internal dissent. Institutional authorization can coexist with a defective evidentiary process.

A responsible process supplies participants with meaningful opportunities to review relevant representations, retains dissent and uncertainty, separates source statements from model-generated synthesis, records material transformations, provides decision makers with access appropriate to the stakes, and preserves an appeal or correction path. Privacy and retaliation risk can justify redaction and tiered access. The resulting record can remain useful while carrying visible conditions.

8.4 Cross-Case Mechanism Matrix

This subsection compares the cases through a shared structure. Table 4 identifies the principal gain, concentration route, provenance need, authority issue, and decisive evidence for each case.

The matrix shows why one policy cannot govern every setting. A private writing tool, community language infrastructure, and institutional evidentiary summary carry different stakes, threat models, and legitimate authorities. The shared framework supplies common questions and keeps the answers case-sensitive.

Table 4: Mechanism comparison across three generative-AI mediation cases

Case	Primary possible gain	Principal concentration route	Provenance and authority focus	Decisive evidence
professional rewriting	feasible access to accepted register, time and correction savings	ranked standard style plus institutional reward	material edits, endorsement, safety relevance, user register choice	randomized suggestions, later unaided writing, rubric and gate effects
minoritized-language support	new tools, domains, accessibility, content, and learning support	uneven variety coverage, standardization, provider dependency	data authority, restricted knowledge, benefit, local capacity, withdrawal	variety-specific performance, correction burden, longitudinal uptake, community-governed outcomes
institutional summary	scalable synthesis, search, comparison, and decision support	compression of disagreement into fluent institutional voice	source links, transformations, mandate, correction, appeal, decision authority	proposition and relation-sensitive fidelity, participant review, process tracing, outcome audit

9 Empirical Research Programme

This section translates the framework into an empirical programme. It specifies units, estimands, controlled and longitudinal designs, participatory validation, institutional process analysis, model-update audits, measurement robustness, and ethical stopping conditions.

9.1 Units, Treatments, and Outcome Profiles

This subsection establishes the observational and causal units. It prevents a model name or binary assistance label from concealing treatment versions, user heterogeneity, and downstream institutional settings.

A study should specify users, groups, languages or varieties, tasks, genres, contexts, interaction partners, model versions, interfaces, suggestion policies, decision stages, institutions, and horizons. Treatment can include access to suggestions, ranking policy, personalization, disclosure, fidelity constraint, variety support, provenance display, or institutional acceptance rule. Each treatment version should be logged or controlled.

For outcome coordinate k , group g , and horizon h , define the causal estimand

$$\Delta_k(g, h) = \mathbb{E}[Y_k(u_1, h) - Y_k(u_0, h) \mid G = g], \quad (21)$$

under an identified intervention contrast (u_1, u_0) . The outcome Y_k can represent task performance, effective repertoire, concentration, fidelity, correction burden, trust, standing judgments, institutional acceptance, or later unaided expression. Interference should be declared because partners and peers can receive mediated language from treated users.

An outcome vector is preferable to a composite:

$$\mathbf{Y}_{i,t} = (Y^{\text{art}}, Y^{\text{access}}, Y^{\text{dist}}, Y^{\text{sem}}, Y^{\text{rel}}, Y^{\text{prov}}, Y^{\text{agency}}, Y^{\text{institution}}, Y^{\text{future}})_{i,t}. \quad (22)$$

Pre-registration should state primary outcomes, category systems, thresholds, missing-data rules, multiplicity handling, and subgroup analyses. Qualitative components can remain open to emergent categories while preserving an audit trail for interpretation.

9.2 Controlled Interface Studies

This subsection describes short-horizon experiments suited to exposure, acceptance, editing, and audience mechanisms. It also states their external-validity boundary.

Randomized studies can vary suggestion availability, order, distributional breadth, style, cultural framing, disclosure, provenance display, and editing friction. The interface should log suggestions shown, time visible, acceptance, edits, rejection, final output, and task duration where consent and privacy permit. A factorial design can separate model content from interface ranking and disclosure.

Immediate outcomes include speed, task quality, semantic and relational-historical fidelity, lexical and syntactic features, acceptance, correction, perceived authorship, trust, and audience interpretation. Baseline unaided tasks support within-person comparison. Partner-level randomization can estimate effects on interlocutors. Delayed tasks without assistance test persistence and transfer.

Existing predictive-text, smart-reply, co-writing, creativity, and cross-cultural experiments establish the feasibility of such designs [3, 17, 19, 12, 1]. P027’s added requirement is the joint measurement of individual repertoire, population distribution, fidelity, and history. The short horizon remains explicit.

9.3 Longitudinal Repertoire and Distribution Panels

This subsection targets durable human adaptation and changing population distributions. It uses repeated assisted and unaided tasks, system withdrawal, and contextual transfer.

A panel can observe users before deployment, during repeated use, after temporary withdrawal, and across new contexts. It should retain raw drafts, shown suggestions, accepted and rejected fragments, final artifacts, later unaided production, and user explanations under an ethically minimized record. System versions and institutional rule changes should be time-stamped.

The analysis can estimate individual feature transitions, effective-repertoire frontiers, persistence after withdrawal, population concentration, within- and between-group change, and correction

burden. Interrupted time series or staggered rollout can support causal inference when randomization is infeasible. Matched comparison sites can improve context control. Qualitative interviews can distinguish convenience, strategic compliance, learning, resistance, identity management, and institutional pressure.

Attrition deserves substantive analysis. Users who experience high correction burden, privacy concern, cultural mismatch, or low value may exit. An analysis limited to continuing users can overstate benefits and understate exclusion. Exit reasons and viable alternatives should be recorded with appropriate confidentiality.

9.4 Participation and Successor-Competence Studies

This subsection develops operational tests for claims that exceed artifact resemblance. It combines trace-only, multimodal, interactive, and situated conditions.

One experimental sequence can compare four training regimes: symbolic traces alone; multimodal traces; interactive simulation with consequence feedback; and situated participation with qualified human correction. Evaluation should include held-out artifact tasks, novel practical situations, breakdown detection, repair, adaptation after environmental change, explanation to a learner, and recognition of conditions that exceed competence.

The comparison asks which additions improve which outcome. Multimodal systems such as Flamingo and PaLM-E illustrate architectures that enlarge the evidence base [2, 13]. Situated-learning, tacit-knowledge, and distributed-cognition traditions motivate tests involving changing participation, correction, and coordinated action [21, 11, 18]. A practical successor claim should identify the target practice and transfer conditions. Interpretive standing requires a separate institutional and normative evaluation.

9.5 Participatory and Community-Governed Evaluation

This subsection specifies research procedures for language and cultural settings in which task definition, data control, and evaluation authority are themselves part of the object.

Affected speakers and communities should participate in defining desired capabilities, varieties, sensitive materials, evaluation tasks, error severity, benefit paths, access rules, and revision procedures. Participation should include compensation, capacity building, dissent recording, and opportunities to withdraw or change terms. Individual privacy and mobility remain protected within collective governance.

The evaluation can compare externally selected metrics with community-selected measures and record disagreement. It can test whether a model supports new speakers, fluent speakers, translators, educators, elders, youth, disabled users, and institutional workers differently. It can also track correction labor, data contributions, provider control, local technical capacity, and portability.

Participatory machine-translation research and CARE provide concrete antecedents for this design orientation [23, 9]. Their principles guide process; each project still requires local authority and context.

9.6 Institutional Process and Authority Studies

This subsection follows mediated artifacts through organizational gates. It evaluates how a text becomes evidence, policy input, public representation, or an official record.

Process tracing should record who commissions the artifact, defines the prompt, selects sources, reviews output, resolves discrepancies, authorizes use, communicates limitations, and handles appeal. Comparative designs can vary whether decision makers receive the final summary alone, linked sources, uncertainty markers, participant annotations, or full provenance. Outcomes include factual accuracy, recognition of disagreement, decision quality, decision time, perceived authority, correction, and burden on source participants.

Institutional rules can also be interventions. A procurement policy can require versioned model records and source linkage. A professional rubric can accept several registers. An archive can retain original and transformed versions under tiered access. A consultation process can give participants a review and response route. Evaluation should examine consequences as well as formal compliance.

9.7 Model-Update and Data-Pool Audits

This subsection addresses the model and training-data edges of recurrence. It separates generated-data composition from human linguistic adaptation.

Audits should record data-source classes, human and generated shares, filtering, deduplication, licensing, model lineage, fine-tuning, retrieval corpora, and update dates. Versioned evaluation can compare tail performance, rare forms, language varieties, calibration, and distribution across releases. Controlled mixture experiments can vary generated-data proportions and preserved source data.

Research on recursive training provides a concrete mechanism warning under specified regimes [25]. Generalization to a deployed model requires matching the training process, data mixture, model initialization, filtering, and evaluation. Generalization to human language requires a separate human and institutional study. Provenance enables these edges to be distinguished empirically.

9.8 Measurement Robustness and Rival Mechanisms

This subsection states the robustness obligations for a multi-level study. It requires sensitivity to category systems, metrics, treatment versions, and causal rivals.

At least three rival mechanisms can produce apparent homogenization: direct suggestion uptake, user selection into assistance, and institutional preference for a standard register. A fourth is composition change in the observed population. Accessibility gains can arise from genuine learning, temporary delegation, reduced interface friction, or institutional acceptance. Fidelity changes can arise from model transformation, human editing, source ambiguity, or evaluator disagreement.

The analysis should therefore use exposure logs, baseline measures, randomization or credible adjustment, negative controls, versioned systems, multiple category schemes, alternative distance metrics, qualitative interpretation, and withdrawal tasks. Measurement invariance across language groups should be tested where meaningful. The model should report uncertainty and failed operationalizations.

9.9 Research Ethics and Stopping Conditions

This subsection defines the ethical boundary of the empirical programme. It treats longitudinal language and provenance research as capable of producing surveillance, identity exposure, and institutional harm.

Continuous text logging can reveal health, politics, relationships, location, disability, minority identity, and restricted knowledge. Research design should minimize collection, separate identifiers, use access controls, define retention and deletion, and provide comprehensible consent. Group harms can arise even when individual records are de-identified. Community protocols and collective impact review may therefore be required.

Predeclared stopping or redesign conditions include disproportionate privacy loss, evidence of retaliation, severe uncorrected output harms, unexpected extraction of restricted materials, rising correction burden without benefit, institutional use beyond consent, or loss of a viable exit path. Research value alone supplies an insufficient warrant for continued exposure.

10 Governance and Design

This section converts the analytical and empirical results into a modular governance portfolio. It presents user control, fidelity-sensitive assistance, plural suggestions, provenance tiers, community authority, institutional procurement, and recurrent monitoring as selectable components matched to stakes.

10.1 Admissible Design Set

This subsection defines the safeguards that constrain technically available interventions. It separates ethical admissibility from evidential adequacy.

Let $\mathcal{U}$ be candidate system, research, or institutional designs. Define

$$\begin{aligned}\mathcal{U}^{\text{adm}} &= \{u \in \mathcal{U} : \mathbf{s}(u) \succeq \boldsymbol{\tau}\}, \\ \mathbf{s}(u) &= (\text{Consent, Refusal, Accessibility, Privacy,} \\ &\quad \text{Authority, Contest, Portability, Revision, Safety})(u),\end{aligned}\tag{23}$$

The predicates require procedures and conflict rules. Consent includes scope and treatment versions. Refusal includes functional alternatives. Accessibility includes disability, language, literacy, price, and correction burden. Authority concerns legitimate control over data and representation. Contestability includes correction and appeal. Portability includes usable records and switching. Revision includes system, policy, and community review. Safety includes task-specific harm thresholds.

For each admissible design, report an evidence profile

$$\mathbf{E}(u) = (E_{\text{construct}}, E_{\text{causal}}, E_{\text{longitudinal}}, E_{\text{heterogeneity}}, E_{\text{community}}, E_{\text{provenance}}, E_{\text{robustness}}, E_{\text{replication}}).\tag{24}$$

An admissible design can have weak evidence, and a powerful experiment can fail an ethical threshold. The two profiles should remain visible.

10.2 User-Controlled Fidelity Modes

This subsection formalizes an optional anti-normalization design. Its purpose is to give users control over semantic, idiolectal, cultural, relational, and mobility objectives.

For source x , context c , candidate output y , task loss $L_{\text{task}}(y)$, correction burden $B_i(y)$, and user-selected relational-fidelity tolerance ε_i , consider

$$\min_{y \in \mathcal{Y}} L_{\text{task}}(y) + \alpha_i B_i(y) \quad \text{subject to} \quad d_{\text{rel}}(x, y; c) \leq \varepsilon_i, \quad (25)$$

when the user activates the constraint. A different mode can prioritize a new register, anonymity, compression, accessibility, or cross-group intelligibility. The user can compare candidates and inspect material shifts in stance, uncertainty, attribution, or local form.

The constraint depends on an imperfect metric. It should therefore support user review and feature-level explanation. A strict preservation mode can retain unwanted markers or expose identity. A mobility mode can erase desired voice. Reversible choice and local adaptation provide a better governance structure than a universal preservation objective.

10.3 Plural Suggestion and Distribution Controls

This subsection addresses interface concentration. It proposes controls over suggestion support, ordering, and objectives while preserving task quality and user choice.

Interfaces can show several functionally adequate alternatives across register, directness, formality, dialect, metaphor, or uncertainty. Ranking can include a diversity regularizer after a task-validity threshold. Users can request local or community models, block styles, retain rejected suggestions locally, and choose whether their accepted output contributes to personalization or training.

Population audits should examine exposure as well as final output. A provider can publish versioned distribution cards for representative prompts and groups, including coverage, concentration, tail behavior, and known correction burdens. Institutions can revise rubrics that reward a single model-shaped style. These measures address different edges and should be tested independently.

10.4 Provenance Tiers

This subsection matches provenance detail to consequence and threat model. It integrates technical integrity with social disclosure, privacy, correction, and retention.

Four illustrative tiers are useful:

1. **Personal drafting tier:** local edit history, model version, user control, minimized provider retention, and easy deletion.
2. **Public communication tier:** disclosure of material generation or transformation, final human or organizational endorsement, accessible history summary, and correction route.
3. **Institutional evidence tier:** source linkage, transformation-event graph, versioned system record, reviewer identity, uncertainty, redaction log, access control, and appeal.

4. **Community-governed knowledge tier:** collective authority, culturally governed access, restricted-material handling, benefit conditions, local stewardship, and withdrawal or revision procedures.

C2PA-compatible records can support selected integrity and lineage functions [10]. Dataset and system documentation provide complementary context [4, 14, 22]. The full governance object spans artifact, dataset, system, institution, and affected community.

10.5 Community Authority and Individual Safeguards

This subsection specifies governance for collective language or cultural resources. It treats community authority as a structured process with internal plurality.

Community governance can control task definition, data contribution, access, sensitive categories, evaluation, model distribution, commercial use, benefit, and revision. Local stewards can maintain provenance, language resources, and model interfaces. Agreements can provide capacity transfer, revenue or benefit paths, portability, deletion, and post-project support.

Internal dissent deserves an explicit route. Communities contain age, gender, class, dialect, religious, geographic, disability, and political differences. Governance should identify representation procedures, conflicts of interest, individual privacy, personal mobility, and appeals. CARE’s principles supply a relevant foundation while local institutions determine application [9].

10.6 Institutional Procurement and Use Controls

This subsection directs attention toward organizations that give mediated artifacts downstream force. It makes procurement, validation, and appeal part of linguistic-mediation governance.

Procurement can require documented intended uses, language and subgroup evaluations, version-change notice, provenance support, privacy and retention terms, portability, incident reporting, audit access, and exit assistance. Institutions can validate each consequential use independently from a vendor’s general performance claim. Human review should carry a defined competence, workload, authority, and responsibility structure.

Use controls can prohibit unsupported representative claims, require linked sources for summaries, preserve dissent, disclose uncertainty, and maintain appeal. Evaluation rubrics can recognize several legitimate registers. Accessibility accommodations can receive protection from stigmatizing disclosure. Model updates can trigger renewed validation when the treatment changes materially.

10.7 Recurrent Monitoring and Revision

This subsection governs the temporal system introduced in Section 6. It connects monitoring to predeclared review, response, and exit.

Monitoring should cover exposure concentration, accepted-output distribution, individual access, correction burden, fidelity, complaints, appeals, language-group performance, model versions, institutional gates, data-pool composition, and dependency. Thresholds can trigger investigation, interface change, rollback, added support, or suspension. Monitoring data should follow minimization and access rules.

Table 5: Governance portfolio for generative-AI linguistic mediation

Governance component	Principal mechanism target	Intended function	Design risk
user-selectable fidelity mode	editing and accepted output	preserve chosen stance, register, history, or mobility goal	metric error; unwanted identity retention
plural suggestion interface	exposure distribution	widen visible alternatives and user comparison	overload; token diversity without functional difference
versioned distribution audit	system and interface update	detect concentration, gaps, and subgroup change	category bias; surveillance
selective provenance	artifact transformation	support interpretation, attribution, and correction	privacy loss; false confidence in partial records
community governance	data, task, evaluation, benefit	align technology with legitimate collective authority	internal exclusion; elite capture
institutional source linkage	downstream use	preserve evidentiary context and contestability	administrative burden; sensitive-source exposure
portability and exit	provider dependency	preserve migration and refusal	incomplete interoperability; lost context
longitudinal review	recurrent feedback	detect durable human, institutional, and model change	measurement burden; delayed response
independent audit and appeal	authority and consequence	create external scrutiny and remedy paths	symbolic compliance; inaccessible procedure

Stable metrics can conceal changing populations or suppressed exit. Review should include qualitative inquiry and outreach to users who stopped using the system. Revision procedures should preserve earlier versions or records needed for audit and allow communities and institutions to migrate. A monitoring system becomes legitimate through measurement joined to response capacity.

10.8 Governance Portfolio

This subsection consolidates the design options and their mechanism targets. Table 5 links each component to the edge it changes and the risk it introduces.

The portfolio supports combination and revision. High-stakes institutional summaries may need provenance, source linkage, qualified review, appeal, and strict privacy. Personal low-stakes drafting may emphasize local control and minimal retention. Community language infrastructure may emphasize authority, capacity, portability, benefit, and long-term stewardship.

11 Objections and Scope Conditions

This section tests the framework against conceptual, empirical, and normative objections. Each response narrows the claim, adds an evidence obligation, or identifies a condition under which revision is required.

11.1 Conceptual Redundancy

This subsection considers whether grounding, situated cognition, AI-mediated communication, and provenance already exhaust the contribution.

The proposed terms overlap with established work, and that overlap is intellectually necessary. The potential contribution lies in a compositional architecture joining layered warrants, exposure and decision kernels, individual and population measures, recurrent feedback, event-graph provenance, and institutional standing. The framework earns continued use only if this composition discriminates cases, improves study design, or localizes governance more effectively than the antecedents used separately. Failure on those tests calls for narrower presentation or abandonment.

11.2 Human Exceptionalism

This subsection addresses the possibility that participation and standing distinctions preserve human priority through definition.

Participation criteria should center on interventionally relevant relations, feedback, and practice. A nonhuman system could satisfy selected participation or practical-competence criteria when evidence supports them. Representative and institutional standing still depend on mandate, accountability, role, and affected-party governance. Moral status and legal personality require their own arguments. The framework preserves open criteria and avoids assigning every warrant through a benchmark score.

11.3 Functional Sufficiency

This subsection considers contexts in which reliable task performance may be enough for deployment.

Bounded functional sufficiency is plausible for many low-stakes uses. A spelling correction or private paraphrase can be judged by task success, user control, and ordinary safety. Broader evidence becomes relevant when the artifact claims source-specific understanding, transfers to novel practical situations, represents participants, or carries institutional consequence. The framework scales its demands with the asserted claim and use.

11.4 Metric Politics

This subsection addresses the power embedded in expression categories, competence thresholds, and fidelity metrics.

Every operationalization selects features and standards. Dominant-language tools can treat local forms as defects, and community categories can conceal internal variation. The response is methodological pluralism with authority and sensitivity analysis: compare category schemes, document who defined them, report disagreement, include qualitative interpretation, and preserve raw or source-linked material where ethically permitted. A disputed metric should appear as a disputed instrument.

11.5 Accessibility and Stylistic Mobility

This subsection considers the risk that preserving idiolect or cultural features becomes a paternalistic constraint.

Users can reasonably seek a prestigious register, anonymity, reduced stigma, cross-group intelligibility, or experimentation. Fidelity controls should be voluntary, revisable, and complemented by mobility modes. Institutions should also reconsider penalties that make one register compulsory. Individual choice occurs within structural conditions, so both user control and gate reform remain relevant.

11.6 Provenance Surveillance

This subsection examines the risk that history recording becomes a tool for monitoring workers, speakers, patients, students, or political participants.

The risk is substantial. Provenance should be purpose-bound, minimized, tiered, access-controlled, retention-limited, and contestable. Cryptographic commitments can sometimes demonstrate integrity without public disclosure of sensitive content. Some settings justify local-only history or a simple endorsement record. A governance design that maximizes trace collection violates the plural provenance profile when privacy, authority, or safety deteriorates.

11.7 Longitudinal Extrapolation

This subsection tests the move from short experiments to language change.

Immediate suggestion effects establish a mechanism under a selected task. Durable linguistic change requires repeated exposure, persistence, transfer, social diffusion, institutional reinforcement, and a suitable comparison. The paper states these as research obligations. Its recurrence supplies candidate edges and conditional dynamics, while empirical direction remains open.

11.8 Recursive-Training Analogy

This subsection separates model collapse from human accommodation and institutional normalization.

Recursive model training involves learned distributions, data mixture, approximation, filtering, initialization, and generational updates. Human language involves agency, interpretation, resistance, interaction, identity, institutions, and innovation. The processes can share an abstract feedback form while differing in mechanism. Evidence from recursive training supports the $D_t \rightarrow \psi_{t+1}$ edge. Human and institutional edges require their own studies.

11.9 Community Romanticization

This subsection addresses internal hierarchy and conflict within community governance.

Community authority remains plural and fallible. A legitimate process should represent plurality, record dissent, disclose conflicts, protect individual rights, and provide review. External expertise

and criticism can contribute evidence. The key requirement is accountable relationship with affected people and legitimate control over collective resources, joined to corrigibility.

11.10 Formal Simplification

This subsection evaluates the finite categories, linear mixtures, and local recurrence.

The finite model omits open-ended expression, context dependence, semantic topology, network diffusion, delayed feedback, strategic behavior, and changing category systems. Its value lies in exact counterexamples and measurement separation. The recurrent model adds changing state but remains local and schematic. Empirical use requires richer models and comparison against simpler baselines. Failure to connect the constructions to measurable distinctions would reduce the formal contribution to illustration.

12 Future Research

This section states open philosophical, formal, empirical, technical, linguistic, and governance obligations. Each problem includes a result target and a revision condition.

Open Problem 12.1 (Participation-sensitive evaluation). *Develop intervention tasks that distinguish trace-based artifact performance, multimodal grounding, interactive practical competence, and relation-specific participation while controlling model scale, data access, and evaluator bias.*

The result should predict transfer or breakdown behavior unavailable from ordinary benchmarks. If richer task data explain the same outcomes with equal adequacy, the participation coordinate should be narrowed.

Open Problem 12.2 (Open expression-space dynamics). *Extend the finite distribution model to changing vocabularies, compositional structures, contextual meaning, innovation, and category birth while retaining identifiable individual and population quantities.*

Candidate tools include marked point processes, evolving semantic graphs, distribution-valued dynamics, and qualitative-to-formal category revision. The model should retain metric and ontology sensitivity.

Open Problem 12.3 (Delayed heterogeneous recurrence). *Estimate recurrent human, social, institutional, system, and data-pool edges with heterogeneous delays, discontinuous releases, strategic response, network interference, and exit.*

The result should discriminate stable plural regimes, lock-in, oscillation, abrupt policy change, and recovery. Local contraction serves as a baseline alongside richer dynamic theories.

Open Problem 12.4 (Relational-historical fidelity). *Develop validated measures for stance, uncertainty, role, temporal history, register, cultural indexicality, disagreement, and addressed audience across languages and communities.*

Evaluation should compare automatic, expert, participant, and audience measures. Persistent disagreement should remain explicit alongside any average.

Open Problem 12.5 (Provenance utility and privacy). *Compare provenance tiers for interpretation, correction, trust calibration, attribution, privacy, security, accessibility, and institutional burden.*

The research should identify which recorded events materially improve decisions and which primarily increase exposure. Technical validation, social comprehension, and institutional response require separate outcomes.

Open Problem 12.6 (Community-controlled language infrastructure). *Study governance arrangements that combine model performance, local authority, dialect plurality, individual mobility, restricted knowledge, capacity transfer, provider exit, and long-term maintenance.*

Comparative cases should be led or co-governed by affected communities and should publish process failures alongside performance results where safe and authorized.

Open Problem 12.7 (Interpretive-standing thresholds). *Specify standing requirements for brainstorming, translation, public communication, research synthesis, testimony, administrative evidence, and adjudication.*

The analysis should distinguish epistemic contribution, representative mandate, procedural authorization, accountability, and legal effect. Domain-specific work is necessary.

Open Problem 12.8 (Institutional style feedback). *Estimate how evaluation rubrics, hiring and promotion practices, editorial policies, administrative templates, and automated classifiers amplify or resist AI-mediated language patterns.*

This problem connects interface design to political economy. Institution-level variation can reveal feedback channels unavailable from user studies alone.

Open Problem 12.9 (Systematic originality review). *Compare the layered and recurrent framework with philosophy of AI, linguistic anthropology, accommodation theory, computer-mediated communication, human–AI co-creation, algorithmic culture, language ideology, participatory design, data governance, and provenance scholarship.*

The framework should be compressed, renamed, or abandoned wherever established theories already perform the same work with greater precision.

13 Conclusion

This section consolidates the paper’s limited position and its implications for research and governance. It returns to generative AI’s dual location as learner from traces and mediator of later expression.

Generative AI can transform traces into useful artifacts and can become part of the relations through which future symbols are produced. Evaluation at this junction requires several objects. Artifact performance concerns the task result. Participation concerns the source relation. Successor competence concerns practical transfer and regeneration. Interpretive standing concerns warranted roles. Distribution concerns population expression. Accessibility concerns feasible competent use by differently situated people. Fidelity concerns semantic and relational-historical transformation. Provenance concerns the history and conditions of the artifact. Institutional authority concerns

consequential use. Future generativity concerns later human, community, institutional, and model conditions.

The paper’s exact results are modest and useful. A finite intervention can expand one person’s effective repertoire while concentrating population output. A final artifact can arise from distinct human and AI histories, leaving output-only provenance underidentified. The dynamic result is conditional: recurrent mediation converges locally under an explicit contraction regime and becomes unstable around a fixed point with an expanding direction. These results reject simple directional or historical inferences inside the stated models while preserving empirical questions.

The practical consequence is a research architecture. Studies should observe suggestion exposure, human decisions, later unaided production, group distributions, institutional gates, model versions, and provenance under ethical constraints. Short experiments, longitudinal panels, participatory inquiry, institutional process tracing, and data-pool audits answer different parts. Their results should remain layered.

The governance consequence is modular. User-controlled fidelity and mobility modes, plural suggestions, selective provenance, privacy, community authority, institutional source linkage, portability, appeal, and recurrent monitoring target different mechanisms. Their combination should reflect stakes, affected groups, and threat models. Historical transparency and legitimate authority require evidence beyond output quality. A responsible system makes its role, uncertainty, history, limits, and routes of correction proportionate to the use.

The framework remains revisable. Its value will depend on whether it improves discrimination, measurement, and institutional design across real languages, users, communities, and systems. Future evidence may merge, divide, or remove its coordinates. That openness is part of the proposal: generative AI should be studied through the changing relations it enters and changes, with confidence matched to the evidence available at each layer.

Acknowledgments

This manuscript developed through dialogue with ChatGPT (OpenAI), which assisted with literature discovery, structural review, formal reconstruction, drafting, and criticism. The author bears responsibility for the definitions, formal constructions, source selection, interpretations, arguments, and remaining errors. The paper also acknowledges the researchers, communities, and practitioners whose verified work supplies its philosophical, empirical, linguistic, participatory, documentation, and provenance antecedents.

References

- [1] Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. AI suggestions homogenize writing toward western styles and diminish cultural nuances. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–21. Association for Computing Machinery, 2025.
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, et al. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

- [3] Kenneth C. Arnold, Krysta Chauncey, and Krzysztof Z. Gajos. Predictive text encourages predictable writing. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 128–138. Association for Computing Machinery, 2020.
- [4] Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018.
- [5] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623. Association for Computing Machinery, 2021.
- [6] Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198. Association for Computational Linguistics, 2020.
- [7] Steven Bird. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519. International Committee on Computational Linguistics, 2020.
- [8] Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. Systematic inequalities in language technology performance across the world’s languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 5486–5505. Association for Computational Linguistics, 2022.
- [9] Stephanie Russo Carroll, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, Rodrigo Sara, Jennifer D. Walker, Jane Anderson, and Maui Hudson. The CARE principles for indigenous data governance. *Data Science Journal*, 19:43, 2020.
- [10] Coalition for Content Provenance and Authenticity. Content credentials: C2PA technical specification, version 2.2. Official technical specification, 2025.
- [11] Harry Collins. *Tacit and Explicit Knowledge*. University of Chicago Press, 2010.
- [12] Anil R. Doshi and Oliver P. Hauser. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28):eadn5290, 2024.
- [13] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, et al. PaLM-E: An embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 8469–8488, 2023.
- [14] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.
- [15] Jeffrey T. Hancock, Mor Naaman, and Karen Levy. AI-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication*, 25(1):89–100, 2020.

- [16] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1–3):335–346, 1990.
- [17] Jess Hohenstein, Dominic DiFranzo, Rene F. Kizilcec, Zhila Aghajari, Hannah Mieczkowski, Karen Levy, Mor Naaman, Jeffrey Hancock, and Malte Jung. Artificial intelligence in communication impacts language and social relationships. *Scientific Reports*, 13:5487, 2023.
- [18] Edwin Hutchins. *Cognition in the Wild*. MIT Press, 1995.
- [19] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. Co-writing with opinionated language models affects users’ views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–22. Association for Computing Machinery, 2023.
- [20] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293. Association for Computational Linguistics, 2020.
- [21] Jean Lave and Etienne Wenger. *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press, 1991.
- [22] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229. Association for Computing Machinery, 2019.
- [23] Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, et al. Participatory research for low-resourced machine translation: A case study in african languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160. Association for Computational Linguistics, 2020.
- [24] John R. Searle. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–424, 1980.
- [25] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631:755–759, 2024.
- [26] S. Shyam Sundar. Rise of machine agency: A framework for studying the psychology of human–AI interaction. *Journal of Computer-Mediated Communication*, 25(1):74–88, 2020.