CHAPTER 1

INTRODUCTION

The age of AI has reached its peak — it's everywhere, shaping industries, communication, and even our understanding of truth. Yet with this rapid advancement come significant risks that demand careful mitigation. One of the most pressing dangers is the rise of misinformation, amplified by AI's ability to generate persuasive but false narratives.

Interestingly, when humans encounter new or conflicting information, we tend to preserve our existing beliefs by detecting and resisting contradictions — for instance, knowing that *Pluto was once considered a planet* and also that *it no longer is*.(c3) This capacity to maintain parallel yet context-bound beliefs allows us to navigate change without losing coherence in our broader understanding. This psychological tension, known as **cognitive dissonance**, arises when we hold two opposing ideas at once. Though uncomfortable, this dissonance serves as a defense mechanism — a way for our minds to protect the coherence and integrity of what we believe to be true. (bjd 2?)

However, this mechanism is largely absent in current large language models (LLMs). When faced with new or conflicting information, LLMs are quick to adapt and learn, but this adaptability comes at a cost: contradictory updates can disrupt previously stable knowledge. Studies show that introducing conflicting data can degrade up to **80% of unrelated knowledge** within an LLM, highlighting a fundamental gap between human cognition and machine learning. (C3)

CHAPTER 2

PROBLEM DEFINITION

LLMs often exhibit variability in their responses — when prompted about the same topic in slightly different ways, they may generate different answers. However, what they typically fail to provide is the *context* behind these differing answers, leaving users without insight into why or how these variations occur. This lack of contextual grounding becomes especially concerning in domains like scientific discovery, where the presentation of contradictory information can have serious implications for accuracy and trust. Moreover, research has shown that LLMs sometimes produce correct answers simply to maximize apparent accuracy, while internally assigning their *belief probabilities* to different, even conflicting, outcomes — a phenomenon known as the **revealed belief framework**. Such discrepancies highlight deeper epistemic instability in LLMs, where surface-level correctness may mask underlying confusion or inconsistency in their learned representations.

These inconsistencies can manifest in various forms of **epistemic dissonance**, reflecting how inconsistencies can emerge from their reasoning and representational limits, especially in scientific domains. In **chemistry**, for instance, a model may predict a reaction as spontaneous when expressed in SMILES notation but label it non-spontaneous under IUPAC format—revealing **representational dissonance**. In **physics**, it might claim that "entropy increases in isolated systems" yet simultaneously predict a decrease in entropy for a cooling star, showing **inferential dissonance**. In **biology**, models sometimes disagree with experimental gene-editing data, a form of **evidential dissonance** arising from conflicting empirical interpretations. Together, these cases illustrate the multifaceted nature of epistemic dissonance in LLMs, mirroring but not mastering the nuanced balance of human reasoning.

To move beyond merely identifying these inconsistencies, we aim to **quantify epistemic dissonance** as a way to assess the reliability and epistemic integrity of LLMs. By developing measurable indicators of when and how dissonance arises—across representational, inferential, evidential,

temporal, and normative dimensions—we can better understand the internal coherence of model reasoning. The goal is not only to **address and mitigate** this dissonance but also to try and **harness it** as a diagnostic and potentially generative tool for scientific research. If properly characterized, epistemic dissonance could reveal where models struggle, adapt, or innovate—offering valuable insights into both the strengths and limits of machine cognition in knowledge discovery.

CHAPTER 3

LITERATURE REVIEW

LLMs are statistical models built on deep neural networks designed to process, understand and generate human-like text, but its a debated fact if LLMs have the aptitude for reasoning or if they if the responses they are giving are an illusion birthed from their rote memorization of data. The most common way LLMs are measured through their performance is through prompting, where the result is compared to the actual fact to see correctness of response, here it was discovered using the Revealed Belief framework that while LLMs often perform well on multiple-choice questions about probability, this success does not reflect a genuine ability to reason about uncertainty. Their **revealed beliefs**—the probability distributions underlying their answers—are inconsistent, often assigning undue weight to unlikely outcomes. Moreover, LLMs fail to **appropriately update their beliefs** when presented with new evidence, and their responses are easily swayed by **irrelevant contextual cues**. Together, these results expose the limits of current prompt-based evaluation methods and underscore the need for more nuanced approaches to assessing reasoning under uncertainty. (C1)

Another key limitation revealed through the *Revealed Belief* framework is the inconsistency in probability allocation across contexts. When an LLM provides an answer without clarifying the epistemic context of that response—such as the evidence or uncertainty underpinning it—it risks producing misleading or even harmful misinformation, particularly within scientific domains where interpretive nuance is critical. This disconnect between surface-level correctness and underlying probabilistic inconsistency can manifest as **epistemic dissonance**, where the model's expressed statements and its internal probability structure conflict. Such dissonance has parallels in the history of scientific discovery, where conflicting data or theoretical uncertainty must be continually negotiated before consensus emerges. For instance, in chemical nomenclature or standards governed by IUPAC, evolving conventions often reflected iterative reconciliation between

contradictory empirical results and theoretical models—a dynamic that LLMs, in their current static form, fail to emulate.(c10)

To date, most attempts to address these limitations in LLMs have relied on **neural network editing** or **machine unlearning**, which focus on removing or revising specific pieces of information within the model's parameters. While these methods can correct localized factual errors, they overlook the broader epistemic structure in which facts coexist, evolve, and sometimes conflict. Consequently, such interventions fail to capture the contextual dependencies that underlie scientific reasoning and knowledge evolution.

A promising future direction lies in developing **contextualized truth modules**, **memory-augmented architectures**, and **dynamic ontology-based reasoning frameworks**.(c3) Contextualized truth modules would enable models to situate each generated claim within its evidentiary and temporal context—tracking whether a statement reflects consensus, speculation, or active debate. Memory-augmented networks could preserve evolving discourse states, allowing the model to remember prior claims and update them coherently when new evidence appears, rather than overwriting or forgetting them. Finally, dynamic ontology reasoning would allow LLMs to map knowledge relationships in real time, dynamically updating connections between concepts as scientific contexts shift.

Rather than treating epistemic dissonance as a flaw to eliminate, this research proposes **harnessing it as a tool for discovery (c4)**—allowing models to expose points of conceptual tension, uncertainty, or evolving consensus that mirror real scientific processes. By embedding contextual awareness and adaptive reasoning mechanisms into LLMs, we aim to transform them from static repositories of information into *interactive epistemic agents* capable of supporting reflective, uncertainty-aware scientific inquiry.

To empirically explore epistemic dissonance, the astrophysics domain provides an ideal testbed due to its evolving and often contested body of knowledge. Fields such as cosmology, where debates

like the *Hubble tension* or early-universe parameter discrepancies remain unresolved, present a natural setting to study how LLMs internalize, reconcile, or fail to reconcile competing scientific claims. Applying LLMs trained or fine-tuned on astrophysical corpora can help reveal whether these models merely reproduce the surface structure of scientific discourse or demonstrate deeper adaptive reasoning about uncertainty and evidence. By analyzing how models respond to conflicting or updated astrophysical data, we can evaluate whether epistemic dissonance manifests as measurable instability in their internal representations of scientific facts.

Recent work suggests that such dissonance can be **quantified** through statistical metrics that capture inconsistency and instability in model outputs. For instance, **semantic entropy** provides a measure of the diversity or ambiguity in a model's semantic predictions—indicating whether it is mixing incompatible conceptual frames when responding to a prompt. Similarly, **sequence-level KL divergence loss** (c10) can be used to evaluate how a model's probabilistic distribution shifts when new information is introduced, thereby revealing whether it appropriately updates its beliefs or remains anchored to outdated priors. Together, these measures offer a principled way to characterize the degree of epistemic dissonance in LLMs across scientific contexts.

Building on this, an effective approach to managing and leveraging such dissonance would involve constructing an **updating knowledge-graph framework** grounded in the three future research directions identified earlier—contextualized truth modules, memory-augmented networks, and dynamic ontology reasoning. In this setup, each scientific claim or observation would be represented as a node within a knowledge graph enriched with contextual metadata such as time of discovery, confidence score, and supporting or contradicting evidence. When new data emerge, **contextualized truth modules** could evaluate how this information aligns or conflicts with existing nodes, adjusting belief weights accordingly. **Memory-augmented networks** would preserve prior reasoning states, ensuring that updates occur through integration rather than catastrophic overwriting. Meanwhile, **dynamic ontology-based reasoning** would allow the graph to restructure itself as the conceptual landscape of the domain evolves—linking or delinking theories as scientific consensus shifts.

Such an adaptive knowledge architecture would not only mitigate the problem of static factual recall in LLMs but also provide a traceable record of epistemic evolution. The model's outputs could thus become more reflective of how scientific knowledge genuinely progresses—through cycles of evidence, contradiction, and resolution—ultimately transforming epistemic dissonance from a liability into a *computational signal* of inquiry and discovery.

This framework could be empirically tested through domains characterized by active scientific debate, such as astrophysics and particle physics. For instance, unresolved questions like the *Hubble tension* or the evolving acceptance of the *Higgs Boson* provide rich ground for examining how LLMs manage conflicting evidence over time. A detailed exploration of these case studies will be discussed in future study.

CHAPTER 4

RESEARCH / TECHNOLOGY GAPS AND CHALLENGES

Despite growing interest in cognitive and epistemic dissonance in large language models (LLMs), current approaches remain fragmented—both in conceptual framing and in measurable evaluation. Several key gaps emerge at the intersection of epistemic reasoning, knowledge evolution, and domain-specific AI modeling.

4.1 Loss of Adaptive Reasoning through Traditional Fixes.

Existing solutions like unlearning, model patching, or neural editing often eliminate conflicting knowledge at the cost of adaptive reasoning. While these methods restore local consistency, they disrupt the model's ability to engage in dynamic epistemic calibration—essential for reconciling conflicting evidence across time or domains. Inspired by works such as *In Praise of Stubbornness* and *Do LLMs Exhibit Cognitive Dissonance?*, there is a need for a **continual epistemic calibration mechanism** that updates internal belief structures without catastrophic forgetting.

4.2 Dissonance Treated as a Defect Rather Than a Feature.

Most studies frame cognitive dissonance in AI as a pathology to be minimized rather than a productive force driving conceptual change. Drawing from *Cognitive Dissonance AI* and epistemological analyses like *Model Choice and Crucial Tests*, we argue that epistemic dissonance should be reframed as a **driver of paradigm evolution**—analogous to how contradictions in science fuel theoretical progress.

4.3 Limited Exploration in Scientific Domains.

Research has largely overlooked the application of epistemic dissonance modeling in fields such as astrophysics and cosmology, where competing models (e.g., ΛCDM vs. Modified Gravity) embody scientific contradiction. Extending this inquiry to domain-specialized models, as hinted by

AstroLLaMA-2 and *Hubble Tension*, could reveal how LLMs navigate evolving scientific uncertainty. This motivates a **scientific discovery–oriented extension** of dissonance modeling.

4.4 Lack of Unified Methodology for Measuring Epistemic Dissonance.

There is currently no single, standardized way to quantify epistemic dissonance. It manifests in multiple forms—conflicting embeddings, contradictory claims, belief gaps, or unstable reasoning trajectories—each requiring distinct yet connected metrics. Prior efforts (e.g., *Do LLMs Exhibit Cognitive Dissonance?*) rely on isolated indicators such as belief divergence or response inconsistency. To move forward, we propose developing **multi-level evaluation metrics**—including conflict frequency, stance entropy, graph inconsistency scores, and contradiction centrality—to capture how inconsistencies arise and propagate within or across belief systems.

4.5 Fragmented Technical Foundations.

Current technical infrastructures lack mechanisms for conflict-aware updating or explicit representation of contradictory claims. Gaps persist in:

- **Belief embedding methods**, which need to represent epistemic stances (agreement, contradiction, uncertainty).
  Conflict-aware updating, where models must reconcile inconsistencies across evolving datasets.
- **Dissonance graphs**, which map contradictions and supports among scientific claims to trace theory evolution.

Together, these gaps point toward the need for **a unified epistemic framework** capable of embedding, updating, and quantifying belief contradictions in LLMs—transforming epistemic dissonance from a liability into a lens for studying reasoning, adaptation, and discovery.

CHAPTER 5

OBJECTIVES

In this paper, we will explore Epistemic Dissonance in Large Language Models (LLMs) — the inconsistencies that arise when models hold or generate conflicting knowledge. As AI systems become integral to scientific discovery, understanding these internal contradictions becomes crucial for ensuring reliability and interpretability. Our work seeks to bridge the gap between human cognitive mechanisms for managing conflicting beliefs and the unstable knowledge representations of LLMs.

We aim to:

1. Understand the nature and forms of epistemic dissonance in LLMs, comparing them to human cognitive dissonance and analyzing how they manifest across scientific domains.
2. Address the impact of epistemic dissonance on model reliability by identifying, measuring, and mitigating contradictions that disrupt learned knowledge.
3. Harness epistemic dissonance as a potential diagnostic and generative tool to improve scientific reasoning and knowledge discovery in AI systems.

CHAPTER 6

DATA EXPLORATION

CHAPTER 7

CONCLUSION OF CAPSTONE PROJECT PHASE - 1

Through this deeper exploration of epistemic dissonance in AI models, we have gained a more nuanced understanding of how conflicting knowledge and adaptive reasoning coexist within large language systems. Our study highlights that dissonance is not merely a defect to be minimized but a potential driver of epistemic growth—mirroring how scientific paradigms evolve through contradiction and reconciliation.

We identified critical gaps in current approaches, including the overuse of unlearning techniques that erase adaptive reasoning, the lack of conflict-aware updating mechanisms, and the absence of unified metrics to measure epistemic dissonance across its various forms. By connecting insights from both cognitive science and domain-specific research in astrophysics and cosmology, we hope to outline pathways toward continual epistemic calibration, belief-embedding methods, and dissonance-aware evaluation frameworks.

In essence, our findings suggest that embracing epistemic dissonance—rather than suppressing it—could enable AI systems to reason more like scientists: dynamically, self-correctively, and open to conceptual change. This perspective lays the foundation for future work aimed at building models capable of both recognizing and productively managing internal contradictions as part of their reasoning process.

CHAPTER 8

PLAN OF WORK FOR

CAPSTONE PROJECT PHASE - 2

BIBLIOGRAPHY

APPENDIX

LLMs-