



LECTURE 2

- A) NEUROSYMBOLIC AI: INTRODUCTION (CONTD)**
- B) MEDICAL/HEALTH AI APPLICATIONS: OPPORTUNITIES, AND RISKS**

JANUARY 7, 2026



HOW DO WE BUILD NEUROSYMBOLIC SYSTEMS ?



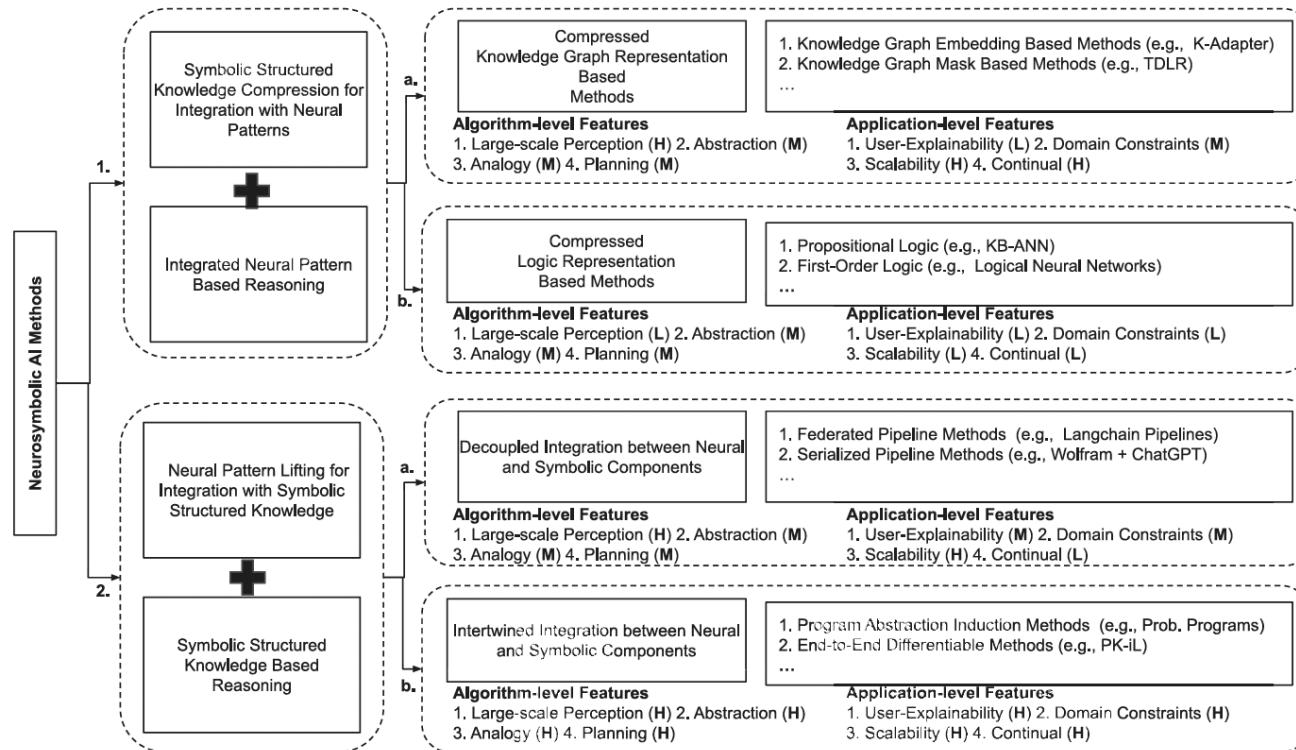
THE NEUROSYMBOLIC PROPOSITION

- Neurosymbolic AI starts from a practical view of intelligence
 - Intelligence requires both perception (processing raw data) and cognition (using background knowledge to reason, plan, and justify decisions).
 - Symbolic structures make background knowledge explicit
They represent concepts, relations, and rules directly, which supports explanation, safety constraints, and reliable evaluation of reasoning.
 - Neural models excel at pattern learning, but their knowledge is implicit
This black-box nature makes it hard to inspect what was “learned,” apply safety standards reliably, or ensure consistent reasoning when consequences matter.
- The field organizes Neurosymbolic methods into two main categories
 - 1) Compress symbolic knowledge into neural patterns (Lowering)
 - The goal is to embed knowledge so the neural system can reason using learned representations.
 - 2) Extract structure from neural patterns into symbolic forms (Lifting)
 - The goal is to map outputs into explicit knowledge structures and then do symbolic reasoning.

NEURAL+SYMBOLIC: MULTIPLE WAYS

- **Category 1 (“Lowering”):** Compress symbolic knowledge into neural patterns
 - 1A: Compressed representations of structured knowledge
 - 1B: Compressed representations of formal logic
- **Category 2 (“Lifting”):** Extract structure from neural patterns for symbolic reasoning
 - 2A: *Decoupled* neural–symbolic pipelines
 - 2B: *Intertwined* neural–symbolic integration
- Balancing SCALE and TRUSTWORTHINESS
 - An inherent tension

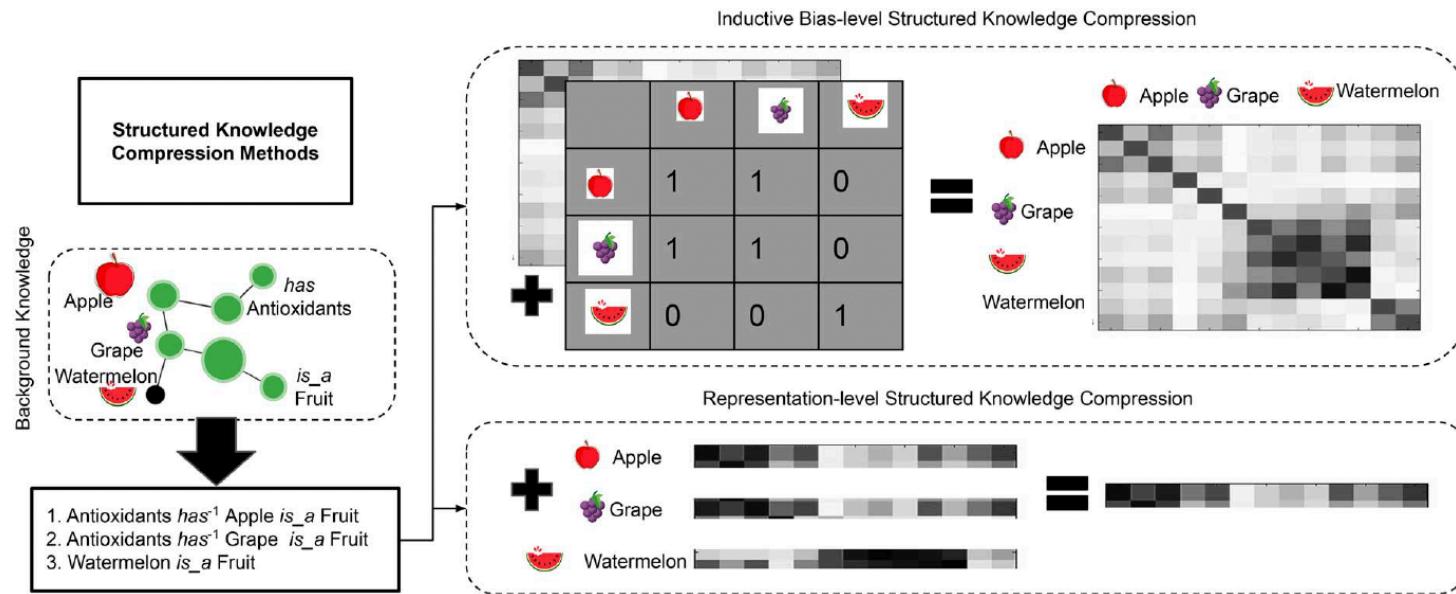
THE FOUR INTEGRATION STRATEGIES



- Just for reference today, it will take us the entire semester to fully understand and appreciate this

CATEGORY 1: LOWERING

- Category 1 (“Lowering”): Compress symbolic knowledge into neural patterns
 - 1A: Compressed representations of structured knowledge
 - 1B: Compressed representations of formal logic



1A: LOWERING, VIA COMPRESSED KNOWLEDGE GRAPH REPRESENTATIONS

- Conceptually: Structured relational knowledge is encoded into continuous representations so that reasoning is performed implicitly by neural pattern matching, rather than explicit symbolic manipulation
- Algorithmically, structured knowledge - in the form of “knowledge graphs” is transformed into “embeddings” that condition the neural representations
- This enables **high-throughput** perception and reasoning
- Advantages
 - Scale !
 - Minimal architectural disruption
- Fundamental limitations
 - Loss of explicit structure **eliminates inspectability**
 - Constraints **cannot be enforced**, only approximated
 - Reasoning behavior is **difficult to validate** or certify
- Best suited for applications where semantic coverage and scale dominate correctness guarantees, such as large-scale retrieval, recommendation, and similarity-driven tasks

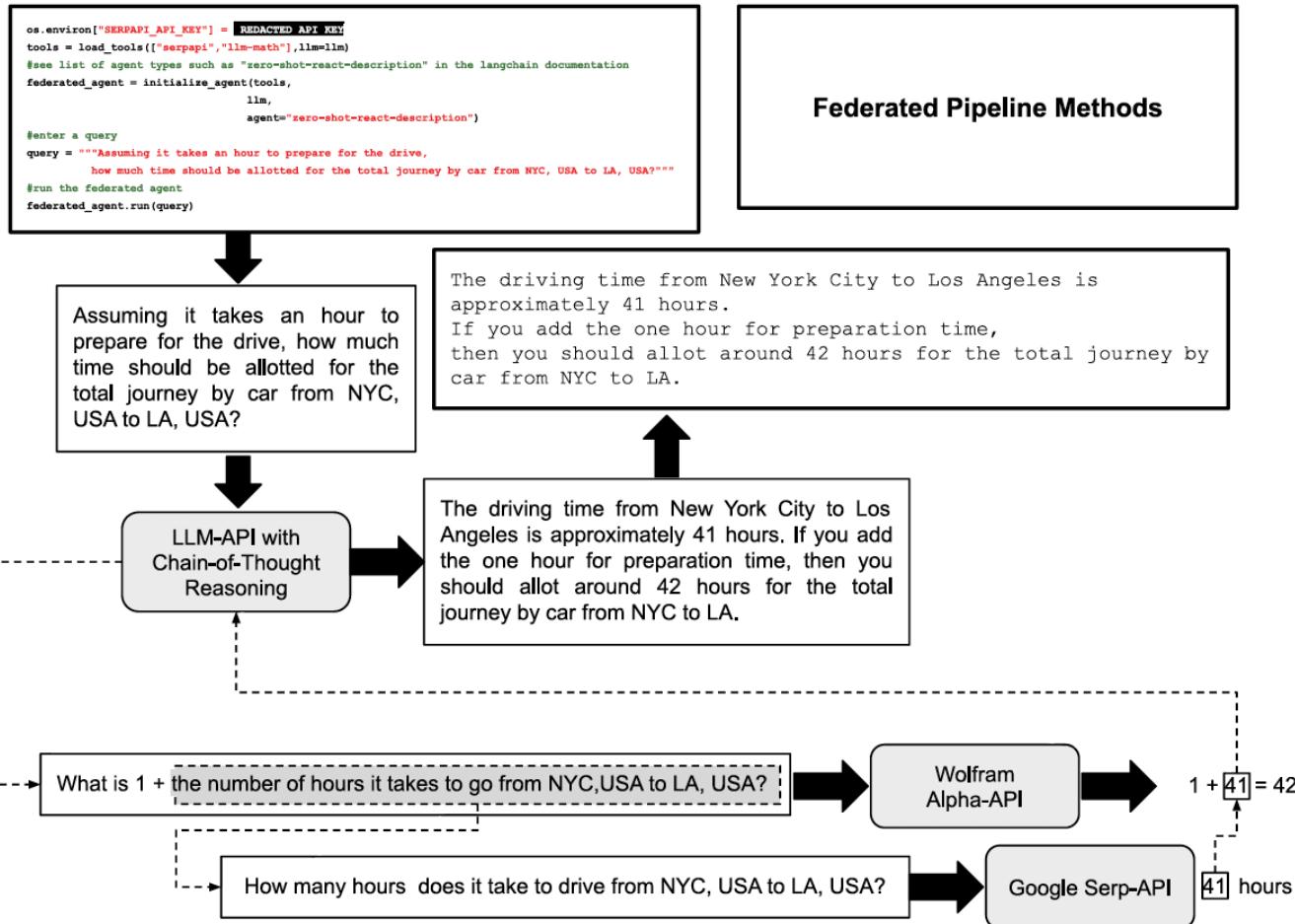
1B: LOWERING, VIA COMPRESSED FORMAL LOGIC REPRESENTATIONS

- **Conceptually:** Logical structure is **approximated** within models, allowing neural networks to emulate reasoning patterns without preserving exact logical semantics.
- Algorithmically, logic is mapped to neural constraints
- This enables
 - The integration of (weak) logical priors into learning
 - Tolerance to noise,incomplete information
- Advantages
 - More structured behavior than purely neural models
 - Supports abstraction beyond memorized rules
- Fundamental limitations
 - Logical guarantees are **no longer strict**
 - Rule **violations are possible** and difficult to bound
 - Interpretability remains **limited**
 - Certification in high-consequence settings is problematic
- Appropriate for scientific exploration, hypothesis generation, and pattern discovery, where approximate reasoning is acceptable and errors are recoverable

2A: LIFTING, VIA DECOUPLED NEURAL-SYMBOLIC PIPELINES

- **Conceptually:** Neural models and symbolic systems retain distinct roles, with explicit interfaces governing information flow and responsibility
- Algorithmically
 - Neural components handle perception, interpretation, and intent recognition
 - Symbolic components perform reasoning, constraint checking, and validation
 - Control logic orchestrates interactions between components
- This enables
 - Explicit reasoning over structured knowledge
 - Traceable decision pathways
 - Clear separation between interpretation and judgment
- Advantages
 - Strong explainability and auditability
 - Explicit enforcement of domain constraints
 - Modular design and incremental development
 - Natural alignment with human-in-the-loop workflows
- Fundamental limitations
 - Increased system complexity and orchestration burden
 - Latency from multi-stage pipelines
 - Requires well-curated symbolic representations
- Well suited for decision-support systems in regulated or safety-critical domains, including healthcare triage, emergency response, compliance, and operational planning
- Most importantly: this is the approach we will employ in our final project !

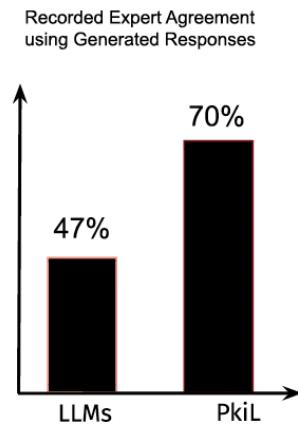
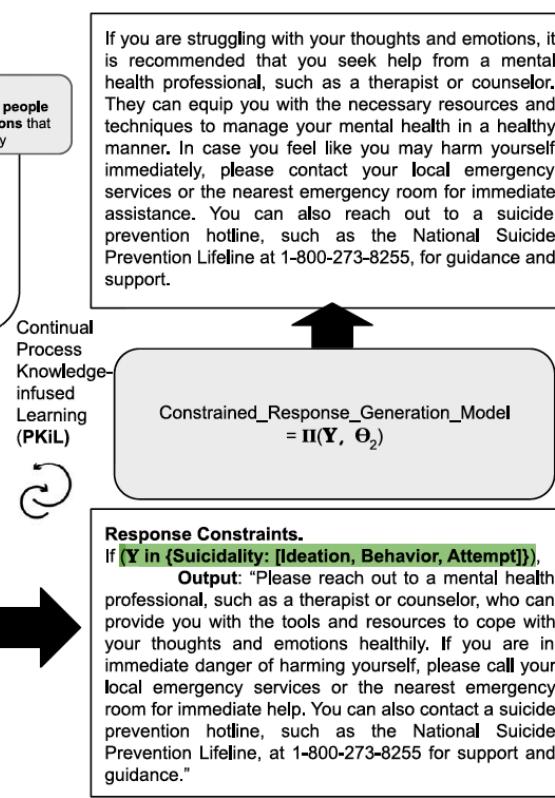
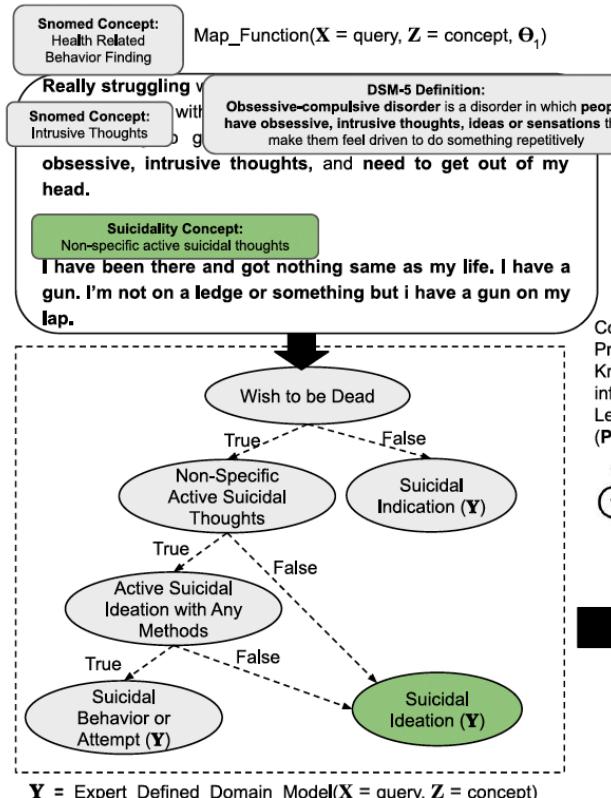
2A: EXAMPLE



2B: LIFTING, VIA INTERTWINED NEURAL-SYMBOLIC INTEGRATION

- **Conceptually:** Symbolic constraints and neural representations are co-resident in the reasoning process
- **Algorithmically**
 - Neural perception and symbolic reasoning interact bidirectionally
 - Constraints influence generation and planning directly
 - Learning and reasoning are tightly coupled
- This *can* enable
 - Consistent multi-step reasoning
 - Explicit abstraction, analogy, and planning
 - Enforcement of invariants during inference
- **Advantages**
 - Highest degree of controllability and explainability
 - Strong alignment with human cognitive processes
- **Fundamental limitations**
 - Significant engineering and modeling complexity
 - Scalability challenges
 - Requires careful knowledge engineering
- **Please note**
 - We are not there quite yet ! Aka this is an active current research area

2B: EXAMPLE



User-explainability
i.e., Clinicians and Patients
+
Domain Constraints
Verify adherence to the clinical guideline on diagnosis which a clinician understands.

IN SUMMARY

- Modern AI systems increasingly participate in decision-making, not just information generation, which changes what “correctness” and “failure” mean
- Large language models are powerful at interpreting inputs and generating fluent responses, but they lack explicit representations of background knowledge, rules, and constraints
- This gap becomes critical in high-consequence domains—such as healthcare, emergency response, and national security - where some actions or responses must *never* occur
- Neurosymbolic AI addresses this gap by deliberately separating:
 - perception and interpretation (handled well by neural models), from
 - reasoning, constraints, and justification (handled by symbolic structures)
- We examined two broad ways neural and symbolic systems interact:
 - compressing symbolic knowledge into neural representations, and
 - extracting structure from neural models to support explicit reasoning
- These lead to four integration strategies, each with different trade-offs in scalability, explainability, enforceability of constraints, and suitability for real-world deployment
- The core question going forward is not *whether* to use neural or symbolic methods, but how responsibility is allocated between them when consequences matter !



AI IN HEALTHCARE



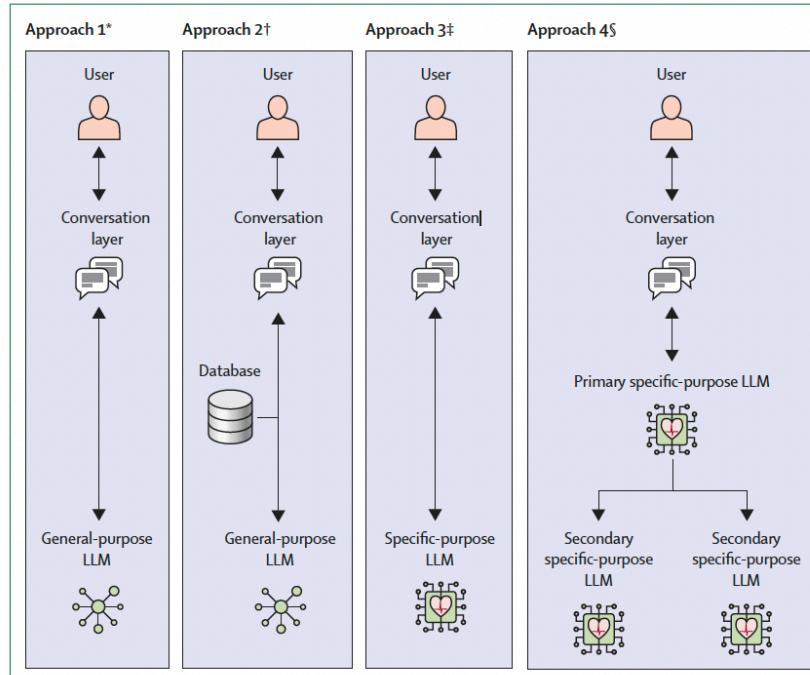
HEALTH CARE AI IS ALREADY HERE (CONCRETE SYSTEMS)

- LLMs are already embedded in real health-facing systems
- Patient-facing conversational agents built on GPT are used for symptom checking and health advice.
- WHO's SARAH chatbot uses a GPT-4–class model for public health guidance.
- Hippocratic AI markets a safety-focused medical assistant trained on clinical conversations.
- The consequence question is no longer theoretical !

CLINICAL DECISION SUPPORT MODELS IN PRACTICE

- Google's Med-PaLM and Med-PaLM 2 were explicitly designed for clinical reasoning tasks
 - Tu, Tao, Shekoofeh Azizi, Danny Driess, Mike Schaeckermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll et al. "Towards generalist biomedical AI." *Njm Ai* 1, no. 3 (2024): Aloa2300138.
- Med-Gemini is positioned as a multimodal medical reasoning system
 - Saab, Khaled, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang et al. "Capabilities of gemini models in medicine." arXiv preprint arXiv:2404.18416 (2024).
- These models achieve high benchmark accuracy on medical QA.
- However, *benchmark performance does not equate to safe clinical deployment!*

ARCHITECTURES



- General-purpose LLM chatbots (e.g., GPT based systems).
- Retrieval-augmented medical assistants (LLM + curated medical sources).
- Domain-specialized medical LLMs (Med-PaLM family).
- Multi-model orchestration systems combining several LLMs.

FAILURE MODE ACROSS ALL ARCHITECTURES

- Despite architectural differences, ALL systems rely on **probabilistic** language generation
- None enforce hard clinical decision boundaries, by construction
- Critical behaviors, for instance mandatory escalation, remain advisory
- Shared **safety risks** across system types

EXAMPLE: EMERGENCY SYMPTOM TRIAGE

- Chest pain, shortness of breath, or neurological symptoms are common triage scenarios.
- GPT based systems often enumerate possible causes accurately.
- However, they frequently *continue conversational probing instead of enforcing escalation* !
- Clinically, the correct action is to halt dialogue and direct immediate care.

LIMITATIONS OF LLMS (EVEN INCLUDING RAG ETC.)

- Retrieval only augments context; it **does not implement a clinical decision** procedure !
 - The system retrieves “escalate if chest pain + shortness of breath,” but retrieval alone does not *trigger* escalation.
- No explicit representation of (**clinical**) **decision state**
 - State variables a triage workflow would require: “red-flag present?”, “acuity level,” “contraindications checked?”, “pending vitals/labs,” “time since onset.”
 - Without an explicit state, the system cannot formally define or enforce what actions are permitted next.
- Generation is **typically unconstrained** with respect to mandatory safety rules
 - Even when red flags are present, a model may produce reassurance, speculative attribution (e.g., anxiety), or non-urgent recommendations
 - The decoding objective is not constraint satisfaction !
 - **Key risk:** action selection that is inconsistent with mandatory escalation criteria.
- Output **stability is not guaranteed**
 - Non-repeatability here refers to *recommendation variability*: the same case description can yield different recommendations due to sampling, prompt differences, or retrieval ordering.
 - In triage-like settings, materially different recommendations for the same presentation are unacceptable.

RETRIEVAL ALONE CANNOT ENSURE SAFE BEHAVIOR

- Clinical guidance is **conditional**, exception-heavy, and order-sensitive
 - “rule out emergent causes first” should dominate the control flow before reassurance or routine care pathways.
- Safety requirements are naturally expressed as **constraints** over actions
 - “If red-flag symptoms → escalate now,” “do not recommend home care under X,” “do not provide reassurance under Y.”
 - Purely neural models have no built-in mechanism to guarantee these constraints are satisfied.
- Medical decision support requires **evidence of coverage** over mandatory checks
 - Before any low-acuity recommendation, the system must evaluate a defined set of red flags and document that evaluation.
 - LLM can provide the checklist; it cannot guarantee the checklist was executed.
- In summary
 - Retrieval-augmented generation is useful for **information support** (summarizing and citing guidance).
 - Medical decision support additionally requires **explicit decision structure** and **constraint-enforced behavior**.

ETHICAL RISK

- Empathy can be clinically miscalibrated
 - Example: "*I know this feels scary, but many people have benign causes of chest discomfort*" can be interpreted as *medical reassurance*, even if the user also reports shortness of breath or faintness.
 - The ethical issue is not the kindness, it's the shift in perceived urgency.
- Automation bias and deference to fluency
 - Example: a user with new unilateral leg swelling after a long flight reads a fluent explanation of "muscle strain or dehydration" and delays seeking care because "the assistant sounded confident."
 - Fluency + specificity can substitute for clinical authority in the user's mind.
- Harm can occur without any single false statement
 - Example: every sentence is defensible ("could be anxiety," "monitor symptoms," "seek care if worse"), but the interaction fails to *prioritize* the red-flag pathway or asks follow-ups instead of escalating.
 - The failure mode is omission/ordering: "continue assessment" when the right action is "stop and escalate."
- Ethical risk is cumulative at the interaction level
 - Example: across 6–8 turns, the system repeatedly normalizes symptoms ("common," "often benign"), the user commits to a low-risk interpretation, and escalation becomes less likely - even if a late "consider urgent care" line appears.
 - Impact emerges from the trajectory of the dialogue, not a single utterance.

HEALTHCARE: COMPELLING NEED FOR NEUROSYMBOLIC AI

- Clinical work is procedural and stateful, not just “question answering”
- Guidelines are conditional, threshold-based, and exception-rich
- Escalation is a safety property, not a stylistic choice
- Accountability requires auditability at the level of decisions

HEALTHCARE: COMPELLING NEED FOR NEUROSYMBOLIC AI

- Medicine is anchored in formal domain knowledge and practice guidelines
 - Clinical care depends on structured knowledge: differential diagnosis frameworks, contraindications, drug interactions, dosage rules, and treatment pathways.
 - “Do not prescribe X if pregnant / if QT prolongation / if on interacting drug Y” is naturally represented as explicit knowledge + rules.
- Guidelines encode obligations
 - Many workflows contain required actions (“must-check” items), not optional advice.
 - Before a low-acuity recommendation, the system must screen for a defined set of red flags; if present, it must escalate.
- High cost of omissions and partial reasoning
 - Harm often arises from what is *not* asked or *not* ruled out (missing a critical symptom question, missing a contraindication, missing a time-critical condition).
- Temporal and sequential reasoning is routine
 - Medical decisions depend on timelines (onset, progression, duration, response to prior treatment) and ordered steps (triage → tests → interpretation → treatment).
 - “worsening over 6 hours” is qualitatively different from “stable for weeks.”
- Clinical environments demand consistency, traceability, and accountability
 - For clinical acceptance (and eventually regulation), systems must support reproducible behavior and an audit trail of the decision basis.
 - “Escalated because criterion A+B met; deferred medication because contraindication C.”

NEURAL AI LIMITATIONS

Desired capabilities	Neural AI Limitations (RAG etc included)
<i>Stateful decision support (explicit clinical/workflow state)</i>	LLMs operate primarily as stateless generators over a prompt; “memory” is typically implicit, lossy, and not a formally defined state model (what’s known/unknown/checked).
<i>Procedural control flow (stepwise, order-sensitive workflows)</i>	LLM generation does not inherently encode or enforce control flow, so it can mis-order steps
<i>Conditional and threshold-based criteria evaluation</i>	LLMs can restate criteria but are not guaranteed to deterministically evaluate conditions across all relevant variables, especially when information is incomplete or distributed across turns.
<i>Exception handling and precedence rules</i>	Pure generation tends to optimize narrative coherence rather than ensuring all exceptions/overrides were applied correctly.
<i>Mandatory-check coverage (“must-check” obligations)</i>	LLMs may omit non-salient but mandatory checks; retrieval can display a checklist but cannot guarantee it was executed.
<i>Safety-critical escalation behavior</i>	LLMs may acknowledge urgency yet continue routine dialogue; without explicit enforcement, escalation can be diluted, delayed, or buried in prose.
<i>Action constraints (“must-not” / contraindication enforcement)</i>	LLMs can mention constraints but cannot guarantee the proposed action set excludes disallowed options without an external constraint-checking mechanism.

NEURAL AI LIMITATIONS

Desired capability	Neural AI Limitations
<i>Temporal reasoning over onset/progression and sequences</i>	LLMs can be inconsistent about timelines and do not reliably maintain temporal variables as first-class, checkable state.
<i>Consistency / stability under equivalent inputs</i>	Recommendations may vary with prompt phrasing, retrieval ordering, or sampling parameters.
<i>Traceability and auditability of decisions</i>	LLMs can produce post-hoc explanations, but those are not equivalent to a verifiable decision trace.
<i>Formal grounding in medical knowledge artifacts (guidelines, pathways, formularies)</i>	LLMs can summarize guidelines, but mapping text to computable representations (criteria, decision tables, contraindication rules) is non-trivial and error-prone; without formalization, the system cannot reliably “apply” the guideline as policy.
<i>Robustness to missing / uncertain information (explicit uncertainty handling)</i>	LLMs may proceed with assumptions or produce plausible defaults unless explicitly constrained to halt/ask under uncertainty.
<i>Verification of outputs against requirements</i>	LLMs do not natively provide proof/verification; any “self-check” remains probabilistic and can fail silently.
<i>Accountable division between information support vs decision authority</i>	LLMs’ conversational style can blur this boundary, leading users to treat outputs as authoritative decisions unless the system has explicit governance and control logic.

HEALTHCARE: A CLEAR CALL FOR NEUROSYMBOLIC AI

Key Capability Requirements for Healthcare Decision Support

Decision Structure & Control

- Stateful Workflow Management
- Procedural Control Flow
- Escalation Enforcement

Rule-Governed Reasoning

- Conditional Criteria Evaluation
- Thresholds & Exceptions
- Contraindication Rules

Obligations & Completeness

- Mandatory Check Coverage
- Handling Missing Information

Knowledge Formalization

- Clinical Guidelines Encoding
- Computable Medical Rules

Reliability & Robustness

- Consistency & Stability
- Temporal Reasoning

Assurance & Accountability

- Traceability & Auditability
- Output Verification