Concept Paper

From Plausible Narrative to Sound Abduction: A Governed Abductive Architecture for Medical Digital Twins in Multiple Sclerosis Care

Emanuel Shirbint * and Alexander Rybalov

Posted Date: 12 May 2026

doi: 10.20944/preprints202605.0728.v1

Keywords: patient digital twins; abductive reasoning; governed abductive architecture; semiosphere; clinical decision support; multiple sclerosis

 doi:10.20944/preprints202605.0728.v1

Concept Paper

From Plausible Narrative to Sound Abduction: A Governed Abductive Architecture for Medical Digital Twins in Multiple Sclerosis Care

Emanuel Shirbint [1,*] and Alexander Rybalov [2]

[1] Independent Researcher, Israel

[2] LAMBDA—Laboratory for AI, Machine Learning, Business & Data Analytics, Tel Aviv University

* Correspondence: shirbintemanuel@gmail.com

Abstract

Large language models (LLMs) embedded in medical Patient Digital Twins (PDTs) exhibit a systemic vulnerability: they can generate fluent, narratively persuasive, yet abductively unsound clinical explanations. In this context, abductive soundness means that an explanation preserves mechanistic plausibility, temporal coherence, explicit handling of missing premises, and sensitivity to counter-evidence. This article reframes the problem as architectural rather than as a mere deficit in training data. We identify three recurrent modes of abductive failure — missing-premise neglect, weak-mechanism support, and counter-evidence discounting. They arise when local semantics, formal world ontology, and the role-specific clinical semiosphere are collapsed into a single surface flow of generation. We propose a governed abductive architecture organised around seven runtime contours and operationalise it in the MS-AGIP platform for multiple sclerosis care. The architecture separates three subsystems: an ontology-guided Research Framework, a clinician-facing Neurologist Digital Twin, and a patient-controlled Patient Digital Twin. We show how disease-specific causal templates, evidence-tiered biomarker reasoning, provenance labels, temporal-coherence checks, molecular-clinical discordance detection, and governed patient-feedback updates jointly transform plausible narrative into sound abduction. The article presents an architectural blueprint and validation protocol aligned with TRIPOD+AI and DECIDE-AI. The architectural-versus-scale distinction has direct implications for safe medical AI: the difference between a fluent and a sound clinical system lies more in architecture and governance than in model size. None of the subsystems has yet been clinically deployed.

Keywords: patient digital twins; abductive reasoning; governed abductive architecture; semiosphere; clinical decision support; multiple sclerosis

1. Introduction and Series Context

1.1. Position of This Paper in the MS-AGIP Series

The MS-AGIP programme consists of three domain specifications and one series-level synthesis. The research companion manuscript defines the ontology-guided hypothesis engine; the clinician-facing companion manuscript defines the Neurologist Digital Twin; the patient-facing companion manuscript defines the Patient Digital Twin. The present paper supplies the cross-cutting abductive-safety argument: why a medical digital twin should be designed as a governed abductive architecture rather than as a monolithic LLM interface. The companion manuscripts provide domain-specific implementation detail; the present paper is intended to stand alone as the series-level architecture and validation protocol.

 doi:10.20944/preprints202605.0728.v1

Table 1. MS-AGIP series structure. The present paper is the synthesis layer: it maps abductive failure modes to architectural controls across the three companion manuscripts.

Art.	Title	User/focus	Data tier	Status
SYN	From Plausible Narrative to Sound Abduction	Series-level abductive safety architecture	Architecture + validation protocol	This paper
I	Research Framework: Architecture and Hypothesis Generation	Researcher/bioinformatician	De-identified IRB cohort	Companion
II	Neurologist Digital Twin: Clinical DSS	Treating neurologist	Identified clinical + biobank	Companion
III	Patient Digital Twin: Personal Disease Narrative	Individual MS patient	Patient-provided only	Companion

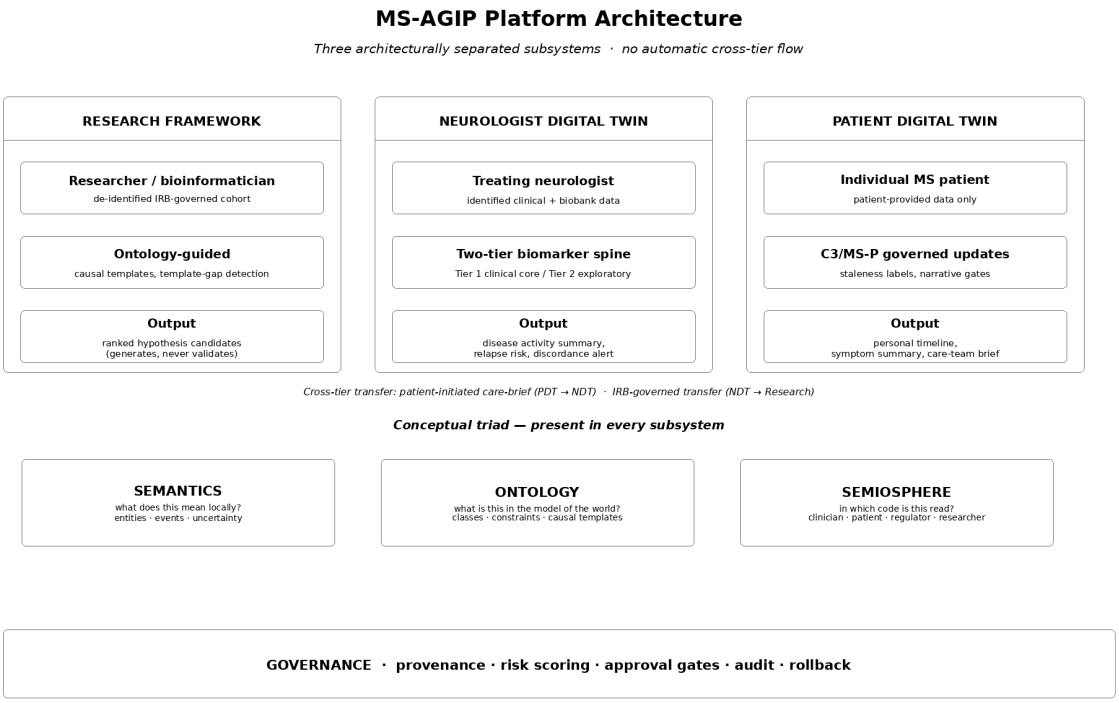

Figure 1. MS-AGIP platform architecture and abductive-safety layers. The platform separates research, clinician-facing, and patient-facing subsystems; semantics, ontology, and semiosphere are present in each subsystem; governance constrains cross-tier transfer and output authority.

1.2. What This Paper Does and Does Not Do

This paper specifies a recurrent failure model for LLM-based medical PDTs — missing-premise neglect, weak-mechanism support, and counter-evidence discounting — and defines a governed abductive architecture. It has seven runtime contours and three separated layers of understanding: semantics, ontology, and semiosphere. It operationalises the architecture in the MS-AGIP platform across research, clinician-facing, and patient-facing subsystems and proposes a validation pathway aligned with TRIPOD+AI and DECIDE-AI.

 doi:10.20944/preprints202605.0728.v1

The contribution is therefore an abductive-safety architecture and staged validation protocol, not an evaluated clinical model. The argument should not be read as claiming that all LLM outputs are clinically wrong or that foundation models cannot contribute to safe clinical systems. Nor does the paper claim to have built or clinically deployed a general-intelligence system; where the term bounded AGI appears, it is used as an engineering label for a domain-bounded, governance-constrained architecture. The paper does not report prospective clinical performance, device approval, or validated patient-outcome improvement, and it does not propose to replace neurologist judgement or patient–clinician dialogue.

1.3. The Central Problem

Industrial monitoring and clinical medicine appear to demand similar functions from artificial intelligence: integrate heterogeneous data streams, detect anomalies, propose causal hypotheses, and recommend actions. The analogy is misleading. Industrial twins usually operate under explicit physical laws, redundant sensors, and closed catalogues of failure modes. Medical twins operate on sparse vital signs, irregular laboratory measurements, patient-reported symptoms, incomplete mechanistic knowledge, and ethical limits on experimentation. Recent work on LLM-enabled and agentic digital twins underlines the urgency of this distinction for healthcare deployments (Makarov et al. 2025; Amad et al. 2025; San et al. 2026). In this setting, a fluent clinical narrative is not equivalent to a sound abductive explanation.

The central claim of this article is therefore architectural. A medical PDT requires local semantic parsing of a single patient trajectory, ontological grounding in a constrained disease model, and semiospheric translation across the codes of clinician, patient, regulator, and researcher. The semiospheric layer follows Lotman's account of meaning as a code-governed space of interpretation rather than a mere collection of signs (Lotman 2005). A monolithic LLM collapses these levels into a single generative stream. The resulting text may sound clinical while failing the abductive tests that clinical safety requires.

2. From Plausible Narrative to Abductive Failure

2.1. Abduction Is Not Prediction

Prediction asks: given these inputs, what is the most likely label? Abduction asks a stronger question: given these observations, which causal process best explains them? Such an explanation must respect known mechanisms, temporal order, missing premises, and counter-evidence. Clinical reasoning is not merely classification under uncertainty. It is the construction and rejection of causal explanations under constraints.

Used as a standalone generator, a monolithic LLM trained on clinical text is optimised to continue or summarise plausible discourse. Without external semantic, ontological, and governance layers, it does not provide a durable representation of what remains unknown. It does not have a disease ontology against which to test mechanistic sufficiency, or a contradiction detector strong enough to interrupt fluent narrative when objective data contradict it. Neurosymbolic and knowledge-graph approaches make clear why explicit grounding and consistency checks are architecturally distinct from language generation alone (DeLong et al. 2023; Galitsky and Rybalov 2026). The result is not random error; it is a structured class of abductive vulnerability.

For the purposes of this paper, abductive soundness is used in a minimal operational sense. A candidate explanation is treated as sound only if it satisfies four constraints: required premises are either verified or explicitly marked as unknown; the proposed mechanism is adequate within the disease ontology; the temporal sequence of causes, observations, and effects is coherent; and available counter-evidence can down-rank, suspend, or defeat the explanation.

 doi:10.20944/preprints202605.0728.v1

2.2. Three Recurrent Failure Modes

Table 2. Three recurrent modes of abductive failure in LLM-based medical PDTs.

Failure mode	Clinical pattern	Architectural absence	Potential harm
Missing-premise neglect	A recommendation is generated while a relevant contraindication, cautionary factor, or premise is absent from the prompt or buried in narrative notes.	No explicit registry of required preconditions, contraindications, and unknown fields.	Unsafe recommendation that appears grammatically and clinically fluent.
Weak-mechanism support	Textual co-occurrence between symptoms and diagnoses substitutes for pathophysiological reasoning.	No disease ontology, causal templates, or temporal-coherence checks.	A common narrative overrides a mechanistically weak explanation.
Counter-evidence discounting	The model continues a plausible diagnosis even when objective tests contradict it.	No discordance detector or belief-update mechanism strong enough to stop the narrative.	Delay, unnecessary testing, or misdirected clinical attention.

2.3. Illustrative Clinical Scenarios

Missing-premise neglect can occur when a patient with type 2 diabetes and elevated HbA1c receives an LLM-generated suggestion to add an SGLT2 inhibitor. At the same time a recurrent urinary-tract-infection history is present only in unstructured notes and should be surfaced as a risk-modifying premise. Weak-mechanism support can occur when fatigue, arthralgia, and mild anaemia are narratively associated with systemic lupus erythematosus despite negative ANA and normal complement. Counter-evidence discounting can occur when chest pain is narratively pulled toward coronary disease despite a normal electrocardiogram, normal troponin, and normal stress testing.

A multiple-sclerosis-specific example is more central to MS-AGIP. A patient reports worsening gait and fatigue after a febrile infection. A fluent LLM-PDT may narrate this as possible relapse. Yet MRI shows no new lesions, serum NfL is not elevated, and the time course is compatible with pseudo-exacerbation rather than new inflammatory activity. MRI and serum NfL are established components of contemporary MS monitoring, although their interpretation remains clinically contextual (Filippi et al. 2016; Disanto et al. 2017; Barro et al. 2018). Sound abduction requires the system to discount the relapse narrative and foreground the discordance between subjective deterioration and objective injury markers. MS-AGIP makes this operation explicit through a molecular-clinical discordance detector: a rule-governed alerting layer that compares patient-reported or clinician-observed deterioration with Tier 1 objective activity markers, flags unresolved discordance for clinician review, and does not diagnose relapse or recommend therapy.

2.4. Why Medical PDTs Are Especially Exposed

Medical PDTs amplify these failures for five reasons. First, they are personalised to a single patient, so atypical trajectories are easily pulled toward common narratives. Second, they are longitudinal, while LLMs remain weak at modelling clinically meaningful temporal order. Third, they are multimodal: laboratory, imaging, wearable, and patient-reported data must cohere across

 doi:10.20944/preprints202605.0728.v1

modalities. Fourth, they must operate under uncertainty rather than suppress it. Fifth, fluent prose creates an asymmetry of trust: clinicians and patients may read smoothness as competence, even when the underlying explanation is unsound (Makarov et al. 2025; Amad et al. 2025).

3. Seven-Runtime Architecture

3.1. From Abstract AGI to a Bounded Operational Contour

In this article, bounded AGI does not denote universal intelligence, autonomous clinical authority, or a metaphysical claim about machine consciousness. It denotes a system of broad competence inside an explicitly delimited operational contour. This use of "bounded" follows a capability-oriented framing in which performance, generality, and autonomy must be specified rather than presumed (Morris et al. 2023). The contour specifies available observations, admissible actions, protected invariants, risk thresholds, and decisions requiring human approval. Restriction is therefore not a weakness; it is the precondition of measurement, safety, and incremental extension of capability.

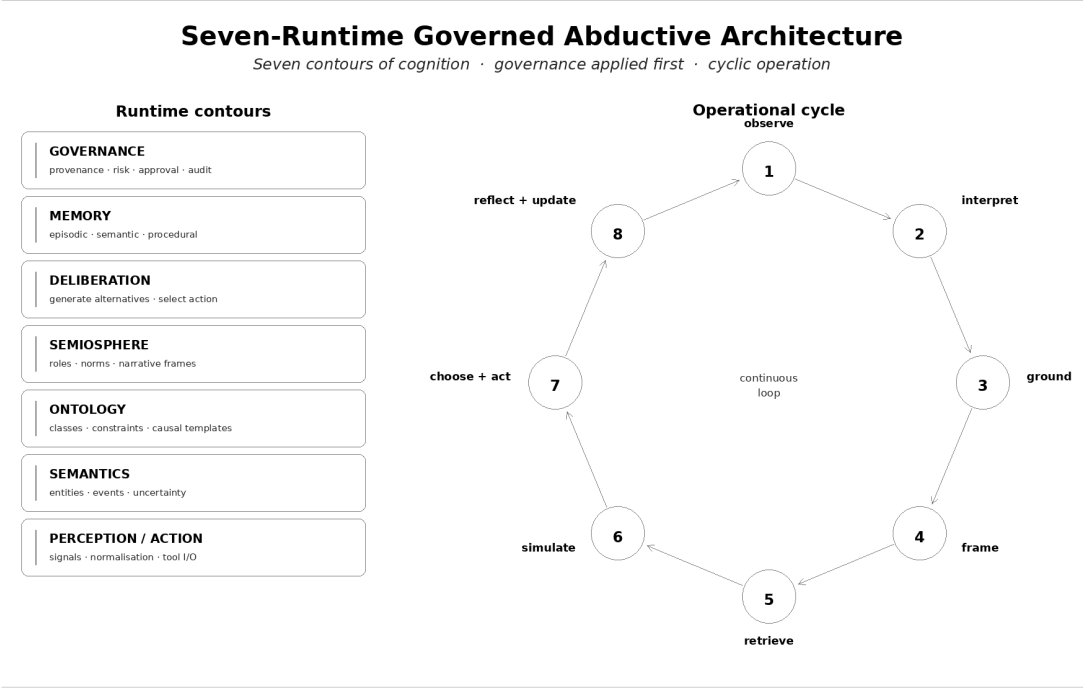

Figure 2. Seven-runtime architecture and operational cycle. Governance is applied first; perception/action is the substrate. The operating loop is observe → interpret → ground → frame → retrieve → simulate → choose + act → reflect + update.

3.2. Runtime Contours

The architecture is composed of seven runtime contours. The perception/action substrate normalises inputs and records tool outcomes. The semantic runtime extracts typed meaning objects. The ontology runtime maps them into classes, constraints, temporal relations, and causal templates. The semiosphere runtime stores roles, norms, narrative frames, and translation maps. The deliberation runtime generates competing hypotheses and action branches. The memory runtime preserves episodic, semantic, procedural, social, and normative experience, a requirement reinforced by work on generative agents and autonomous-agent memory (Park et al. 2023; Du 2026). The governance runtime enforces provenance, risk scoring, approval gates, audit, and rollback.

 doi:10.20944/preprints202605.0728.v1

3.3. The Conceptual Triad

The central conceptual contribution is the separation of three layers that monolithic LLM systems usually collapse. The semantic layer asks what a fragment means locally. The ontological layer asks what it is in the formal model of the world, a distinction central to neurosymbolic and knowledge-graph reasoning (DeLong et al. 2023). The semiospheric layer asks in which institutional, clinical, regulatory, or patient narrative code it is read (Lotman 2005).

Table 3. Distinguishing the semantic, ontological, and semiospheric levels.

Layer	Guiding question	Primary objects	Risk if absent
Semantics	What does this mean locally?	Entities, events, relations, themes, intentions, uncertainty, contradictions.	Smooth prose without durable meaning objects.
Ontology	What is this in the model of the world?	Classes, properties, constraints, causal templates, temporal relations.	An unstable world model and logically inconsistent explanations.
Semiosphere	In which code is this read?	Roles, norms, institutions, narrative frames, translation maps.	Formal correctness coupled with social blindness.

In MS care, the same fact — for example, a serum NfL z-score 1.6 standard deviations above baseline — carries multiple readings. For the neurologist it may trigger interval MRI review and for the regulator it raises threshold-calibration and provenance questions. For the patient it should become a prompt for dialogue with the care team, not a reason to self-modify therapy; for the researcher it may become a hypothesis candidate (Lotman 2005). In this architecture, that translation function is treated as a safety layer.

4. Operationalisation in the MS-AGIP Platform

4.1. Three Subsystems, Three Governance Regimes

MS-AGIP instantiates the architecture in a concrete clinical domain: multiple sclerosis. It separates the platform into three subsystems with different users, data tiers, outputs, and governance regimes (see Figure 1). The Research Framework serves the researcher or bioinformatician; it operates exclusively on a de-identified IRB-governed cohort and produces ranked cross-modal hypothesis candidates with experimental designs. Its governance principle is that hypotheses are generated, never validated automatically, and a principal-investigator review gate is required before any candidate is considered actionable for downstream study.

The Neurologist Digital Twin serves the treating MS neurologist; it operates on identified clinical and biobank data under separate consent and produces a disease activity summary, a relapse-risk signal, and molecular-clinical discordance alerts. Its governance principle is to support evidence synthesis without diagnosis, prescribing, or autonomous clinical decision.

The Patient Digital Twin serves the individual MS patient; it operates only on patient-provided data — patient-reported outcomes, diary entries, wearable streams, and optional shared documents. It produces a personal timeline, symptom-pattern summary, care-team brief, and disease narrative profile. Its governance principle is that the patient controls collection, narrative updates, sharing, pause, deletion, and retirement.

No automatic cross-tier flow exists. The Research Framework does not access identified clinical records; the Neurologist Digital Twin does not send data to the Patient Digital Twin without explicit

 doi:10.20944/preprints202605.0728.v1

patient action; the Patient Digital Twin operates only on patient-provided data. This separation is the operational expression of the semiosphere: the same patient-related fact is read differently by a researcher, a treating neurologist, and the patient (Lotman 2005).

4.2. Research Framework: Ontology-Guided Hypothesis Generation

The Research Framework uses the semantic layer to type molecular and clinical events, including the ExpressionEvent object with relapse-cycle phase tags. Its ontology encodes MS causal templates as directed relations with preconditions and constraints. Its semiospheric layer contains research frames — immunogenetic, transcriptomic, microbiome, imaging, clinical, and therapeutic — each with explicit epistemic boundaries. The design is described in a companion manuscript on the MS-AGIP Research Framework (companion manuscript, under review) and draws on graph-based retrieval principles for connecting local observations to global sensemaking (Edge et al. 2024). The system generates structured hypothesis candidates, not confirmed findings.

4.3. Neurologist Digital Twin: Evidence Synthesis with Provenance

The NDT is a clinician-facing evidence synthesis tool. Its clinical core uses evidence-mature biomarkers: MRI, serum NfL, GFAP, EDSS slope, annualised relapse rate, and therapy context, while diagnosis remains governed by established clinical criteria rather than by the NDT itself (Montalban et al. 2025; Filippi et al. 2016; Disanto et al. 2017; Barro et al. 2018). Its exploratory module contains research-grade IFN score and microbiome annotations for biobank-enrolled patients only; the microbiome components remain observational and exploratory in line with the current evidence base (Berer et al. 2017; Rothhammer et al. 2016). The relapse-risk component is specified in a companion manuscript (companion manuscript, under review) as a Cox-LASSO design on Tier 1 inputs only (Tibshirani 1997). Every output carries provenance: data source, model version, calibration tier, confidence, and vintage, with reporting and evaluation aligned to TRIPOD+AI and DECIDE-AI (Collins et al. 2024; Vasey et al. 2022). Pharmacogenetic HLA × drug-response annotations are excluded pending independent replication in an external MS cohort. The NDT does not diagnose, prescribe, initiate, modify, or discontinue therapy; its regulatory status would require device-specific analysis under EU MDR and FDA CDS guidance (European Parliament and Council 2017; FDA 2026).

Operationally, the discordance layer treats patient-reported or clinician-observed change, MRI activity, serum NfL/GFAP dynamics, EDSS slope, relapse history, and therapy context as separately provenance-labelled signals. It can therefore emit review flags for possible under-estimated activity, possible over-estimated activity, or unresolved biomarker-clinical discordance. These flags are structured prompts for neurologist review, not diagnoses or treatment recommendations.

4.4. Patient Digital Twin: Frozen Clinical Semiosphere and Governed Updates

The PDT addresses the patient side of the disease semiosphere (Lotman 2005). A living patient continuously revises the meanings of relapse, remission, disability, agency, fatigue, and adaptation. A digital model necessarily captures a temporal snapshot of those codes: the frozen clinical semiosphere. MS-AGIP responds through explicit staleness labels and the Governed Patient-Feedback Update Contour (C3/MS-P), which updates narrative frames only through patient-data consistency, temporal coherence, and explicit patient confirmation for structural narrative changes. The patient-facing design also includes psychological safeguards, reflecting the recognised burden of depression and anxiety in MS (Siegert and Abernethy 2005).

 doi:10.20944/preprints202605.0728.v1

5. How the Architecture Constrains Abductive Failure

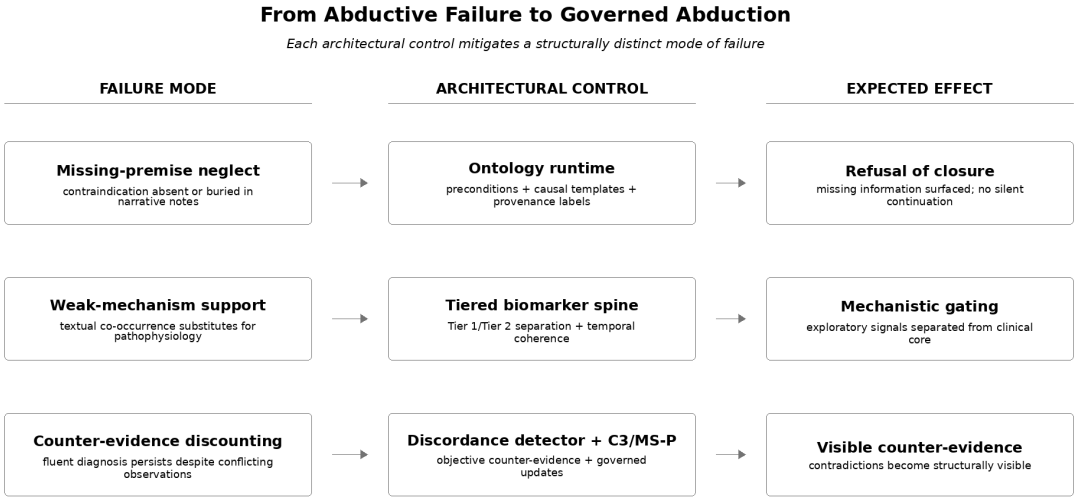

Figure 3. From abductive failure to governed abduction. Each failure mode is mitigated by a corresponding architectural control rather than by a larger language model alone.

The mapping from failure mode to architectural control is as follows. Missing-premise neglect has its root cause in the fact that the model has no explicit representation of required but absent premises. MS-AGIP addresses this through an ontology runtime with preconditions, contraindications, causal templates, and "unverified precondition" states, together with provenance labels on every clinical output. The expected effect is that the system can withhold abductive closure and surface missing information rather than silently continuing a recommendation.

Weak-mechanism support arises when textual co-occurrence is mistaken for pathophysiological explanation. MS-AGIP constrains this through a two-tier biomarker architecture, ExpressionEvent temporal annotation, evidence maturity labels, and cross-frame interpretability. The expected effect is that exploratory signals are separated from clinical-core signals and hypotheses must pass mechanistic and temporal coherence checks.

Counter-evidence discounting occurs when no contradiction detector interrupts the fluent narrative. MS-AGIP addresses this through the molecular-clinical discordance detector, over-estimated activity alert, and C3/MS-P consistency and temporal-coherence gates. The expected effect is that objective counter-evidence becomes structurally visible and can override the narrative trajectory.

The important point is that each mitigation mechanism is architectural. The system is constrained so that certain errors are designed not to pass silently: missing preconditions are labelled, exploratory biomarkers are quarantined from clinical-core reasoning, and discordant objective data trigger structured alerts.

6. Evidence Status and Validation Programme

6.1. Current Evidence Status

The claims advanced in this paper differ in their epistemic standing, and we separate them explicitly. The three-mode abductive failure model is a conceptual proposal supported by clinical examples and the literature on LLM and digital-twin risk; its required next step is architecture-level comparison against published LLM-PDT approaches. The seven-runtime governed abductive

 doi:10.20944/preprints202605.0728.v1

architecture is a design specification at blueprint level, awaiting ablation-style demonstration in a bounded prototype.

Three components have prospective validation protocols already specified. The Research Framework is a design-and-protocol contribution that has not yet been prospectively validated; its required next step is Milestone 0 — a retrospective proof-of-concept with blinded expert-panel assessment against a PubMed-RAG baseline. The Neurologist Digital Twin's relapse-risk signal is a Cox-LASSO design on Tier 1 inputs, currently a provisional site-calibrated pilot; it requires DECIDE-AI silent deployment and MSBase transport validation. The molecular-clinical discordance detector defines three alert classes with provisional retrospective estimates and requires prospective adjudicated validation per alert class.

The Patient Digital Twin's Governed Patient-Feedback Update Contour (C3/MS-P) is currently a design-only specification; its required next step is the planned N = 30 usability study with PHQ-9 monitoring and multilingual patient-facing materials. None of these components has yet been clinically deployed, and every output described in this paper is labelled accordingly.

Section 6.2 translates these status labels into a staged evaluation programme: retrospective feasibility first, then silent clinical evaluation, visible deployment, and external transport validation, in keeping with TRIPOD+AI and DECIDE-AI expectations for prediction models and early-stage clinical decision support (Collins et al. 2024; Vasey et al. 2022).

6.2. Validation Roadmap

Table 4. Series-level validation roadmap.

Stage	Subsystem	Study/milestone	Primary endpoint	Timing
Milestone 0	Research Framework	Retrospective proof-of-concept against PubMed-RAG baseline with blinded MS expert panel.	Novel and testable hypothesis candidates; ontological consistency; time to hypothesis.	2026 Q2 – Q3
Silent deployment	NDT	Background NDT run with outputs hidden from neurologists.	Concordance between signals and subsequent clinical action.	2026 Q3 – 2027 Q2
Visible deployment	NDT	Randomised crossover study with neurologists using NDT vs standard synthesis.	Clinical data synthesis time, confidence, MRI ordering, alert fatigue.	2027 Q1 – Q4
Transport validation	NDT	External validation of Cox-LASSO in MSBase Israeli cohort.	Calibration slope/intercept, Brier score, C-statistic.	2027
Usability pilot	PDT	N = 30 RRMS patients; mixed-methods usability and psychological safety monitoring.	PRO completion, SUS score, comprehension, PHQ-9 safety.	2027 Q1 – Q3

 doi:10.20944/preprints202605.0728.v1

Stage	Subsystem	Study/milestone	Primary endpoint	Timing
Integrated evaluation	All	Concurrent operation of Research, NDT, and PDT under governance.	Data-flow integrity, governance compliance, combined research/clinical/patient impact.	2028

The roadmap operationalises an honest sequence rather than a single proof event. Retrospective feasibility precedes silent clinical deployment; silent deployment precedes visible evaluation; visible evaluation precedes external transport validation; integrated platform evaluation comes last. Each stage produces evidence that conditions the next; none stands alone, and any negative result triggers architectural revision before further progression. This is the standard logic of TRIPOD+AI and DECIDE-AI applied at series scale (Collins et al. 2024; Vasey et al. 2022).

7. Societal and Governance Dimension

The technical failure described in this article also has a societal dimension. First, fluency creates an asymmetry of trust. A model presented by an institution and embedded into a clinical interface receives a presumption of competence that its internal reasoning may not deserve. When a fluent output is wrong, the cost of overriding it is borne by clinicians as cognitive and institutional friction, and by patients as delayed or misdirected care.

Second, a single fact must be translated across several codes. An elevated serum NfL carries multiple readings: it is a clinical monitoring signal, a regulatory calibration question, a patient-facing conversation prompt, and a research hypothesis candidate (Lotman 2005). A system without a semiosphere defaults to one register and becomes socially blind in all others.

Third, abductive failure has distributive consequences. Patients with rare, atypical, or underrepresented presentations are precisely those for whom additional reasoning capacity should help. A system optimised for textual likelihood tends to pull them into common narratives. The architectural response is therefore not merely a technical refinement; it is part of the equitable distribution of clinical reasoning effort across the patient population a system serves.

8. Irreducible Limits

The architecture presented here is a blueprint and protocol, not a clinically deployed system; this is stated throughout, and validation remains a prospective programme goal rather than a claim already achieved. Several further limits deserve explicit acknowledgement.

Patient-reported outcomes, wearable streams, electronic-health-record fields, and biobank measurements all carry missingness, bias, and calibration problems. The architecture responds with provenance and vintage labels, confidence indicators, multiple imputation by chained equations where appropriate, and prospective adjudication, but data quality remains a structural rather than a technical limitation. The ontology, similarly, reflects the current knowledge boundary: causal templates not yet encoded cannot be detected as template gaps. Thus, the system is designed for periodic literature review, ontology revision, and frame update rather than for one-time specification. Several proposed analyses begin with small cohorts and retrospective pilots; all such outputs are labelled provisional, and external validation and prospective studies are required before any clinical claim.

Two limits are conceptual rather than empirical. First, the frozen clinical semiosphere is unavoidable: no digital twin can fully capture the moving lived experience of illness. Second, staleness labels and the C3/MS-P contour render that lag visible and governable rather than abolish it. Indicative regulatory analysis under EU MDR and FDA CDS guidance is not a device-specific

 doi:10.20944/preprints202605.0728.v1

regulatory decision; a full intended-use statement, hazard analysis, legal review, and regulatory memo are required before deployment.

9. Conclusions

Monolithic LLM-based medical digital twins can speak medicine without reliably reasoning medically. Their fluency may conceal missing premises, weak-mechanism support, and ignored counter-evidence. These are not merely prompt-engineering defects. They arise when semantics, ontology, and semiosphere are collapsed into one generative stream.

MS-AGIP offers a different design posture. Semantics extracts local meaning. Ontology imposes disease structure, constraints, and causal templates. The semiosphere translates facts across the roles of clinician, patient, researcher, and regulator. Memory, deliberation, and governance then transform the architecture from a text generator into a bounded abductive system. The task is not to make the model sound more clinical. The task is to make the system structurally capable of refusing unsound clinical narratives.

References

1. Amad H, Astorga N, van der Schaar M (2025) Continuously Updating Digital Twins using Large Language Models. Proceedings of Machine Learning Research 267:1343–1366
2. Barro C, Benkert P, Disanto G et al (2018) Serum neurofilament as a predictor of disease worsening and brain and spinal cord atrophy in multiple sclerosis. Brain 141(8):2382–2391. https://doi.org/10.1093/brain/awy154
3. Berer K, Gerdes LA, Cekanaviciute E et al (2017) Gut microbiota from multiple sclerosis patients enables spontaneous autoimmune encephalomyelitis in mice. Proceedings of the National Academy of Sciences USA 114(40):10719–10724. https://doi.org/10.1073/pnas.1711233114
4. Collins GS, Moons KGM, Dhiman P et al (2024) TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning. BMJ 385:e078378. https://doi.org/10.1136/bmj-2023-078378
5. DeLong LN, Mir RF, Whyte M et al (2023) Neurosymbolic AI for reasoning over knowledge graphs: a survey. arXiv:2302.07200
6. Disanto G, Barro C, Benkert P et al (2017) Serum neurofilament light: a biomarker of neuronal damage in multiple sclerosis. Annals of Neurology 81(6):857–870. https://doi.org/10.1002/ana.24954
7. Du P (2026) Memory for autonomous LLM agents: mechanisms, evaluation, and emerging frontiers. arXiv:2603.07670
8. Edge D, Trinh H, Cheng N et al (2024) From local to global: a Graph RAG approach to query-focused summarization. arXiv:2404.16130
9. European Parliament and Council (2017) Regulation (EU) 2017/745 on medical devices. Official Journal of the European Union L117:1–175
10. FDA (2026) Clinical Decision Support Software: Guidance for Industry and Food and Drug Administration Staff. U.S. Food and Drug Administration, Center for Devices and Radiological Health, issued 6 January 2026 and re-issued 29 January 2026. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software
11. Filippi M, Rocca MA, Ciccarelli O et al (2016) MRI criteria for the diagnosis of multiple sclerosis: MAGNIMS consensus guidelines. Lancet Neurology 15(3):292–303. https://doi.org/10.1016/S1474-4422(15)00393-2
12. Galitsky B, Rybalov A (2026) Neuro-Symbolic Verification for Preventing LLM Hallucinations in Process Control. Processes 14(2):322. https://doi.org/10.3390/pr14020322
13. Lotman YM (2005) On the semiosphere. Sign Systems Studies 33(1):205–229
14. Makarov N, Bordukova M, Quengdaeng P et al (2025) Large language models forecast patient health trajectories enabling digital twins. npj Digital Medicine 8:588. https://doi.org/10.1038/s41746-025-02004-3
15. Montalban X, Lebrun-Frénay C, Oh J et al (2025) Diagnosis of multiple sclerosis: 2024 revisions of the McDonald criteria. Lancet Neurology 24(10):850–865. https://doi.org/10.1016/S1474-4422(25)00270-4

 doi:10.20944/preprints202605.0728.v1

16. Morris MR, Sohl-Dickstein J, Fiedel N et al (2023) Levels of AGI for operationalizing progress on the path to AGI. arXiv:2311.02462
17. Park JS, O'Brien JC, Cai CJ, Morris MR, Liang P, Bernstein MS (2023) Generative agents: interactive simulacra of human behavior. arXiv:2304.03442
18. Rothhammer V, Mascanfroni ID, Bunse L et al (2016) Type I interferons and microbial metabolites of tryptophan modulate astrocyte activity and CNS inflammation via the aryl hydrocarbon receptor. Nature Medicine 22(6):586–597. https://doi.org/10.1038/nm.4106
19. San O, Rasheed A, Bozdemir E, Deng J (2026) The evolution of digital twins from reactive to agentic systems. Nature Computational Science 6(1):6–10. https://doi.org/10.1038/s43588-025-00944-0
20. Siegert RJ, Abernethy DA (2005) Depression in multiple sclerosis: a review. Journal of Neurology, Neurosurgery & Psychiatry 76(4):469–475. https://doi.org/10.1136/jnnp.2004.054635
21. Tibshirani R (1997) The lasso method for variable selection in the Cox model. Statistics in Medicine 16(4):385–395
22. Vasey B, Nagendran M, Campbell B et al (2022) Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nature Medicine 28(5):924–933. https://doi.org/10.1038/s41591-022-01772-9