LECTURE 1

NEUROSYMBOLIC AI: INTRODUCTION

JANUARY 5, 2026

DATASCI290

INTRODUCTION

- Introduce yourself
  - **Anything** you would like to share about yourself
- Live survey, on your interests, expectations from the course, and your professional & academic background

CONSEQUENCE-AWARE COGNITIVE AGENTS

- A consequence-aware cognitive/AI agent is an AI system that does not mess up
- We see Neurosymbolic AI as the methodology, technology, and toolkit for building consequence-aware cognitive agents

WHAT THIS IS COURSE IS ABOUT

- Building AI agents that
  - Behave (correctly)
  - You, as a designer and developer can
    - Defend
    - Explain (the behavior of)
- Learning and investigating Neurosymbolic AI as a *means towards building safet and trustworthy cognitive agents*
- By grounding AI Agents in *structured knowledge* and *explicit reasoning*

WHAT WE WILL ENOUNCTER, BUT NOT WHAT THIS COURSE IS PRIMARILY ABOUT

- Generative AI (in itself)
  - Our interest is in *using* GenAI/LLMs
- AI Agents or Models for some actively investigated problems
  - Such as agents for autonomous scientific discovery, or models for specific topics such as geometry, matrix multiplication etc.
  - We *are* interested in some aspects of the design of such models however, especially the reasoning capabilities
- Regulations, Ethics, Law
  - We will certainly encouter rules and regulations and safety constraints, the above are however disciplines in themselves
- Building “fully autonomous” systems
  - Our primary interest in this course is in “human-in-the-loop” AI systems

QUESTION

HOW DO **HUMANS** INTERACT WITH THE ENVIRONMENT, THINK, LEARN, ACT ?

HUMAN COGNITION

- Humans ….
  - Interact with the environment using a combination of
    - Perception: transforming sensory inputs from their environment into symbols, and
    - Cognition:
      - mapping symbols to knowledge (about the environment) for supporting
        - abstraction
        - reasoning by analogy
        - long-term planning
- Humans *“can”* be trusted to behave, be safe, follow rules
  - In this context the term “can” merely refers to potential capability, not an affirmation :)

QUESTION

HOW DO (YOU THINK) **MACHINES** INTERACT WITH THEIR ENVIRONMENT, “THINK”, LEARN, ACT ?

MACHINES

- Machine cognition relies on learning from data
  - Pattern recognition
  - **next-token prediction**
- What is next-token prediction anyway
- Fundamentally, next-token prediction is the one-single thing that GenAI/ LLMs can currently do
  - That with “auto-regression” gives you the ChatGPT like “intelligence”

GENERATIVE AI (LLM)

- Modern AI models increasingly **shape decisions**
  - Versus merely producing answers
- Foundation models such as GPT, Claude, Gemini, LLaMA, DeepSeek, ….
  - Demonstrate *very* powerful perception and language competence
  - *Super* useful in a variety of personal and professional tasks !
- However, the fluent generation alone does not ensure correct behavior
  - An issue when decisions carry consequences

QUESTION

WHAT IS A TASK WHERE YOU FIND AN LLM (CHATGPT, GEMINI, YOUR FAVORITE…) VERY USEFUL OR ADMITTEDLY INDESPENSABLE ?

WHY ARE INCORRECT AND/OR NOT EXHAUSTIVE RESPONSES OK ?

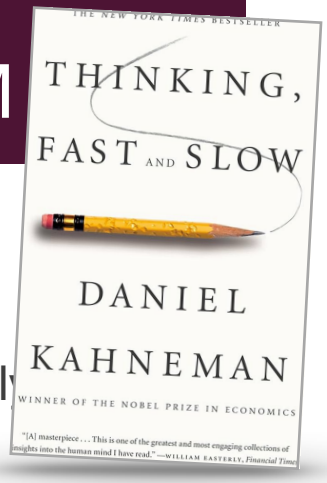
THE CENTRAL QUESTION

- Can large generative models acquire cognitive functionality **purely from data** ?
- Hypothesis: next-token prediction over massive corpora induces an *emergent world model*
  - This view is supported by the “surprising” capabilities of state of the art AI models (such as as GPT5, Gemini 3, Claude…)
- However, internal representations remain **opaque** and (thus) **difficult to validate**
- And that …. limits **TRUST** !

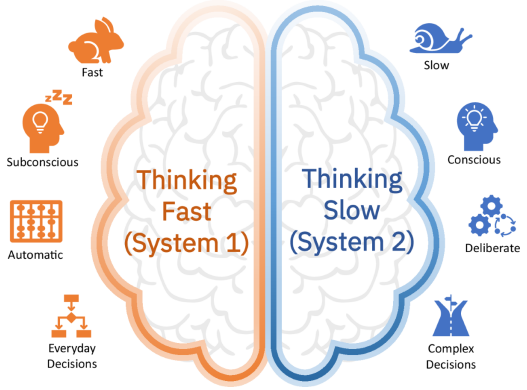
HUMAN INTELLIGENCE AS A REFERENCE POINT

- Human intelligence separates **perception** from **structured cognition**
- Perception: Maps raw input to **symbols** - objects, words, events …
- Cognition **reasons** over **knowledge structures**
  - To support planning, abstraction
- Humans can **articulate** rules, exceptions, justifications, …

NEUROSYMBOLIC AI: A **HYBRID** SYSTEM



- Behavioral Psychology: Humans think **fast** *and* **slow**, simultaneously
- Adaptation to AI systems
  - System 1/S1 (Fast) : intuitive, reflexive, pattern-driven processing, **breadth**
  - System 2/S2 (Slow) : deliberative reasoning using explicit rules and constraints, **depth**
- Neural systems (GenAI) resemble System 1 behavior
- Symbolic reasoning mirrors System 2 processes
- Effective AI systems require **both** operating together
- This hybrid/dual is the essence of **Neurosymbolic** AI !



WHAT NEURAL MODELS DO EXCEPTIONALLY WELL

- Large-scale perception and pattern recognition from raw data.
- Self-supervised learning over text, images, and code.
- Examples: GPT-5 for language, AlphaFold for protein folding.
- Neural models interpolate well within learned distributions.

LIMITS OF PURELY NEURAL APPROACHES

- Implicit knowledge representations are difficult to inspect or debug.
- Reasoning behavior varies with prompts and context.
- Hard constraints and invariants are not enforced by default.
- Failures concentrate at edge cases and rare conditions.

HOW DO WE BUILD NEUROSYMBOLIC SYSTEMS ?

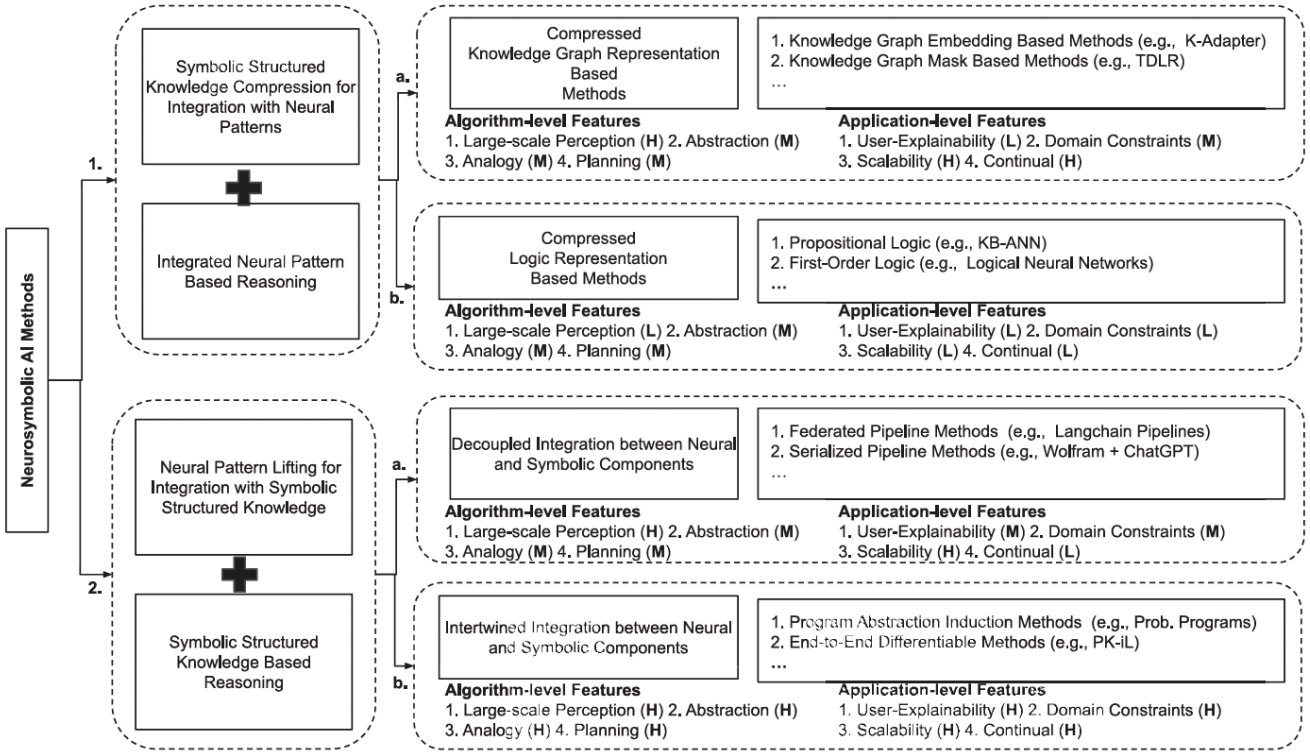
THE NEUROSYMBOLIC PROPOSITION

- 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

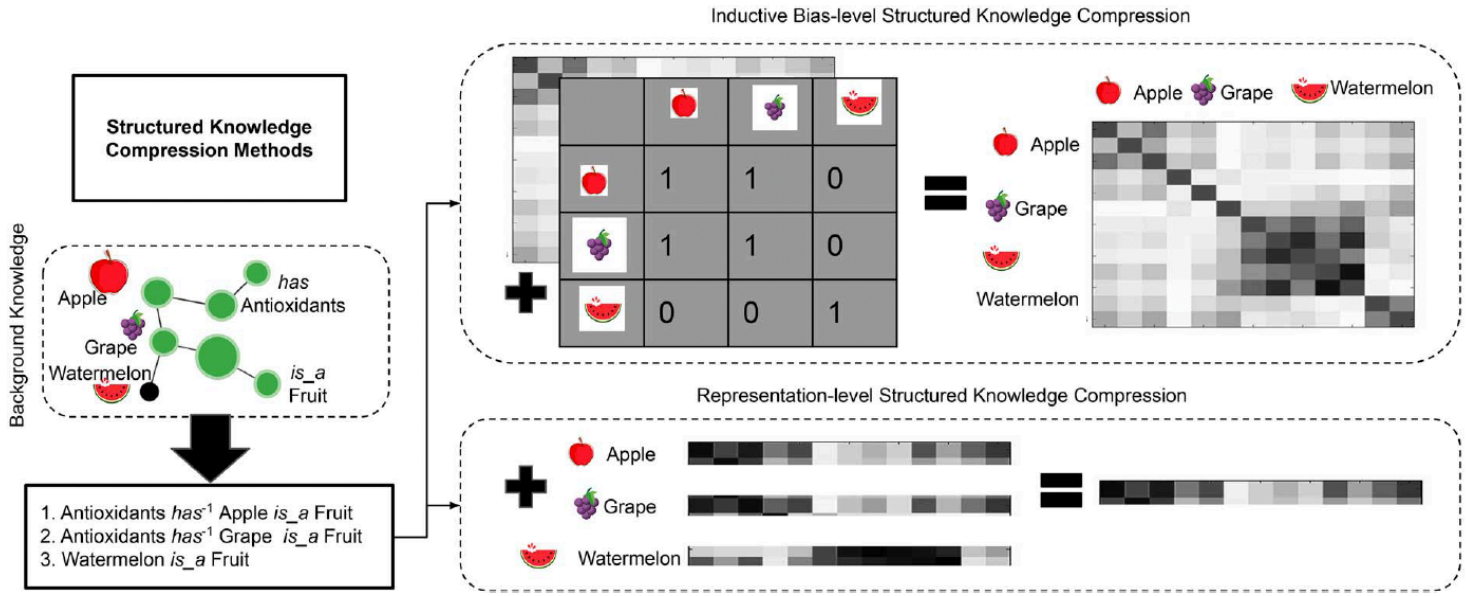
THE FOUR INTEGRATION STRATEGIES



- Just for reference today, it will take us the entire semster 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

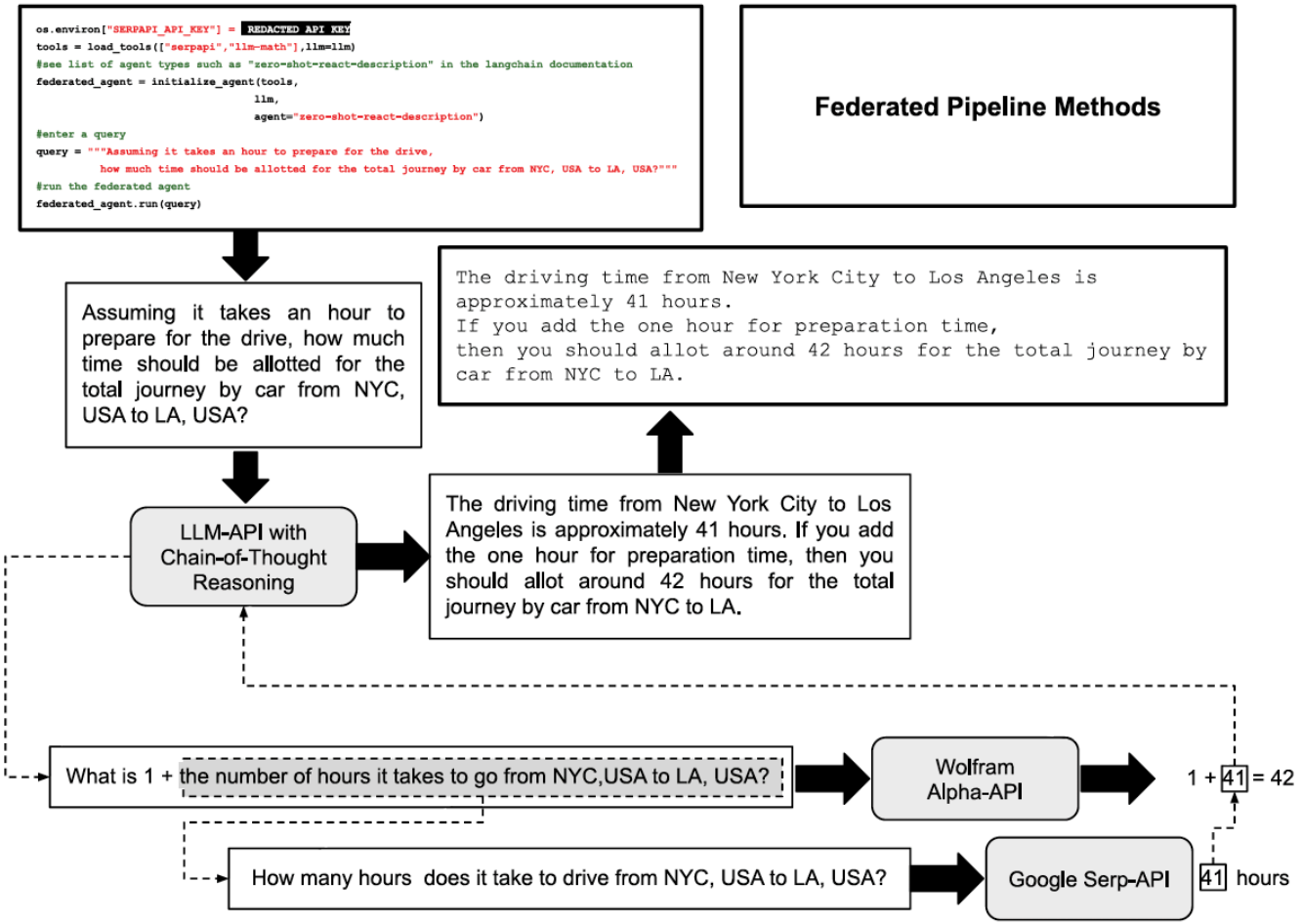
1B: LOWERING, VIA COMPRESSED FORMAL LOGIC REPRESENTATIONS

- **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 !

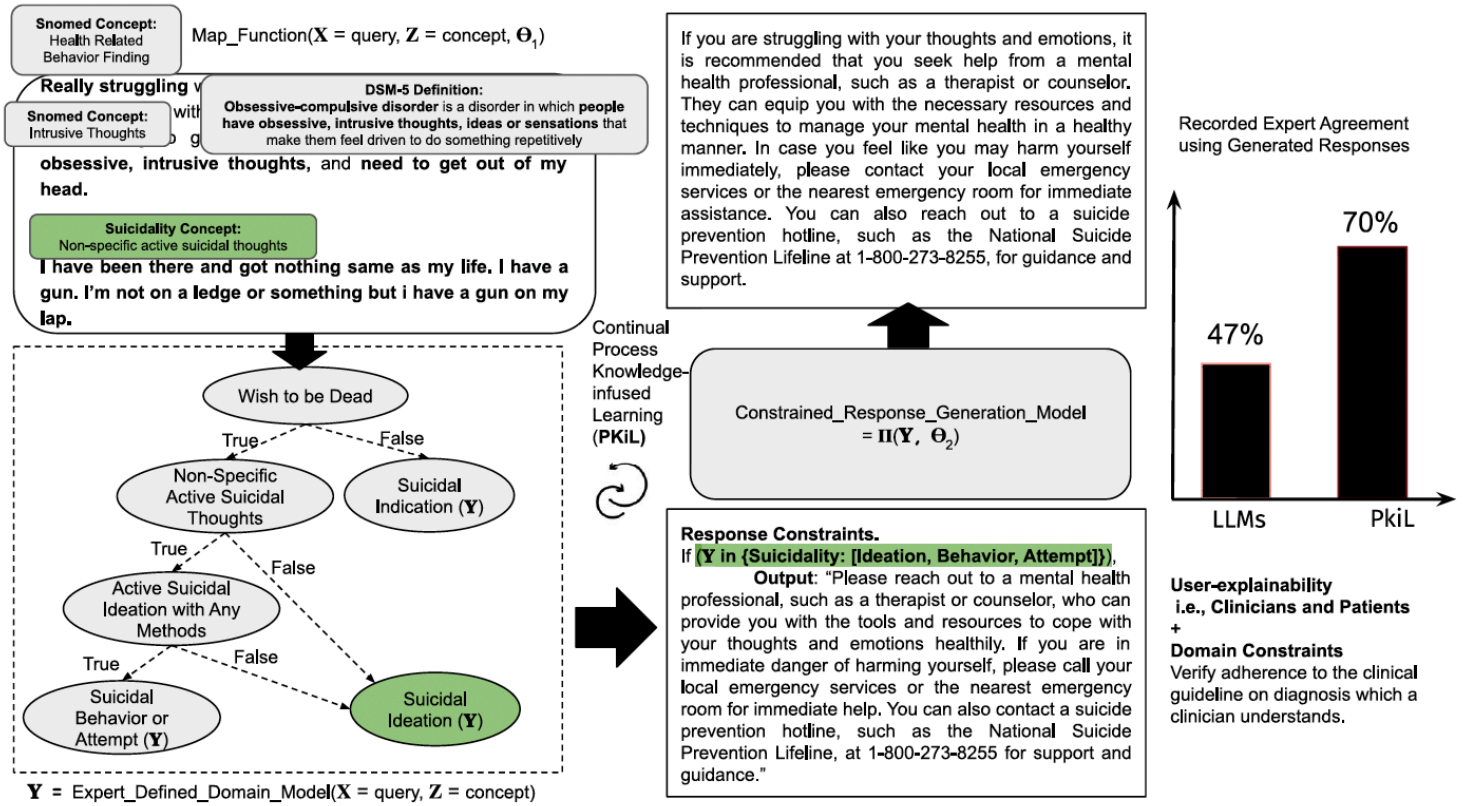
2A: EXAMPLE


os.environ["SERPAPI_API_KEY"] = REDACTED API KEY
tools = load_tools(["serpapi","llm-math"],llm=llm)
#see list of agent types such as "zero-shot-react-description" in the langchain documentation
federated_agent = initialize_agent(tools,
llm,
agent="zero-shot-react-description")
#enter a query
query = """Assuming it takes an hour to prepare for the drive,
how much time should be allotted for the total journey by car from NYC, USA to LA, USA?"""
#run the federated agent
federated_agent.run(query)
Federated Pipeline Methods
Assuming it takes an hour to prepare for the drive, how much time should be allotted for the total journey by car from NYC, USA to LA, USA?
The driving time from New York City to Los Angeles is approximately 41 hours.
If you add the one hour for preparation time,
then you should allot around 42 hours for the total journey by car from NYC to LA.
LLM-API with Chain-of-Thought Reasoning
The driving time from New York City to Los Angeles is approximately 41 hours. If you add the one hour for preparation time, then you should allot around 42 hours for the total journey by car from NYC to LA.
What is 1 + the number of hours it takes to go from NYC,USA to LA, USA?
Wolfram Alpha-API
1 + 41 = 42
How many hours does it take to drive from NYC, USA to LA, USA?
Google Serp-API
41 hours

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


Snomed Concept: Health Related Behavior Finding
Map_Function(X = query, Z = concept, Θ1)
Really struggling
Snomed Concept: Intrusive Thoughts
DSM-5 Definition: Obsessive-compulsive disorder is a disorder in which people have obsessive, intrusive thoughts, ideas or sensations that make them feel driven to do something repetitively
obsessive, intrusive thoughts, and need to get out of my head.
Suicidality Concept: Non-specific active suicidal thoughts
I have been there and got nothing same as my life. I have a gun. I'm not on a ledge or something but i have a gun on my lap.
Wish to be Dead
True
False
Non-Specific Active Suicidal Thoughts
Suicidal Indication (Y)
Active Suicidal Ideation with Any Methods
Suicidal Behavior or Attempt (Y)
Suicidal Ideation (Y)
Y = Expert_Defined_Domain_Model(X = query, Z = concept)
Continual Process Knowledge-infused Learning (PKiL)
Constrained_Response_Generation_Model = Π(Y, Θ2)
If you are struggling with your thoughts and emotions, it is recommended that you seek help from a mental health professional, such as a therapist or counselor. They can equip you with the necessary resources and techniques to manage your mental health in a healthy manner. In case you feel like you may harm yourself immediately, please contact your local emergency services or the nearest emergency room for immediate assistance. You can also reach out to a suicide prevention hotline, such as the National Suicide Prevention Lifeline at 1-800-273-8255, for guidance and support.
Response Constraints.
If (Y in {Suicidality: [Ideation, Behavior, Attempt]}),
Output: "Please reach out to a mental health professional, such as a therapist or counselor, who can provide you with the tools and resources to cope with your thoughts and emotions healthily. If you are in immediate danger of harming yourself, please call your local emergency services or the nearest emergency room for immediate help. You can also contact a suicide prevention hotline, such as the National Suicide Prevention Lifeline, at 1-800-273-8255 for support and guidance."
Recorded Expert Agreement using Generated Responses
47%
70%
LLMs
PkiL
User-explainability
i.e., Clinicians and Patients
+
Domain Constraints
Verify adherence to the clinical guideline on diagnosis which a clinician understands.