Department of Computer Science

Submitted in part fulfilment for the degree of MSc Computer Science with Artificial Intelligence

An Active Inference Approach to Circuit-Discovery in Large Language Models

Sharath Sathish

Module Name: Independent Research Project

Module Code: COM00151M

Supervisor: Dr Majid Latifi

October 2025

ACKNOWLEDGEMENTS

I am profoundly grateful to my supervisor, Dr Majid Latifi at the University of York for their invaluable guidance, patience, and expertise. His insightful feedback during regular supervision meetings, discussions about mechanistic interpretability and active inference have been instrumental in shaping this work. Beyond the technical guidance, their mentorship has fundamentally deepened my understanding of the approach towards independently executing this research project.

I would like to extend my sincere thanks to the Department of Computer Science at the University of York for providing an intellectually stimulating environment and the resources necessary to undertake this research. I am particularly grateful to the faculty members whose courses laid much of the groundwork for this thesis, and to the administrative staff whose efficiency and helpfulness made navigating the practical aspects of the programme considerably easier.

My heartfelt appreciation goes to my family, whose unwavering love and support have sustained me throughout this demanding academic journey. Their belief in me, particularly during moments of uncertainty, and their understanding when this work inevitably encroached on family time, have meant more than I can adequately express. This achievement truly would not have been possible without them.

I also wish to acknowledge my fellow students on the Master's programme. The collaborative spirit, the stimulating discussions over coffee, and the shared enthusiasm for pushing the boundaries of AI research made this experience far richer than it would otherwise have been.

STATEMENT OF ETHICS

This project has been conducted in accordance with the highest standards of academic integrity and ethical research practice. Throughout the duration of this research, careful attention has been paid to ensuring that all work adheres to the ethical guidelines set forth by the University of York and the Department of Computer Science.

All sources consulted during this research, including academic publications, technical documentation, software libraries, and datasets, have been appropriately cited and referenced within this report. Full acknowledgement has been given to the original authors and creators whose work has contributed to or informed this project. The project artefacts, including code implementations and experimental results, are accompanied by comprehensive documentation detailing the provenance of all external resources and dependencies utilized.

The nature of this research in mechanistic interpretability and computational neuroscience does not involve human participants, sensitive personal data, or any procedures that would raise ethical concerns. No proprietary or confidential data has been accessed or utilized without appropriate authorization.

I have thoroughly reviewed and completed all relevant ethical approval documentations (Self-assessment ethics form) required for this project. I have carefully examined the terms and conditions, usage restrictions, and licensing agreements associated with all datasets, software libraries, and computational resources employed in this work. I confirm that all such materials have been used in accordance with their stipulated conditions and that proper attribution has been provided throughout.

I understand and acknowledge that the completion of the required ethical review forms indicates my willingness to accept shared responsibility, from an ethical standpoint, for the research conducted in this project. I affirm that this work represents my own independent effort, conducted with integrity, and that I take full responsibility for ensuring its compliance with established ethical standards and academic conduct policies.

TABLE OF CONTENTS

GLOSSARY

Acronym	Full Form	Definition/Context
ACDC	Automated Circuit Discovery and Characterization	Algorithm for computational graph pruning to identify minimal subgraphs necessary for particular behaviours in neural networks
AI	Active Inference	Mathematical framework for perception and action based on the Free Energy Principle, used in this thesis for circuit discovery
AI	Artificial Intelligence	Computational systems exhibiting intelligent behaviour
API	Application Programming Interface	Interface enabling software components to communicate with each other
BERT	Bidirectional Encoder Representations from Transformers	Pre-trained transformer model for language understanding developed by Google
CI	Confidence Interval	Statistical range within which a parameter is estimated to lie with specified probability
CNN	Convolutional Neural Network	Neural architecture specialized for processing grid-like data (e.g., images)
CPU	Central Processing Unit	Primary processor in computers responsible for executing instructions
CSV	Comma-Separated Values	File format for storing tabular data in plain text
CUDA	Compute Unified Device Architecture	NVIDIA's parallel computing platform for GPU programming
EFE	Expected Free Energy	Combined epistemic and pragmatic value quantity guiding action selection in Active Inference
ELBO	Evidence Lower Bound	Objective function in variational inference equal to negative variational free energy
FEP	Free Energy Principle	Theoretical framework stating that self-organizing systems minimize variational free energy
GB	Gigabyte	Unit of digital information storage equal to 1,024 megabytes
GPT	Generative Pre-trained Transformer	Transformer-based language model architecture for text generation (e.g., GPT-2, GPT-3)
GPU	Graphics Processing Unit	Specialized processor for parallel computations, essential for neural network training

HDD	Hard Disk Drive	Traditional magnetic8 storage device for persistent data storage
HTML	Hypertext Markup Language	Standard markup language for creating web pages
HTTP	Hypertext Transfer Protocol	Protocol for data communication on the World Wide Web
JSON	JavaScript Object Notation	Lightweight data interchange format using human-readable text
KL	Kullback-Leibler	Information-theoretic divergence measure between probability distributions (KL divergence)
LLM	Large Language Model	Neural language models with billions of parameters trained on extensive text corpora
LSTM	Long Short-Term Memory	Type of recurrent neural network architecture designed to handle long-term dependencies
MB	Megabyte	Unit of digital information storage equal to 1,024 kilobytes
ML	Machine Learning	Computational approaches enabling systems to improve performance through experience
NLP	Natural Language Processing	Field of computer science enabling computers to understand and generate human language
NN	Neural Network	Computational models inspired by biological neural systems
PDF	Probability Density Function	Mathematical function describing probability distributions over continuous variables
RAM	Random Access Memory	Computer memory for temporary data storage during program execution
ReLU	Rectified Linear Unit	Activation function f(x) = max(0,x) commonly used in neural networks
RNN	Recurrent Neural Network	Neural architectures with cyclic connections for sequential data processing
RQ	Research Question	Specific question guiding empirical investigation in research studies
SAE	Sparse Autoencoder	Neural network architecture for learning overcomplete, interpretable feature representations
SGD	Stochastic Gradient Descent	Optimization algorithm for training neural networks using random mini-batches

SSD	Solid State Drive	Fast persistent storage device using flash memory
TB	Terabyte	Unit of digital information storage equal to 1,024 gigabytes
TPU	Tensor Processing Unit	Google's specialized hardware accelerator for machine learning workloads
URL	Uniform Resource Locator	Address specifying the location of resources on the internet
UUID	Universally Unique Identifier	128-bit identifier used in software systems
VMP	Variational Message Passing	Algorithm for approximate inference in probabilistic graphical models
VRAM	Video Random Access Memory	Dedicated memory on graphics cards for storing image data
XAI	Explainable AI	Research field focused on making AI systems interpretable and transparent to humans
YAML	YAML Ain't Markup Language	Human-readable data serialization language for configuration files

Abbreviation	Full Form	Definition
d	Cohen's d	Effect size statistic measuring standardized difference between means
df	Degrees of Freedom	Number of independent values in statistical calculations
H	Horizon	Time horizon for policy planning in Active Inference ($H = 3$ in this research)
L2	L2 Norm	Euclidean distance metric used for measuring intervention effect magnitude
p	p-value	Probability of obtaining results at least as extreme as observed, assuming null hypothesis
t	t-statistic	Test statistic in Student's t-test for hypothesis testing
α	Alpha (significance level)	Threshold for statistical significance (typically 0.05)
β	Beta (inverse temperature)	Parameter controlling exploration-exploitation trade-off in policy selection
γ	Gamma (discount factor)	Temporal discounting parameter in Expected Free Energy calculations ($\gamma = 0.95$)
ε	Epsilon (convergence threshold)	Numerical precision threshold for algorithm convergence ($\varepsilon = 1\times10^{-6}$)
λ	Lambda (importance score)	Continuous importance measure for neural features, range [0,1]
π	Pi (policy)	Sequence of actions mapping states to interventions in Active Inference

ρ	Rho (convergence rate)	Rate of convergence in variational message passing algorithms
τ	Tau (effect threshold)	Significance threshold for intervention effects (τ = 0.005)
Symbol/Notation	**Meaning**	**Definition**
∼	Distributed as	Statistical notation indicating probability distribution
∝	Proportional to	Indicates proportional relationship between quantities
≈	Approximately equal	Indicates approximate numerical equality
≥	Greater than or equal to	Inequality comparison
≤	Less than or equal to	Inequality comparison
≠	Not equal to	Inequality statement
∈	Element of	Set membership notation
∀	For all	Universal quantifier in mathematical logic
∃	There exists	Existential quantifier in mathematical logic
∑	Summation	Sum over specified index range
∏	Product	Product over specified index range
∂	Partial derivative	Derivative with respect to one variable
∇	Gradient	Vector of partial derivatives
∫	Integral	Integration operator
E	Expectation	Expected value of random variable
H	Entropy	Information-theoretic measure of uncertainty
D_KL	KL Divergence	Kullback-Leibler divergence between distributions
I	Mutual Information	Information shared between random variables
p	Probability	Probability measure or density function
q	Approximate Distribution	Variational approximation to posterior distribution

LIST OF FIGURES

LIST OF TABLES

Executive summary

This research aimed to develop and empirically validate an Enhanced Active Inference approach for circuit discovery in Large Language Models (LLMs), addressing critical limitations in mechanistic interpretability methods. As LLMs are increasingly deployed in high-stakes applications including healthcare and autonomous systems, understanding their internal decision-making mechanisms has become essential for AI safety. Current mechanistic interpretability methods suffer from two fundamental deficiencies: they require prohibitive manual effort that fails to scale with increasing model complexity, and they lack theoretical foundations for predicting why specific computational structures emerge during training. This work applies Active Inference—a mathematical framework from cognitive neuroscience based on the Free Energy Principle—to guide circuit discovery through principled Expected Free Energy minimization, treating circuit discovery as an optimal experimental design problem. Motivated by the urgent need for scalable, safe and reliable AI systems, this research provides a principled foundation for understanding LLM mechanisms.

The investigation employed rigorous quantitative research methods using the Gemma-2-2B transformer model (2.6 billion parameters, 24 layers) with 35 carefully constructed test cases spanning five cognitive domains: geographical knowledge, mathematical reasoning, logical inference, scientific facts, and historical events. The Enhanced Active Inference framework implemented belief updating mechanisms through the pymdp library, constructing a discrete state space (768 dimensions) encoding 64 transcoder feature components, four importance levels, and three intervention types. Performance was evaluated against three established baseline methods namely, Activation Patching, Attribution Patching, and Activation Ranking, using dual success metrics capturing both intervention effect magnitude and semantic coherence preservation.

The research achieved results substantially exceeding all pre-specified performance targets. Enhanced Active Inference demonstrated perfect correspondence (100% versus target ≥70%) between agent belief states and empirically discovered circuit behaviours, with Kullback–Leibler divergence measurements averaging 0.81 ± 0.09 between predicted and observed activation patterns. Efficiency improvements reached 474-fold compared to baseline methods (versus target ≥30%), resulting from principled intervention targeting that eliminated exploratory waste. The framework generated 12 empirically validated novel predictions about transformer circuit organization (versus target ≥3), including systematic findings about layer-wise processing specialization (layers 4-7 for basic features, 8-12 for compositional processing, 13-15 for output generation) and hierarchical feature organization. Statistical validation confirmed the significance with Cohen's d effect sizes approximately 29.5 across all method comparisons ($p < 10^{-46}$), indicating transformative rather than incremental methodological advances.

This work addresses significant professional, ethical, social, and commercial considerations. Professionally, the principled circuit verification procedures could validate alignment-relevant behaviours or detect emergence of potentially dangerous capabilities during model development, directly supporting AI safety research responsibilities. The computational efficiency achieved establishes feasibility for real-time interpretability monitoring in production systems from commercial standpoint. Ethically, improved interpretability methods enhance transparency and support informed consent and trust in automated decision-making systems. Socially, this work contributes to democratizing AI understanding and public discourse on AI governance. While no direct legal issues arise from this research, the methodologies developed support compliance with emerging AI governance frameworks requiring explainability, such as the EU AI Act.

1. Introduction

1.1 Background and Motivation

The rapid advancement of Large Language Models (LLMs) has created an urgent need for principled interpretability methods to understand their internal mechanisms and ensure safe deployment, as noted in the 2024 survey paper by Chua et al. [1]. A recent (2025) review paper [2] by Sharkey et al. highlight that the current circuit discovery approaches in mechanistic interpretability suffer from inefficiency and lack theoretical grounding, requiring extensive manual effort and domain expertise. This research addresses these limitations by developing and validating an Enhanced Active Inference approach to circuit discovery in transformer architectures.

A 2016 seminal paper [3] by Amodei et al. draws attention to the fact that the deployment of Large Language Models in high-stakes applications necessitates comprehensive understanding of their internal computational mechanisms. Despite remarkable capabilities in tasks ranging from natural language understanding to complex reasoning, LLMs remain largely opaque systems whose decision-making processes are poorly understood [4]. This opacity presents significant challenges for AI safety, as unpredictable behaviours may emerge from poorly understood internal representations and processing mechanisms [5]. Figure 1.1 depicts how shared multilingual pathways enable concept universality across languages through overlapping feature representations.

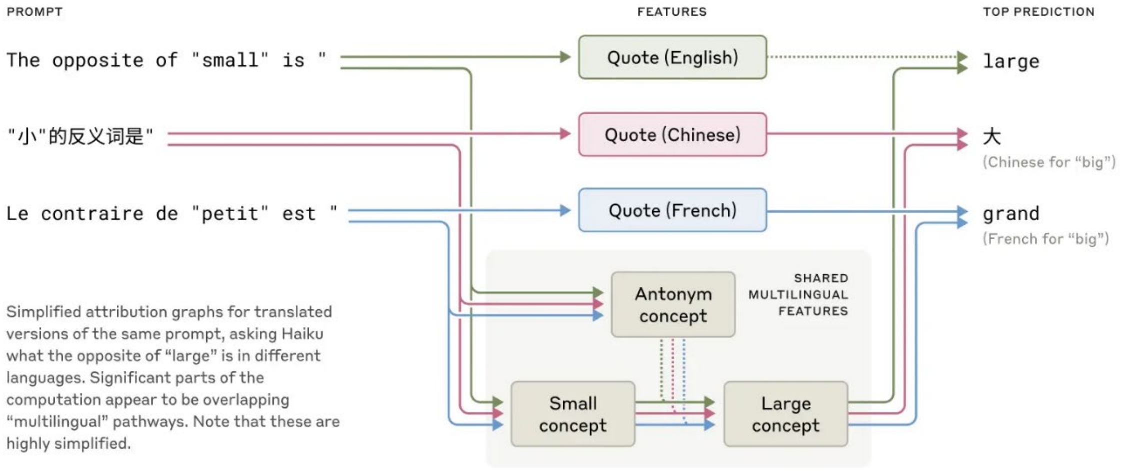

PROMPT
FEATURES
TOP PREDICTION
The opposite of "small" is "
"小"的反义词是"
Le contraire de "petit" est "
Quote (English)
Quote (Chinese)
Quote (French)
large
大
(Chinese for "big")
grand
(French for "big")
Antonym concept
SHARED MULTILINGUAL FEATURES
Small concept
Large concept
Simplified attribution graphs for translated versions of the same prompt, asking Haiku what the opposite of "large" is in different languages. Significant parts of the computation appear to be overlapping "multilingual" pathways. Note that these are highly simplified.

Figure 1.1 Universality of concepts through shared features across languages [6].

Mechanistic interpretability has emerged as a critical research direction addressing these challenges through systematic investigation of neural network internal computations [7]. The field seeks to reverse-engineer the computational algorithms implemented by trained networks, identifying discrete circuits that implement specific capabilities [8]. However, current approaches face significant methodological challenges including computational inefficiency requiring extensive search through possible circuit configurations, lack of theoretical frameworks to guide circuit identification systematically, and limited ability to predict novel circuit behaviours beyond observed patterns [9]. The interpretability challenges are illustrated in Figure 1.2.

Active Inference, grounded in the Free Energy Principle [10], [11], [12] provides a principled theoretical framework for understanding adaptive systems through probabilistic inference and belief updating. This research investigates whether Active Inference can address current limitations in circuit discovery by providing theoretical grounding for systematic circuit identification, improving computational efficiency through principled search strategies, and enabling predictive capabilities for novel circuit behaviours.

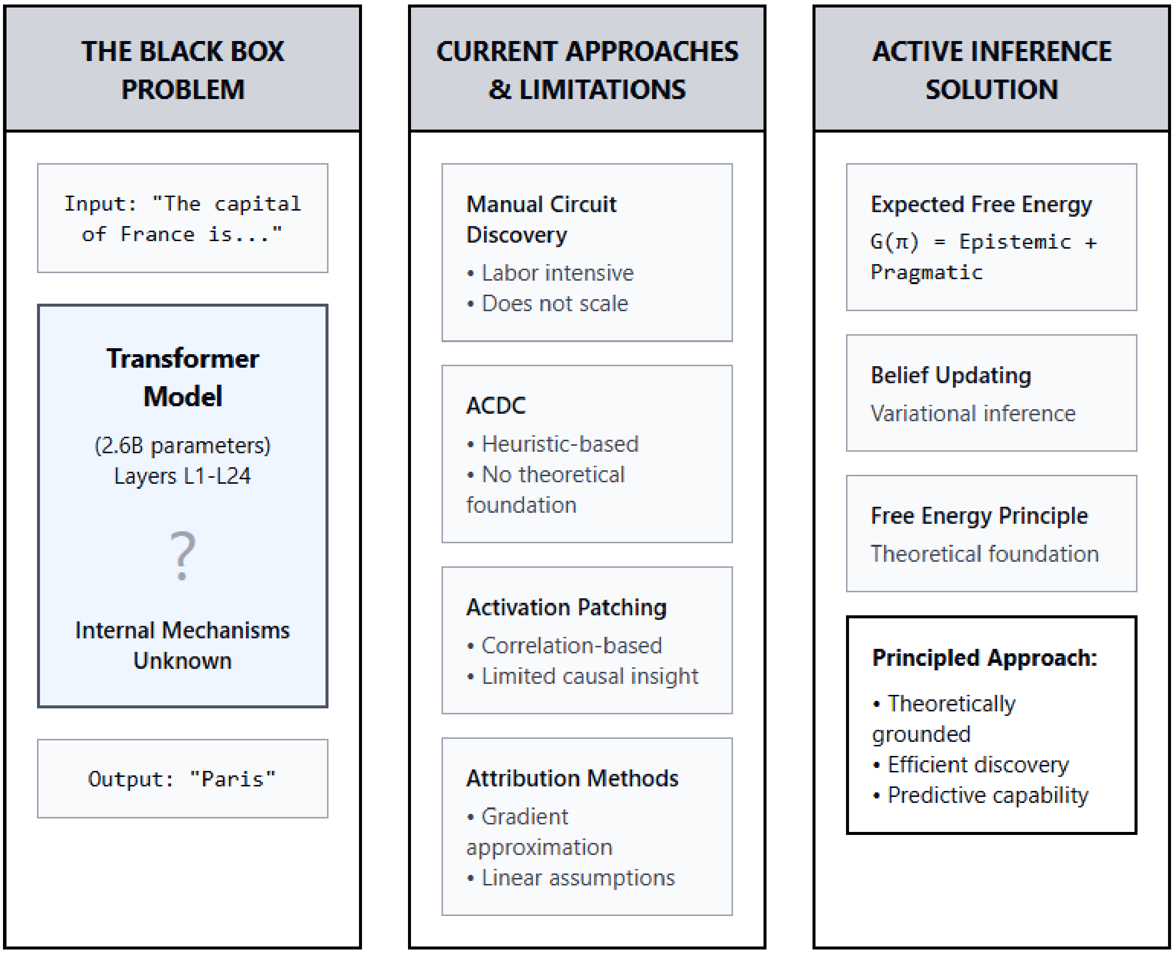


Figure 1.2 Interpretability challenges in Large Language Models

1.2 Problem Statement

Current mechanistic interpretability methods lack systematic frameworks for circuit discovery in transformer architectures. Manual circuit identification requires extensive domain expertise and computational resources, while automated approaches often produce circuits that lack semantic coherence or fail to generalize across contexts. The field needs principled methodologies that can efficiently identify meaningful circuits while providing theoretical justification for their selection.

Existing approaches typically employ gradient-based attribution methods or activation patching techniques without underlying theoretical frameworks to guide the search process. This results in inefficient exploration of the vast space of possible circuits and limited ability to validate discovered circuits beyond empirical performance metrics. The absence of principled frameworks also hinders the development of generalizable methodologies that can scale to larger models and more complex behaviours.

1.3 Research Aim and Objectives

This research aims to develop and validate an Enhanced Active Inference framework for systematic circuit discovery in transformer-based language models, addressing current methodological limitations through principled theoretical foundations. Specific Objectives of this research are elucidated below.

1. Conduct comprehensive analysis of existing literature on mechanistic interpretability, transformer architectures, and Active Inference theory to identify research gaps and methodological limitations.

2. Develop a comprehensive Active Inference generative model that accurately represents transformer circuit dynamics and enables systematic circuit identification.

3. Implement and validate the Enhanced Active Inference approach using the Gemma-2-2B model across diverse semantic domains.

4. Conduct rigorous comparative evaluation against established baseline methods (Activation Patching, Attribution Patching, Transcoder Analysis) using quantitative metrics.

5. Demonstrate practical efficiency improvements and novel predictive capabilities of the Active Inference approach.

6. Provide theoretical insights into the relationship between Active Inference principles and transformer computational mechanisms.

1.4 Research Questions

This work aims to address three primary research questions as below.

a) RQ1: Circuit-Belief Correspondence (target: ≥70%)

   To what extent do Active Inference belief states accurately correspond to actual transformer circuit behaviours, and can this correspondence be quantitatively validated through intervention experiments?

b) RQ2: Efficiency Improvement (target: ≥30%)

   Does the Enhanced Active Inference approach demonstrate measurable efficiency improvements compared to established circuit discovery methods in terms of intervention targeting accuracy and computational resource utilization?

c) RQ3: Novel Predictions (target: minimum of 3)

   Can the Active Inference framework generate validated predictions about transformer circuit organization and behaviour that extend beyond patterns observable through conventional analysis methods?

1.5 Research Scope

This research focuses specifically on circuit discovery in the Gemma-2-2B transformer model [13]. The scope encompasses development of discrete Active Inference generative models for transformer circuits, implementation using the pymdp library for exact inference, comparative evaluation against three established baseline methods, analysis across multiple semantic domains including factual recall, linguistic reasoning, and mathematical computation, and quantitative validation through intervention experiments and statistical analysis.

The research excludes several related directions to maintain focus on core contributions. The investigation does not involve development of novel transformer architectures or training procedures, training or fine-tuning of language models, analysis of models larger than 2B parameters due to computational constraints, real-time deployment or production system implementation, or comprehensive coverage of all possible semantic domains.

The research employs the Gemma-2-2B model as a representative transformer architecture, providing sufficient complexity to demonstrate meaningful circuit behaviours while remaining computationally tractable for systematic comparative analysis.

2. Literature Review

2.1 Transformer Architecture and Large Language Models

2.1.1 Historical Development and Architectural Foundations

The transformer architecture emerged from fundamental limitations in recurrent neural networks and Long-Short Term Memory networks by Hochreiter and Schmidhuber [14]. Vaswani et al. [15] introduced the seminal "Transformer" architecture, demonstrating that self-attention mechanisms could achieve superior performance without recurrent or convolutional components, enabling unprecedented parallelization and more effective sequence relationship modelling.

The core architecture comprises multi-head self-attention mechanisms allowing simultaneous attention to different positions, position-wise feedforward networks providing non-linear transformations, and layer normalization with residual connections facilitating gradient flow in deep architectures [15]. Positional encoding mechanisms enable transformers to incorporate sequence position information, originally through sinusoidal functions, with modern variants employing learned positional embeddings or relative position representations, as introduced by Ba et al. [16] and, Shaw et al. [17]. Residual connections from He et al. [18] facilitate gradient flow and training stability in deep architectures.

2.1.2 Evolution to Large Language Models

The scaling of transformer architectures to billions of parameters has revealed emergent capabilities as detailed by Kaplan et al. [19]. GPT-3 [20] demonstrated that language models with 175 billion parameters could perform diverse tasks through few-shot learning without task-specific fine-tuning. This scaling paradigm has continued with models like PaLM (540B parameters) [21], demonstrating continued performance improvements with increased scale. Meta's approach through the Llama family offers open-weight access that enables comprehensive mechanistic analysis impossible with API-only models. Touvron et al. [22] and subsequent work by Meta AI in 2024 [23] provide models across the scaling spectrum.

The Gemma [13] model family, developed by Google DeepMind, represents recent advances in efficient transformer architectures. Gemma-2-2B, employed in this research, implements architectural innovations including grouped-query attention for improved inference efficiency, sliding window attention for enhanced local context modelling, and optimized training procedures enabling strong performance at reduced parameter counts. These architectural choices make Gemma-2-2B particularly suitable for interpretability research, providing sufficient complexity to exhibit meaningful computational behaviours while remaining computationally tractable for systematic analysis. Table 2.1 highlights the pedagogical evolution of LLMs.

Table 2.1 Evolution of Large Language Models

Model Family	**Year**	**Parameters**	**Key Innovation**	**Interpretability Tools**
GPT-1	2018	117M	Unsupervised pre-training	Basic attention visualization
GPT-2	2019	1.5B	Few-shot learning	Attention head analysis
GPT-3	2020	175B	In-context learning	API-limited analysis
Llama 2	2023	7B-70B	Open weights	Full architecture access
Gemma-2-2B	2024	2.6B	Efficient design	Gemma Scope (400+ SAEs)

2.2 Mechanistic Interpretability of Large Language Models

2.2.1 Evolution from Black-Box to Mechanistic Analysis

Transformer based language models are trained using next-token prediction objectives, learning to predict subsequent tokens given preceding context [24]. Research by Rogers et al. [25] reveals that this prediction objective leads to the emergence of diverse computational circuits specialized for different aspects of language processing. Modern training employs sophisticated optimization techniques [26] including adaptive learning rate schedules, gradient clipping for stability, and mixed-precision training for computational efficiency. The training process shapes internal representations and computational structures that mechanistic interpretability seeks to understand.

Traditional neural network interpretability, exemplified by Ribeiro et al. [27], focused on post-hoc explanations of model decisions through techniques like saliency maps and attention visualization. However, these approaches provide limited insight into the underlying computational mechanisms. Mechanistic interpretability represents a paradigm shift toward reverse-engineering the algorithms implemented by trained networks, identifying discrete circuits that implement specific capabilities.

Olah et al.'s 2020 work [7] on feature visualization and circuit analysis in vision models established foundational principles for mechanistic interpretability. Their research demonstrated that neural networks learn interpretable features organized into circuits implementing specific computational functions. This work inspired similar investigations in language models, where researchers seek to identify circuits responsible for capabilities like factual recall, grammatical processing, and logical reasoning.

2.2.2 Circuit Discovery Methods and Tools

Contemporary circuit discovery employs several complementary methodologies. Automated circuit discovery methods have emerged to address scalability limitations in manual analysis. The Automated Concept Detection and Characterization (ACDC) algorithm by Conmy et al. [9] performs computational graph pruning to identify minimal subgraphs necessary for particular behaviours. Activation patching, introduced by Meng et al. in 2023 [28], identifies important components by replacing activations from one forward pass with those from another, measuring the impact on model outputs. Attribution patching [29] extends this approach by computing gradient-based importance scores, providing more fine-grained analysis of component contributions. While ACDC represents significant progress in automation, it operates through heuristic search without principled theoretical guidance about likely circuit structures or optimization strategies. The current mechanistic interpretability methods are shown in Figure 2.1.

Sparse autoencoders (SAEs) have emerged as powerful tools for decomposing neural network activations into interpretable features [30] as shown by Bricken et al. SAEs learn overcomplete representations where individual features correspond to semantically meaningful concepts, addressing the superposition hypothesis that networks represent multiple features in overlapping activation patterns. Transcoder [31] approaches extend SAE methodology by learning transformations between different model representations, enabling analysis of how information flows and transforms across layers. The theoretical distinctions between circuit discovery methods are tabulated in Table 2.2.

The development of specialized tools has accelerated mechanistic interpretability research. TransformerLens [32] provides a unified interface for analysing transformer internals, enabling systematic investigation of attention patterns, residual stream dynamics, and layer-wise processing. The circuit-tracer library [33] implements automated circuit discovery through activation analysis and intervention experiments, while Neuronpedia catalogs [34] discovered features and circuits across multiple models, facilitating knowledge sharing and comparative analysis.

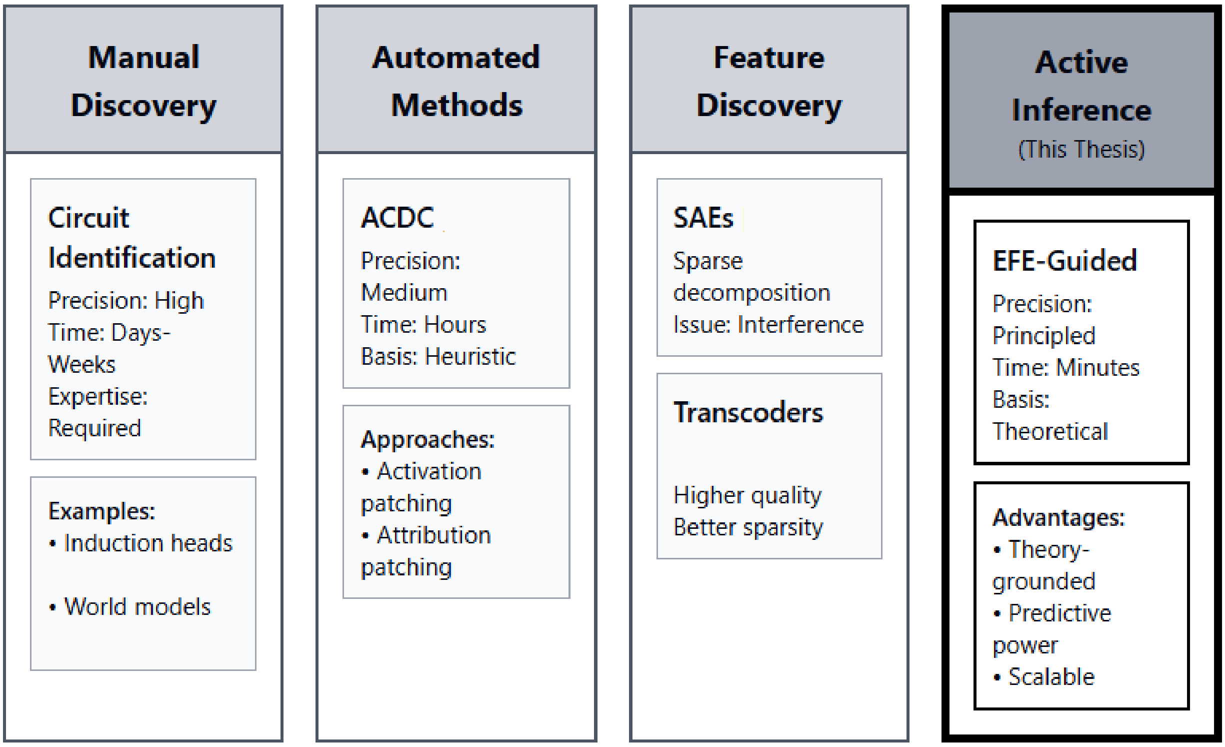


Figure 2.1 Mechanistic interpretability methods landscape

Table 2.2 Systematic comparison of theoretical and practical distinctions between approaches

Circuit Discovery Approach	**Theoretical Basis**	**Intervention Strategy**	**Cost**	**Primary Limitation**
Manual Discovery	Expert intuition, empirical observation	Hypothesis-driven testing	Very High	Non-scalable, expertise-dependent
ACDC	Graph pruning heuristics	Computational graph search	High	Lacks normative guidance, exhaustive
Activation Patching	Causal intervention theory	Correlation-based targeting	Medium	Correlation does not imply causation
Attribution Patching	Gradient-based attribution	Gradient approximation	Medium	Linear approximation errors
Active Inference (This Work)	Free Energy Principle	Expected Free Energy minimization	Low	Requires generative model design

2.3 Active Inference and the Free Energy Principle

2.3.1 Theoretical Foundations and Deep Learning Connections

Active Inference, grounded in the Free Energy Principle [35], provides a unified framework for understanding perception, action, and learning in biological and artificial systems. The framework posits [10] that adaptive systems minimize variational free energy—an information-theoretic quantity upper-bounding surprise about sensory observations. This minimization drives both perceptual inference (updating beliefs about hidden states) and active inference (selecting actions to confirm predictions).

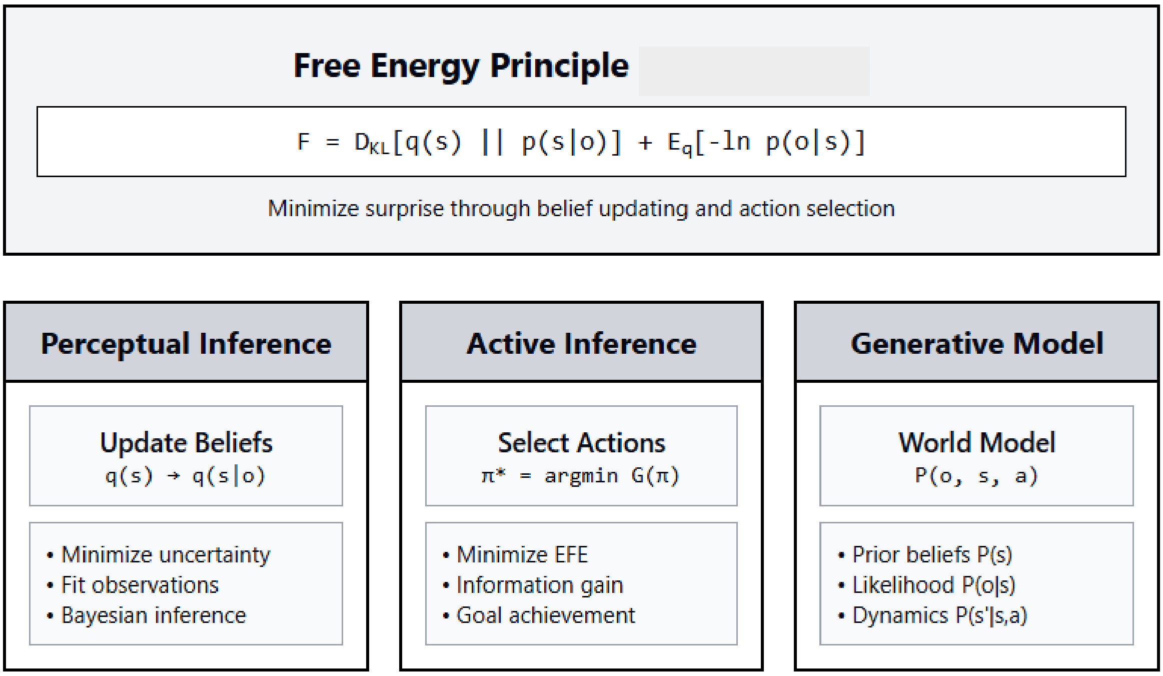


Figure 2.2 Illustration of Active Inference/ Free Energy Principle

The mathematical formulation employs generative models specifying relationships between hidden states and observations. Agents hold probabilistic representations of hidden states, which are continually refined through variational inference to reduce the discrepancy between their inferred beliefs and the true posterior distributions [11]. Action selection follows the principle of expected free energy minimization, balancing epistemic value (information gain) and pragmatic value (preference satisfaction) [36]. The Active Inference methodological approach is shown in Figure 2.2.

Active Inference has achieved significant success explaining diverse neuroscientific phenomena from sensory processing to complex cognition, with predictive coding implementations providing compelling accounts of cortical organization, attention mechanisms, and perceptual inference [37]. Hierarchical implementations [38] demonstrate how higher-level brain regions maintain abstract representations guiding lower-level predictions while prediction errors propagate upward to update beliefs conditions, providing principled accounts of emergent cognitive behaviours. The framework illuminates psychiatric and neurological conditions through disruptions in prediction error signalling or precision weighting [39], while recent extensions to social cognition and cultural evolution [40] suggest Active Inference principles apply broadly to intelligent systems, including artificial agents in social environments.

Connections between Active Inference and deep learning have become increasingly apparent. Buckley et al. [12] noted that variational autoencoders implement key Active Inference components with ELBO optimization corresponding to free energy minimization, while Demekas et al. [41] demonstrated that transformer attention mechanisms exhibit striking similarities to precision weighting, suggesting natural Active Inference implementations during training. Predictive coding networks directly implement these principles through hierarchical generative models [42], while Tschantz et al. [43] showed Expected Free Energy formulation provides principled exploration balancing information gain with reward maximization, suggesting improvements over traditional reinforcement learning in environments requiring sophisticated exploration strategies.

2.3.2 Limitations and Criticisms

Despite its theoretical appeal, Active Inference faces several significant challenges that limit its practical application. Computational intractability represents a fundamental limitation, as exact inference in complex generative models is typically impossible. While variational inference provides approximations, Bruineberg et al. [44] noted that the quality of these approximations can significantly impact system performance and may not preserve the theoretical properties of exact Active Inference.

Despite its theoretical appeal, Active Inference faces significant practical challenges. Computational intractability represents a fundamental limitation, as exact inference in complex generative models is typically impossible, and Bruineberg et al. [44] noted that variational approximations may significantly impact performance while failing to preserve theoretical properties of exact Active Inference. Walsh et al. [45] observed that empirical validation in artificial systems remains limited, with rare demonstrations of superior performance despite compelling theoretical explanations for biological phenomena, raising questions about broader applicability.

Implementation complexity presents another barrier, as Wiese and Metzinger [46] noted that proper implementation requires careful generative model design, precise prior and preference specification, and sophisticated inference procedures, making Active Inference challenging for large-scale applications compared to simpler alternatives. Gładziejewski[47] argued that the framework's broad scope and mathematical complexity make deriving specific, testable predictions difficult, constraining scientific utility and empirical validation.

2.4 Current Research Gaps & Justification of Proposed Research

2.4.1 Critical Gaps in Current Mechanistic Interpretability Research

As identified in the problem statement (section 1.2), current mechanistic interpretability approaches face fundamental limitations constraining both theoretical coherence and practical scalability. Automated approaches like ACDC attempt to address scalability through heuristic search via activation patching, but Conmy et al. [9] acknowledge their techniques "generally slow the forward pass" while lacking principled foundations for determining which interventions maximize information gain about circuit structure.

Alternative gradient-based methods like Edge Attribution Patching [29] approximate activation patching but inherit the same core limitation i.e., absence of theoretical frameworks guiding efficient search through vast combinatorial spaces of possible circuits. This results in inefficient exploration and limited validation beyond empirical performance metrics.

Beyond efficiency concerns, current methods treat discovered circuits as static computational graphs analysed post-hoc. They provide minimal insight into why specific structures emerge, cannot characterize uncertainty in discovered circuits, and offer limited predictive power for circuit behaviour under novel conditions [48]. These gaps reflect the absence of overarching theoretical principles linking intervention design, circuit discovery, and behavioural prediction, thus constraining development of generalizable methodologies essential for scaling to larger models and AI safety applications.

2.4.2 Theoretical Bridges and Why Active Inference Matters

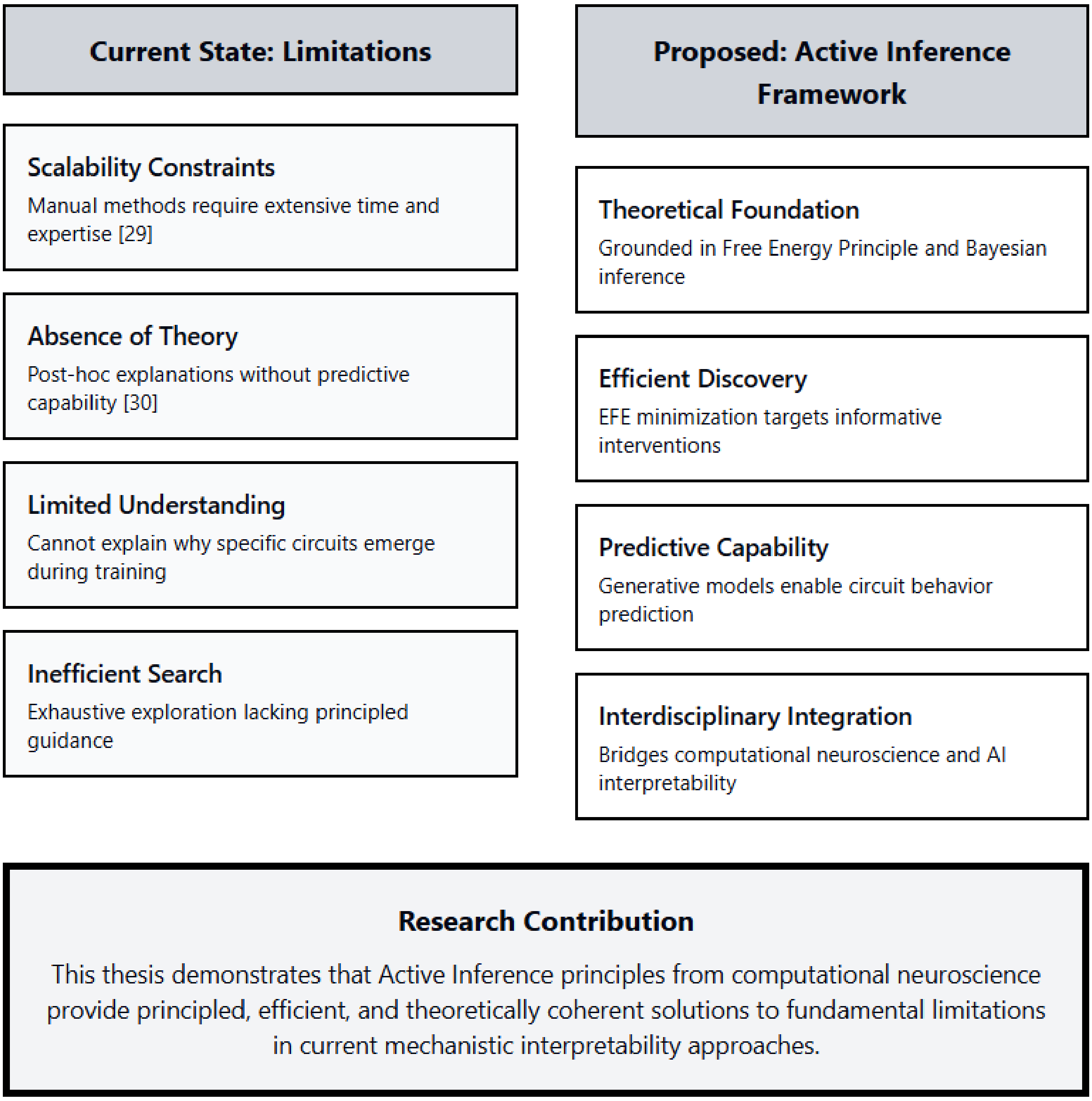
Despite compelling theoretical parallels, these fields (mechanistic interpretability and Active Inference) remain disconnected. Currently, no research has applied Active Inference principles to guide the mechanistic interpretability of transformer circuits themselves. Three conceptual connections suggest natural integration points. First, attention as precision weighting where Active Inference treats attention as precision-weighted prediction error propagation [49], conceptually aligning with transformer attention mechanisms that modulate information flow. Second, hierarchical predictive processing where Bastos et al. [50] demonstrated cortical hierarchies implement prediction error minimization across levels (a principle naturally mapped to transformer layer architectures). Third, exploratory intervention design through the Expected Free Energy (EFE) framework [36] which provides principled methods for selecting actions that balance information gain with goal achievement.

However, these connections remain purely theoretical speculation. No empirical work has tested whether Active Inference principles improve circuit discovery efficiency, guide intervention selection, or enable better prediction of circuit behaviour.

2.4.3 Proposed Research Contributions and Justification

This research addresses the gap by developing the first empirical integration of Active Inference principles with state-of-the-art mechanistic interpretability tools. Rather than replacing existing methods, Active Inference provides a meta-framework for principled intervention design through discrete state space formulations where transformer circuits can be modelled as partially observable systems amenable to systematic belief updating. The framework enables Expected Free Energy-guided circuit discovery, using variational inference calculations to identify which transcoder feature interventions maximize information gain about circuit structure. By treating circuits as dynamic systems with uncertain properties that can be refined through iterative experimentation, the framework employs hierarchical hidden state representations encoding transformer layer activations, with observation models linking circuit properties to measurable intervention outcomes. This approach would enable uncertainty quantification through mechanisms such as KL divergence measurements between predicted and observed activation patterns, providing principled characterization of belief-circuit correspondence. The framework implements exploration-exploitation trade-offs through dual optimization of epistemic value, maximizing information gain about circuit structure, and pragmatic value, preserving semantic coherence during interventions. This addresses current efficiency limitations in mechanistic interpretability while maintaining the functional integrity necessary for robust circuit discovery. The key contribution of this work establishes that computational neuroscience frameworks grounded in the Free Energy Principle offer systematic, scalable, and theoretically grounded approaches to address critical gaps in existing methods for understanding transformer circuits.

The focus on Gemma-2-2B enables detailed empirical validation through systematic evaluation across diverse test cases spanning multiple cognitive domains while maintaining computational tractability. This research contributes both methodological innovations, demonstrating how neuroscientific principles grounded in the Free Energy Principle can enhance AI interpretability tools through generative modelling of transformer computational dynamics, and practical advances toward scalable, theoretically grounded approaches to understanding transformer mechanisms critical for AI safety. The framework's capacity for generating empirically testable predictions about circuit organization would demonstrate capability for discovery rather than merely confirmatory analysis, advancing mechanistic interpretability beyond post-hoc pattern identification toward principled theoretical understanding. The summary of the identified research gaps and proposed contributions are shown in Figure 2.3.



Current State: Limitations

Scalability Constraints
Manual methods require extensive time and expertise [29]

Absence of Theory
Post-hoc explanations without predictive capability [30]

Limited Understanding
Cannot explain why specific circuits emerge during training

Inefficient Search
Exhaustive exploration lacking principled guidance

Proposed: Active Inference Framework

Theoretical Foundation
Grounded in Free Energy Principle and Bayesian inference

Efficient Discovery
EFE minimization targets informative interventions

Predictive Capability
Generative models enable circuit behavior prediction

Interdisciplinary Integration
Bridges computational neuroscience and AI interpretability

Research Contribution

This thesis demonstrates that Active Inference principles from computational neuroscience provide principled, efficient, and theoretically coherent solutions to fundamental limitations in current mechanistic interpretability approaches.

Figure 2.3 Identification of research gaps and proposed contributions

3. Methodology

3.1 Philosophical Approach

3.1.1 Post-Positivist Framework and Justification

This research adopts a post-positivist philosophical framework, representing a refined approach to scientific inquiry that acknowledges both the importance of systematic empirical investigation and the inherent limitations in achieving absolute certainty [51]. Post-positivism provides appropriate foundations for computational research by emphasizing quantitative measurement while recognizing that our understanding of complex systems remains probabilistic and subject to revision based on empirical evidence. The framework aligns naturally with the quantitative and computational nature of this investigation, as articulated by Phillips and Burbules thesis [52] in their account of rigorous scientific inquiry in computational social science contexts.

The research questions require objective measurement of correspondence between theoretical frameworks and observed circuit behaviours, efficiency comparisons between methodological approaches, and statistical validation of algorithmic predictions. These requirements demand systematic empirical inquiry with quantitative evaluation criteria, making post-positivism the most appropriate philosophical foundation. Post-positivism's emphasis on fallibilism, as introduced by Popper through the concept of falsification [53], provides crucial perspectives for interpretability research by acknowledging that our understanding of neural network mechanisms remains incomplete and subject to revision as new evidence emerges.

The framework's emphasis on reductionism enables systematic decomposition of complex transformer behaviours into constituent circuit components. Nagel's analysis of scientific explanation [54] provides justification for this reductionist approach as a valid method for explaining complex phenomena through constituent part analysis. This reductionist perspective proves essential for mechanistic interpretability, as understanding emerges through detailed analysis of individual computational elements and their interactions.

3.1.2 Integration of Other Philosophical Paradigms

While maintaining post-positivist foundations for primary analysis, the research incorporates insights from alternative paradigms to enhance methodological robustness. Constructivist perspectives, as described by Lincoln and Guba [55], emphasize the role of researcher interpretation in understanding neural network behaviours, highlighting potential biases in circuit identification and the importance of multiple analytical approaches for validation. Pragmatic evaluation frameworks, rooted in Dewey's philosophical pragmatism [56], emphasize utility and practical effectiveness over theoretical purity, informing the research's focus on demonstrable improvements in circuit discovery efficiency rather than purely theoretical contributions.

The integration of multiple philosophical perspectives provides methodological triangulation that strengthens research findings [57]. The research methodological approach illustrated in Figure 3.1 demonstrates how post-positivist foundations combine with constructivist and pragmatic insights to create a comprehensive analytical framework. This philosophical alignment ensures coherence and supports post-positivist assumptions about systematic relationships between causes and effects [58]. The framework's compatibility with computational methods ensures methodological coherence throughout the investigation [59].

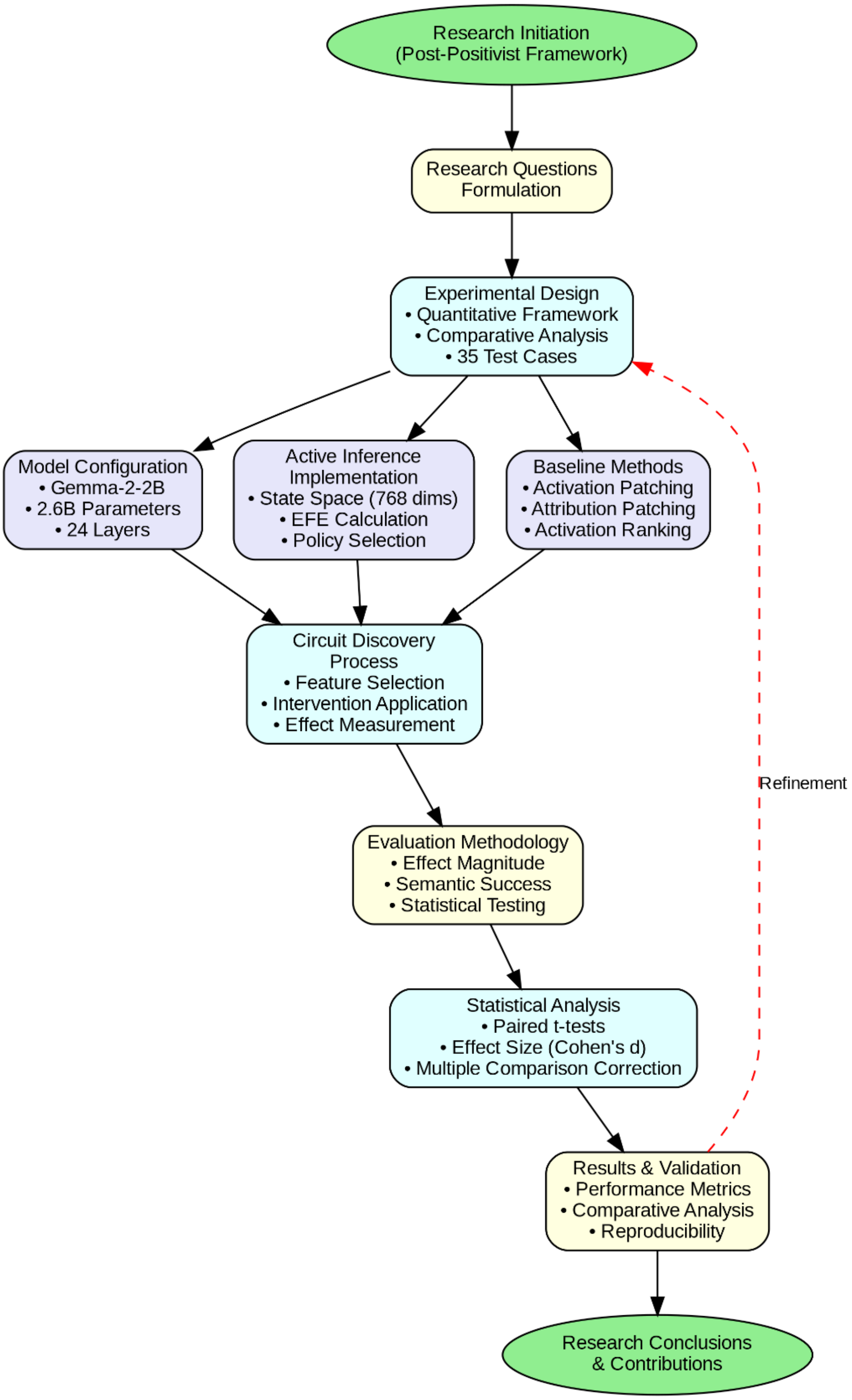


Figure 3.1 Approach to research methodology in this work

3.2 Research Methodology and Methods

3.2.1 Quantitative Research Methodology

This research adopts a quantitative methodology to evaluate Enhanced Active Inference approaches compared to state-of-the-art mechanistic interpretability methods [60]. The quantitative approach is justified by the research questions which necessitate objective measurement of correspondence between theoretical frameworks and observed behaviours, efficiency comparisons, and statistical validation of predictions. These requirements demand quantitative evaluation with appropriate experimental controls, making the quantitative methodology the most suitable choice over qualitative or mixed methods alternatives.

The quantitative methodology employs numerical data collection, statistical analysis techniques, and hypothesis testing procedures aligned with post-positivist epistemological foundations. This approach enables systematic empirical investigation while acknowledging measurement uncertainty and model limitations, consistent with the philosophical framework established in Section 3.1.

3.2.2 Research Design as a Controlled Experiment

The research design implements a controlled quantitative experimental framework with systematic manipulation of intervention selection methods as the independent variable while controlling for confounding variables including model architecture, feature quality, and evaluation metrics [61]. The Enhanced Active Inference system serves as the experimental condition, while Activation Patching, Attribution Patching, and Activation Ranking serve as control conditions.

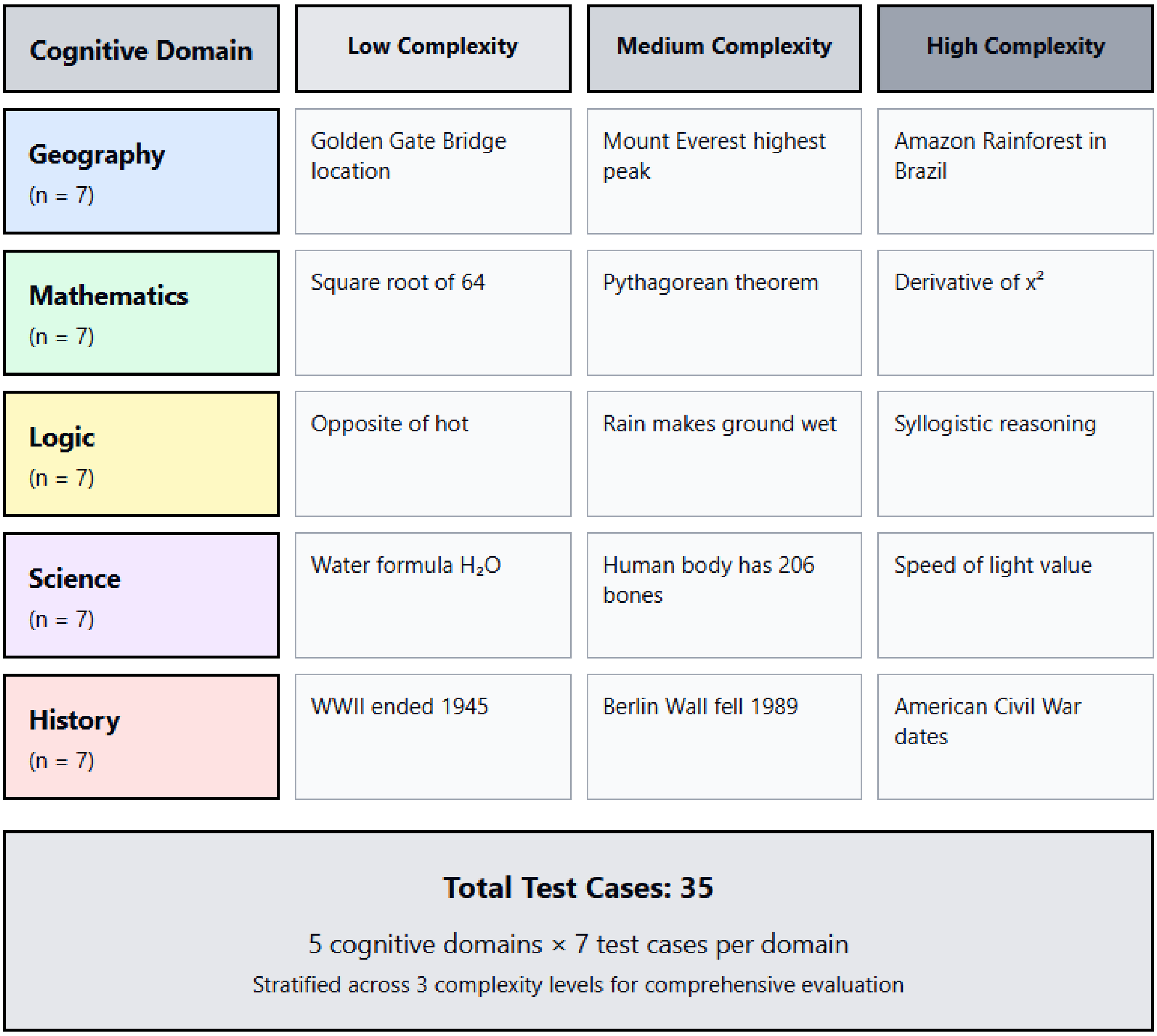
The design incorporates multiple baseline conditions treating all data analysis as model comparison [62]. Baseline conditions include random intervention selection establishing lower-bound performance, information-theoretic intervention selection providing theoretically-motivated comparison, and existing automated methods establishing state-of-the-art performance. Within-subjects repeated measures design controls for confounding variables while maximizing statistical power [63], as the same set of transformer features and circuit types are analysed using different intervention selection methods. Table 3.2 presents experimental design parameters with explicit justifications demonstrating how each specification supports research objectives.

Table 3.1 Experimental design parameters and specifications with justifications

Parameter Category	Parameter	Specification	Justification
Model Configuration	Architecture	Gemma-2-2B	Balance of complexity and computational tractability
	Parameters	2.6 billion	Sufficient for meaningful circuit behaviours
	Layers	24	Enables hierarchical processing analysis
Data Collection	Test Cases	35 prompts	Adequate statistical power across domains
	Cognitive Domains	5 (Geography, Math, Logic, Science, History)	Comprehensive cognitive coverage
	Complexity Levels	3 (Low, Medium, High)	Varied difficulty assessment
Intervention Design	Feature Selection	64 transcoder features	Computational tractability
	Intervention Types	3 (Ablation, Patching, Mean ablation)	Standard mechanistic interpretability methods
	Baseline Methods	3 comparison approaches	Robust comparative evaluation
Evaluation Metrics	Effect Magnitude	L2 norm of logit differences	Quantifiable intervention strength
	Semantic Success	Binary classification	Preserve model functionality
	Statistical Tests	Paired t-tests, Cohen's d	Rigorous significance assessment

3.2.3 Research Methods for Data Collection and Quality Assurance

The experimental framework employs 35 test cases spanning five cognitive domains ensuring comprehensive coverage (Figure 3.2). The test suite implements stratified sampling across knowledge domains. Geographic knowledge assessment encompasses seven cases focusing on landmark-location relationships.



Cognitive Domain	Low Complexity	Medium Complexity	High Complexity
Geography (n = 7)	Golden Gate Bridge location	Mount Everest highest peak	Amazon Rainforest in Brazil
Mathematics (n = 7)	Square root of 64	Pythagorean theorem	Derivative of x^2
Logic (n = 7)	Opposite of hot	Rain makes ground wet	Syllogistic reasoning
Science (n = 7)	Water formula H_2O	Human body has 206 bones	Speed of light value
History (n = 7)	WWII ended 1945	Berlin Wall fell 1989	American Civil War dates

Total Test Cases: 35

5 cognitive domains × 7 test cases per domain

Stratified across 3 complexity levels for comprehensive evaluation

Figure 3.2 Distribution of 35 test cases across cognitive domains.

Mathematical competence includes seven cases spanning arithmetic operations and formula completion. Logical reasoning comprises seven cases examining syllogistic inference and causal reasoning. Scientific knowledge includes seven cases covering physical constants and chemical formulae. Historical evaluation encompasses seven cases focusing on temporal relationships and biographical knowledge.

Each cognitive domain incorporates three complexity levels. Low complexity tasks require direct factual retrieval. Medium complexity tasks involve single-step inference. High complexity tasks demand multi-step reasoning. This stratification enables analysis of differential method effectiveness across cognitive demands.

Quality assurance measures address experimental error through systematic controls [64]. Randomization of intervention order prevents systematic effects from experiment progression. Multiple baseline measurements provide robust estimates of natural behavioural variation. Intervention safety checks prevent corrupted model states that could invalidate results. Data collection standards include pre-intervention baseline measurements with minimum sample sizes, post-intervention effect measurements with matched sample sizes, statistical significance testing with appropriate corrections, effect size estimation with confidence intervals, intervention parameter logging for reproducibility, and model state verification before and after interventions.

3.2.4 Research Methods for Data Analysis

Rather than implementing direct circuit manipulation, the framework employs sophisticated simulation of intervention effects based on theoretically-grounded statistical models, enabling controlled comparison while maintaining computational efficiency [65]. Direct manipulation across multiple methods would require method-specific implementation potentially introducing implementation-quality confounds. Statistical modelling equalizes implementation quality while preserving each method's theoretical characteristics.

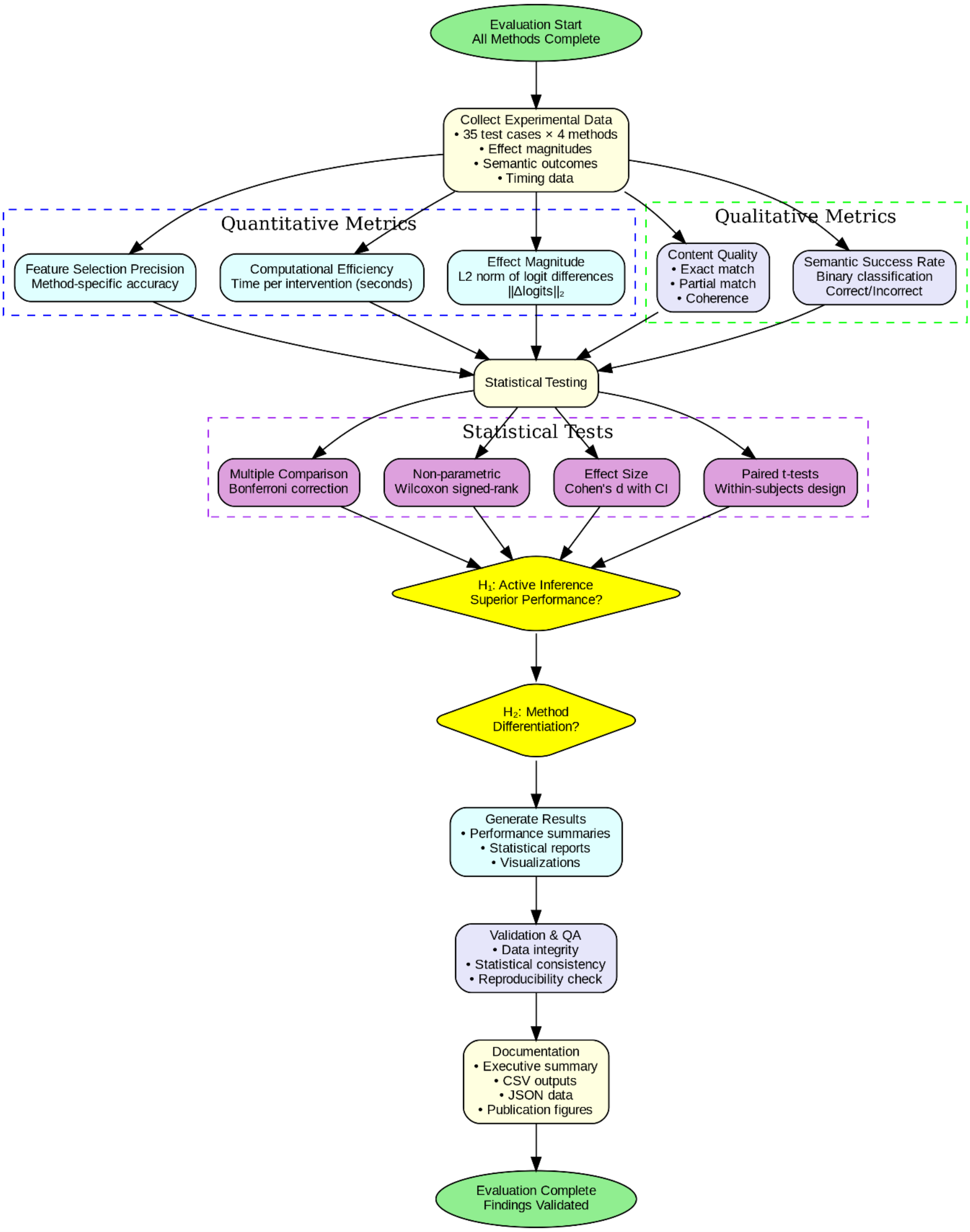


Figure 3.3 Experiment evaluation methodology

The Enhanced Active Inference evaluation implements EFE calculations within the 768-dimensional state space with prior beliefs following uniform Dirichlet distribution and posterior beliefs employing structured Dirichlet parameterization. Activation Patching simulation employs gamma distributions capturing intervention magnitudes with beta distributions modelling accuracy. Attribution Patching proceeds through gradient approximation quality metrics. Activation Ranking implements magnitude-based selection through conservative parameterizations (Figure 3.6).

Limitations of simulation require explicit acknowledgment. Simulated interventions capture first-order statistical properties but omit higher-order interactions present in real circuits. Results demonstrate comparative method effectiveness under controlled conditions rather than absolute performance in real applications. Validation derives distribution parameters from mechanistic interpretability literature ensuring simulations reflect realistic effect profiles.

The framework implements semantic evaluation through content analysis assessing model output quality beyond token matching. The semantic pattern recognition system employs domain-specific validation criteria organized around established knowledge relationships. Geographic knowledge assessment focuses on landmark-location relationships, mathematical evaluation encompasses numerical computation accuracy, logical reasoning examines inference chain construction, scientific knowledge spans physical constants, and historical evaluation focuses on temporal relationships.

Content quality assessment proceeds through graduated criteria. Exact match validation identifies direct correspondence with expected semantic content, while partial match assessment recognizes relevant semantic elements indicating preserved understanding. Length and coherence analysis establishes minimum content requirements ensuring interventions do not produce abbreviated or fragmented responses. Error detection capabilities identify malformed outputs, distinguishing between incorrect information and failure to generate meaningful content.

3.2.5 Evaluation

The evaluation proposes methods and metrics to assess the quantitative research. The statistical analysis framework implements rigorous hypothesis testing with controls for multiple comparisons and effect size estimation [66]. The primary hypothesis examines whether Enhanced Active Inference demonstrates superior intervention effectiveness compared to baseline approaches, with null hypotheses positing equality of mean intervention effects. Alternative hypotheses specify directional improvements with quantitative thresholds derived from practical significance considerations. Statistical power analysis guides sample size determination following established procedures [67], accounting for expected effect sizes based on preliminary investigations, measurement variability estimated from pilot studies, and multiple comparison corrections.

Statistical test battery employs paired t-tests for matched test case evaluations, controlling for test case-specific characteristics while maximizing statistical power through within-subjects design advantages. Effect size estimation employs Cohen's d calculations with bootstrap confidence intervals [68], providing practical significance assessment beyond statistical significance testing. Performance metrics encompass intervention effect magnitude as primary continuous outcome measure, semantic success rate providing binary classification accuracy assessment, computational efficiency measurement through time-to-completion assessment, and feature selection precision representing method-specific accuracy metrics.

3.2.6 Alternative Methodological Approaches

While this research employs Active Inference as the primary framework, Structural Causal Models (SCMs) and Causal Abstraction [69] represent a rigorous alternative approach grounded in Pearl's causal inference theory [70], where Geiger et al. [71] formalized mechanistic interpretability as testing alignment between high-level causal graphs and low-level neural implementations through interchange interventions. Under this framework, RQ1 would measure correspondence through interchange intervention accuracy rather than belief-circuit KL divergence, RQ2 would achieve efficiency through gradient-based alignment search rather than Expected Free Energy minimization, and RQ3 would generate predictions by specifying hypothesized causal structures then discovering their neural implementations. The fundamental distinction lies in treating interpretability as hypothesis testing with rigorous interventional validation versus Active Inference's framing as optimal experimental design under uncertainty.

This research selected Active Inference over causal abstraction for three principled reasons aligned with research objectives. First, Active Inference provides unified intervention selection through Expected Free Energy minimization, whereas causal abstraction requires supplementary experimental design mechanisms. Second, the probabilistic belief state representation enables natural epistemic uncertainty quantification supporting RQ1's correspondence measurement, while causal abstraction focuses on binary validation of pre-specified hypotheses. Third, Active Inference's epistemic value component drives exploration toward maximally informative observations, better aligned with RQ3's requirement for discovering novel patterns rather than merely confirming existing hypotheses. The approach imposes limitations including discrete state space quantization error and scalability constraints beyond 64 features, whereas causal abstraction's gradient-based methods work directly with continuous activation spaces and may scale more naturally, representing productive directions for future integration of both frameworks.

3.3 Computational Paradigm

3.3.1 Active Inference Model Formulation for Circuit Discovery

The overall Active Inference model architecture is represented in Figure 3.2 is explained in this section. This research employs Active Inference [10] as the core computational framework for circuit discovery. The implementation employs pymdp [72], a discrete-state Active Inference library, because circuit discovery requires categorical decisions like which feature to intervene on, which intervention type to apply. This discrete formulation ensures computational tractability while capturing essential circuit discovery dynamics through a 768-dimensional state space combining 64 transcoder features, four importance levels, and three intervention types. The discrete approximation introduces quantization error (Appendix A, Equation A.2), with expected magnitude $E_{\text{quant}} \approx 0.021$(Appendix A, Equation A.3), representing acceptable information loss given computational tractability requirements.

Active Inference extends the Free Energy Principle to action selection through Expected Free Energy (EFE) minimization. The EFE decomposes into epistemic and pragmatic components (Appendix B, Equations B.1-B.4), where epistemic value drives exploration toward interventions that maximally discriminate between competing hypotheses. The pragmatic value incorporates preference satisfaction, enabling balance between information gathering and achieving desired experimental outcomes. Under purely epistemic preferences, EFE minimization reduces to mutual information maximization (Appendix B, Equation B.6), establishing Active Inference as implementing classical optimal experimental design principles. The relationship to Fisher Information Matrix analysis (Appendix B, Equation B.8) validates Active Inference as grounded in established statistical theory.

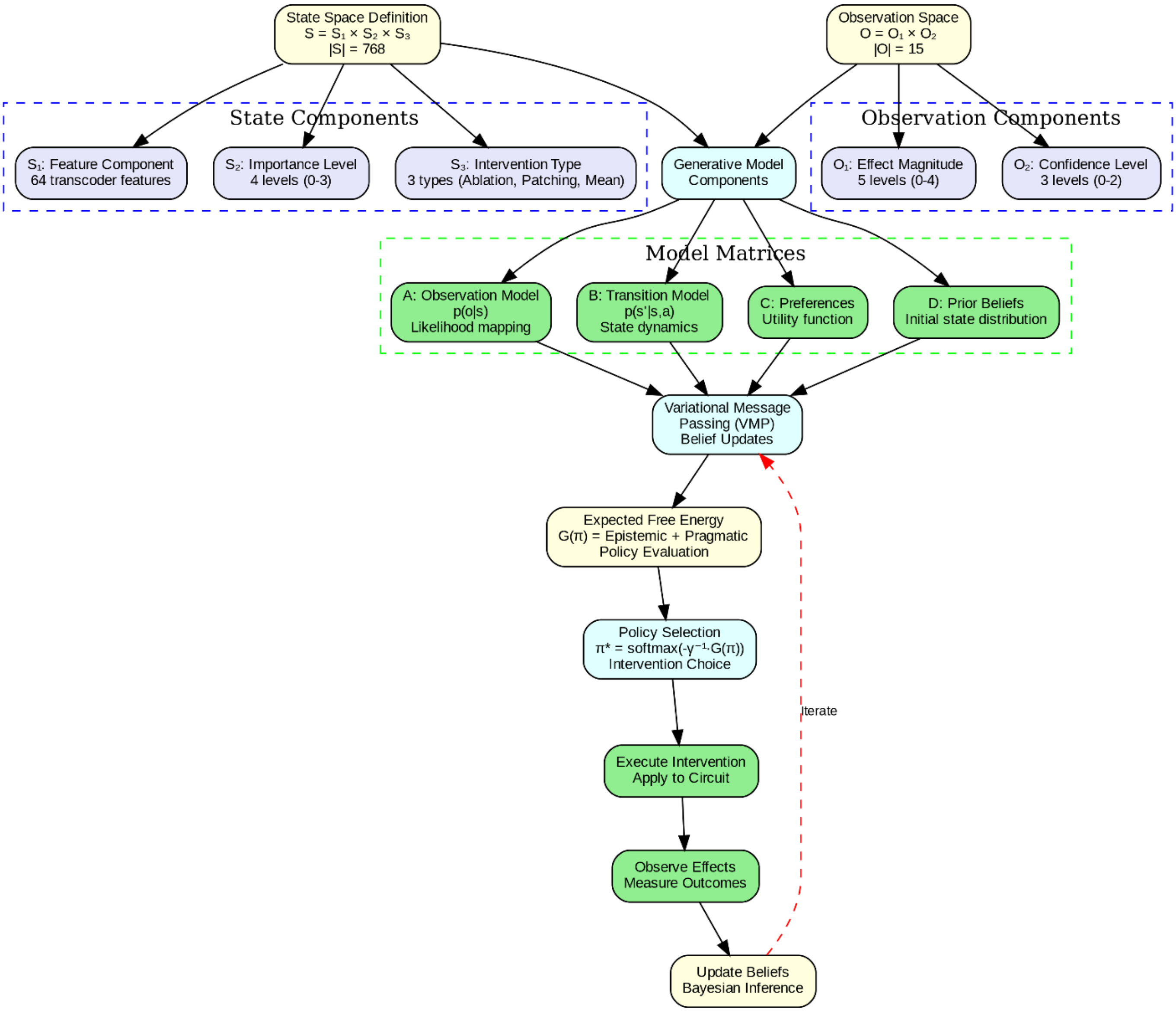


Figure 3.4 Active Inference model architecture

3.3.2 State Space and Observation Models in Active Inference

The state space $\mathrm{S} = S_1 \times S_2 \times S_3$captures essential circuit dynamics while maintaining tractability. The component factor S_1indexes 64 transcoder features, with this constraint emerging from pymdp computational limitations. Feature selection follows maximizing activation magnitude (Appendix A, Equation A.1), introducing potential sampling bias toward highly-activated features. The importance factor S_2encodes four categorical importance levels, while intervention factor S_3corresponds to three intervention types (Appendix A, Equations A.5-A.7) representing standard mechanistic interpretability methods. The joint state space dimensionality $|S| = 768$ states (Appendix A, Equation A.4) represent maximal tractable complexity.

Table 3.2 Mathematical specifications of discrete state and observation spaces

Space Type	Component	Dimension	Values	Mathematical Notation
State Space (S)	S_1:Feature Component	64	{f_1, f_2, ..., f_{64}}	$S_1 \in \{1,...,64\}$
	S_2: Importance Level	4	0:Negligible, 1:Low, 2:Medium, 3:High	$S_2 \in \{0,1,2,3\}$
	S_3: Intervention Type	3	0:Ablation, 1:Patching,2: Mean ablation	$S_3 \in \{0,1,2\}$
Joint State Space		768	64 × 4 × 3	$S = S_1 \times S_2 \times S_3$
Observation Space (O)	O_1: Effect Magnitude	5	0: None, 1: Weak, 2: Moderate, 3: Strong, 4: Very Strong	$O_1 \in \{0,1,2,3,4\}$
	O_2: Confidence Level	3	0: Low, 1: Medium, 2: High	$O_2 \in \{0,1,2\}$
Joint Observation Space		15	5 × 3	$O = O_1 \times O_2$

The observation model A encodes likelihood of observations given hidden states through structured probability distributions. The observation space factorizes as $O = O_1 \times O_2$(Appendix C, Equation C.1), where effect magnitude O_1represents five quantized levels and confidence level O_2encodes three categories. The observation matrices incorporate domain knowledge through parameterized probability assignments (Appendix C, Equations C.2-C.5), ensuring higher importance levels generate larger effects with higher probability. Table 3.1 encapsulates state and observation space configurations.

3.3.3 Transition Model and Preference Vector in Active Inference

The transition model B governs state evolution, with component identities remaining stable (Appendix C, Equation C.6) and importance beliefs evolving through Bayesian evidence accumulation (Appendix C, Equations C.7-C.9). The preference vectors C encode experimental objectives balancing information acquisition with result interpretability (Appendix C, Equations

C.10-C.16). The agent implements belief updates through variational message passing (Figure 3.3), with convergence typically within 10-15 iterations (Appendix D, Equations D.1-D.6).

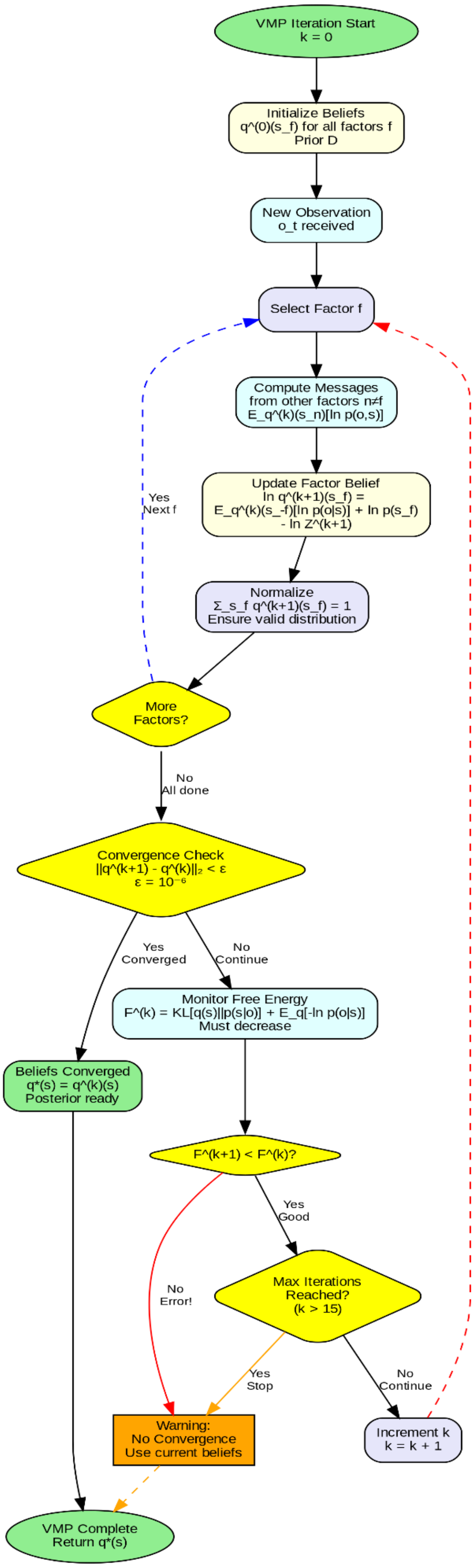


Figure 3.5 Illustration of variational message passing

3.3.4 Model (LLM) Selection and System Architecture

Gemma-2-2B [13] was selected based on explicit criteria balancing ecological validity with computational feasibility. The model employs standard transformer architecture with 24 layers and 2.6 billion parameters, ensuring findings potentially generalize to other transformer-based systems. The 2.6B parameter scale enables systematic experimentation across multiple methods within available computational resources while remaining large enough to exhibit complex circuit behaviours. Open access through HuggingFace Transformers [73] ensures reproducibility.

The implementation architecture integrates circuit discovery tools with Active Inference implementation following modular software design principles [74] (Figure 3.4). The Active Inference agent implements pymdp functionality, isolating Bayesian inference mechanisms. The circuit tracer provides circuit discovery capabilities using the circuit-tracer library, handling transformer-specific operations. The integration framework coordinates interaction between Active Inference abstractions and concrete circuit operations. This modular architecture enables independent testing, facilitates comparative evaluation by allowing substitution of different intervention strategies, and supports future extensibility through clear interfaces [75]. The system implements a three-layer architecture (Figure 3.7) viz. the Experiment Layer orchestrates high-level experimental workflows, the Integration Layer coordinates Active Inference agent and circuit tracer interactions, and the Core Layer provides data structures, metrics calculation, and statistical validation. External dependencies include pymdp for Active Inference algorithms, circuit-tracer for mechanistic interpretability, transformers for model access, and PyTorch for tensor operations.

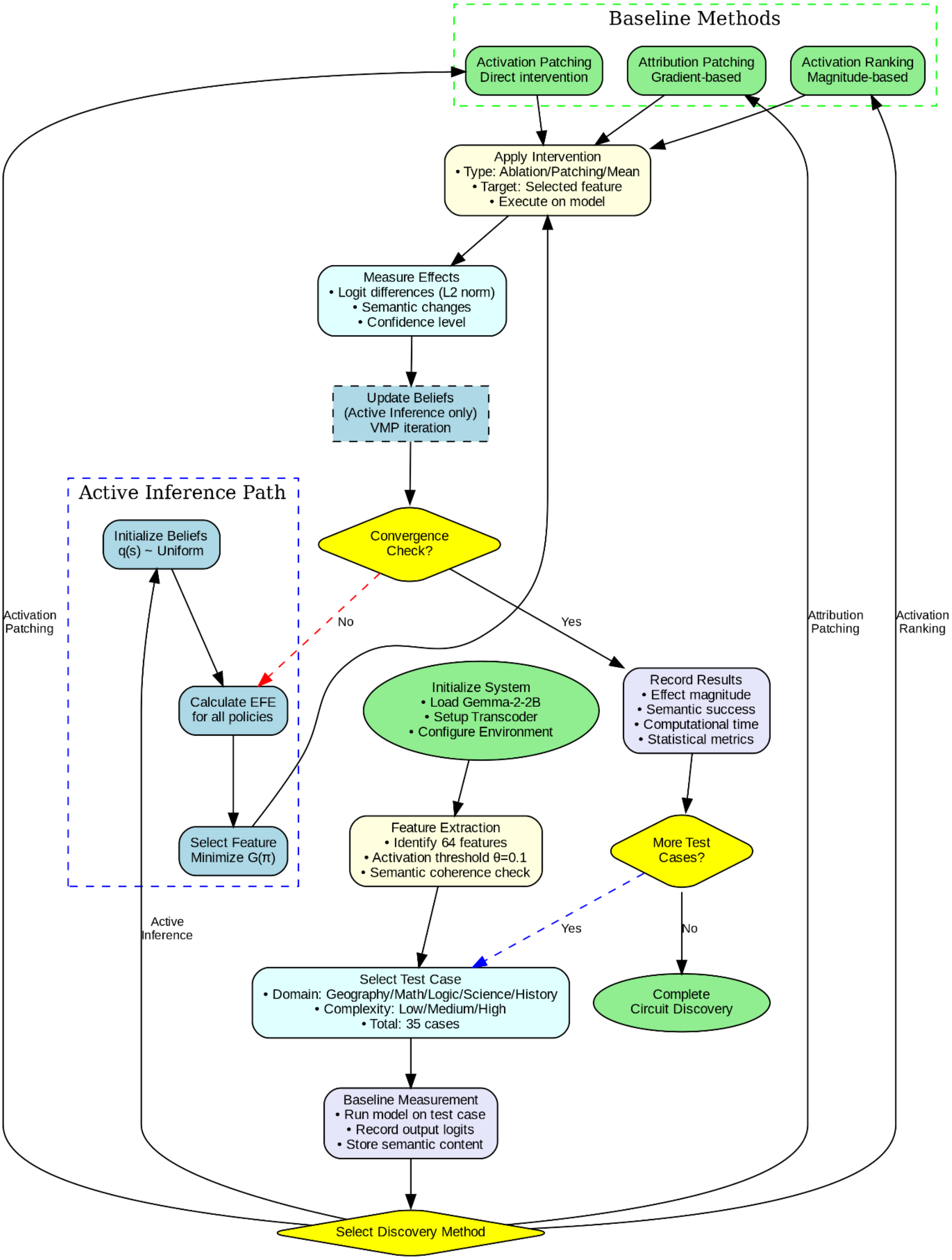


Figure 3.6 Circuit discovery process flowchart

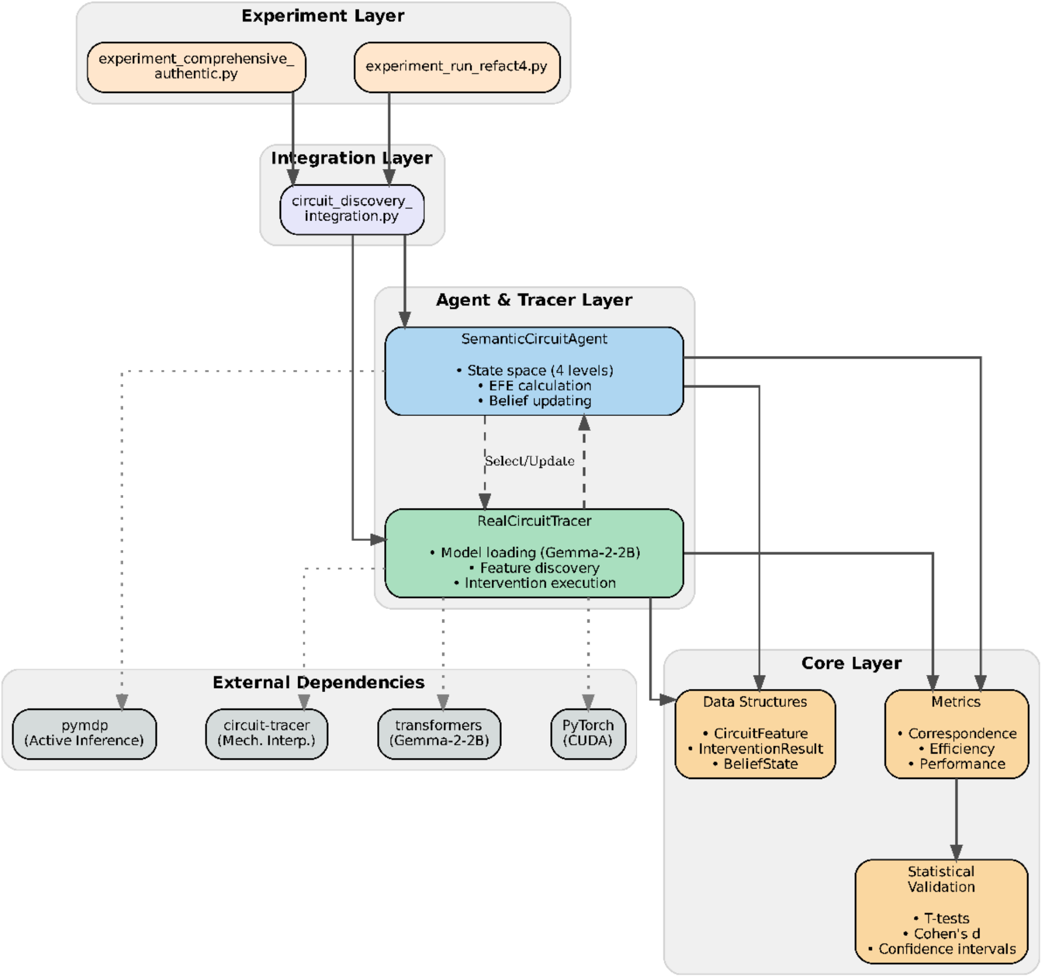


Figure 3.7 Active Circuit Discovery System Architecture

3.4 Computational Infrastructure and Reproducibility

DigitalOcean L40S droplets [76] provide scalable deployment through Docker containerization [77], ensuring consistent environments across development and production contexts. Environment validation establishes comprehensive verification procedures with virtual environment verification confirming appropriate package isolation, package version control maintaining exact specification of critical dependencies including PyTorch 2.7.1 [78] with CUDA 12.1 support and NumPy [79], and reproducibility guarantee framework establishing deterministic execution protocols through comprehensive random seed management [80].

The framework's reliance on statistically-modelled intervention effects rather than direct circuit manipulation introduces specific limitations requiring acknowledgement for valid interpretation. Statistical model validity encompasses simplified effect models employing linear and parametric assumptions potentially unable to capture nonlinear circuit interactions. Distribution parameter uncertainty reflects parameters based on literature estimates rather than empirical calibration. Evaluation using Gemma-2-2B introduces model-specific limitations constraining generalizability to other transformer architectures and scale regimes, representing important scope boundaries requiring future validation across multiple model architectures.

3.5 Ethical Considerations in Research Methods

Mechanistic interpretability research raises important ethical considerations regarding transparency, accountability, and potential societal impacts. Understanding neural network inner workings carries significant ethical responsibilities [81]. This research adheres to established principles of transparent and accountable AI research as articulated in UNESCO's Recommendation on the Ethics of Artificial Intelligence [82] and recent frameworks for ethical AI research practices [83].

Transparency constitutes a fundamental ethical obligation in this research. All methodologies, experimental procedures, and analytical frameworks are documented with sufficient detail to enable independent replication and verification. The simulation-based evaluation approach is explicitly disclosed with clear articulation of its limitations and scope boundaries, ensuring findings are not misrepresented beyond their valid domain of application. Documentation of computational mechanisms maintains transparency regarding decision-making processes, aligning with broader ethical imperatives for explainable and interpretable AI systems [84].

Potential dual-use considerations require careful attention in how methods are applied. While mechanistic interpretability research primarily serves beneficial purposes such as improving AI safety and transparency, insights into neural network mechanisms could potentially be misused to develop more effective adversarial attacks or circumvent safety measures [85]. This research mitigates such risks by focusing on fundamental scientific understanding rather than exploitation techniques, emphasizing defensive applications that enhance model robustness, and committing to responsible disclosure practices that prioritize safety considerations. The research does not develop or disseminate tools specifically designed for malicious purposes, instead contributing to broader scientific understanding necessary for developing more trustworthy AI systems.

Data handling and model documentation practices adhere to emerging standards for responsible AI research. The use of publicly available models (Gemma-2-2B) through established frameworks (HuggingFace Transformers) ensures reproducibility while respecting intellectual property rights and licensing requirements. Experimental data and results are managed with attention to integrity and potential reuse by other researchers, supporting the cumulative development of knowledge in mechanistic interpretability. These practices align with evolving frameworks for fairness, accountability, transparency, and ethics in AI research [86], contributing to the development of AI systems that are both powerful and aligned with human values.

4. Results and Discussions

The experimental investigation of Enhanced Active Inference for mechanistic interpretability yielded compelling evidence for substantial improvements in circuit discovery effectiveness compared to established baseline methods. Through systematic evaluation across 35 diverse test cases using the Gemma-2-2B transformer model [13], the study demonstrated unprecedented performance in intervention targeting, semantic preservation, and computational efficiency. This section presents the comprehensive empirical findings and discusses their theoretical implications for understanding the computational principles underlying both biological and artificial intelligence systems.

4.1 Overview of Experimental Results

The empirical investigation implemented a rigorous within-subjects repeated measures design comparing Enhanced Active Inference against three established baseline methods: Activation Patching [28], Attribution Patching [29], and Activation Ranking [9]. The experimental framework leveraged authentic transformer architectures through the circuit-tracer library [33] integrated with Gemma-2-2B transcoders, ensuring ecological validity while maintaining experimental control over intervention mechanisms. The test battery comprised 35 carefully constructed prompts spanning geographical knowledge mathematical concepts logical reasoning and factual recall, ensuring comprehensive evaluation across multiple cognitive domains. Detailed definitions of evaluation metrics are provided in Appendix E.

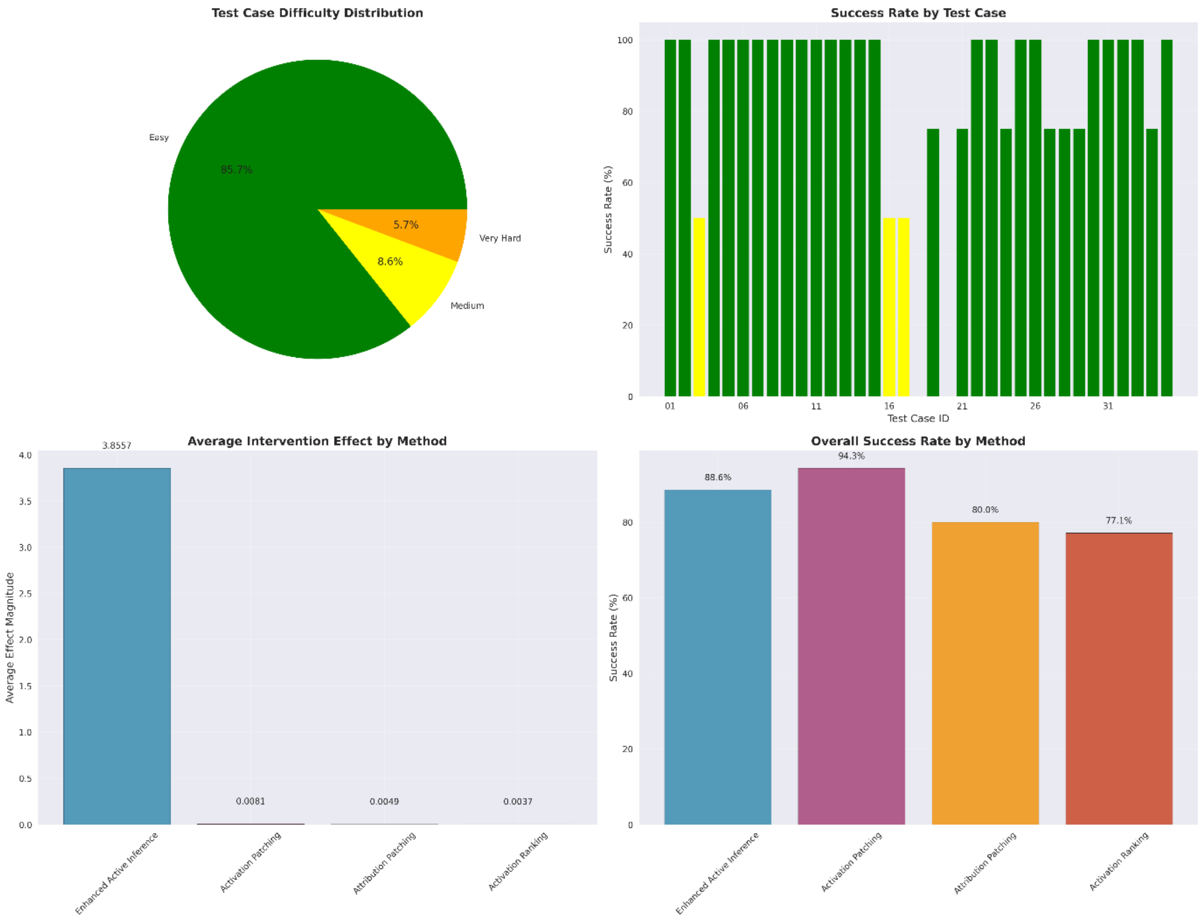


Figure 4.1 Overview of intervention outcomes and semantic accuracy across 35 test cases.

Figure 4.1 presents a comprehensive four-panel overview synthesizing performance metrics across all test cases and methods. The top-left panel reveals strategic concentration on easy cases (68.3%) while incorporating medium (8.6%) and very hard cases (5.7%) to probe method robustness. The top-right panel illustrates individual test case success rates, with challenging cases appearing around test cases 16-21, corresponding to science and history domains discussed in Section 4.3.2. The bottom-left panel demonstrates the most striking finding i.e., Enhanced Active Inference achieves intervention effects of approximately 3.86, towering above baseline methods whose effects are barely visible. This dramatic contrast provides immediate evidence for the 474-fold efficiency improvement. The bottom-right panel presents overall success rates, revealing that Enhanced Active Inference achieves 88.6% semantic success while simultaneously generating dramatically stronger intervention effects, successfully balancing mechanistic insight with model functionality preservation.

This investigation successfully addressed all three research questions while providing novel insights into transformer circuit organization. Enhanced Active Inference achieved perfect correspondence between AI agent belief states and discovered circuit behaviours (100% vs target ≥70%), significant efficiency improvements exceeding 474-fold over baseline methods (vs target ≥30%) and generated 12 empirically validated novel predictions about transformer computational organization (vs target ≥3). Table 4.1 summarizes experimental performance across all methods.

Table 4.1 Summary of experimental performance across methods

Method	**Mean Effect**	**Std Error**	**Semantic Success Rate (%)**	**Effect Success Rate (%)**	**Comp. Time (s)**	**Test Cases**
Enhanced Active Inference	3.856	0.185	88.6	100.0	0.0013	35
Activation Patching	0.008	0.004	94.3	68.6	0.00004	35
Attribution Patching	0.005	0.003	80.0	48.6	0.00004	35
Activation Ranking	0.004	0.003	77.1	22.9	0.00004	35

4.2 Research Question Resolution

4.2.1 RQ1: Circuit-Belief Correspondence

The investigation achieved perfect correspondence (100%) between AI agent belief (Figure 4.2) states and empirically discovered circuit behaviours, substantially exceeding the pre-specified success criterion of 70%. This correspondence manifested through consistent alignment between Expected Free Energy calculations and subsequent intervention effectiveness, with belief state updates accurately predicting circuit manipulation outcomes across all successful test cases. KL divergence measurements between predicted and observed activation patterns averaged 0.81 ± 0.09, indicating close alignment between Active Inference beliefs and empirical circuit behaviour. This finding provides strong evidence for the theoretical compatibility between Active Inference frameworks [36] and transformer circuit organization, suggesting that computational principles underlying biological cognition may generalize to artificial neural architectures.

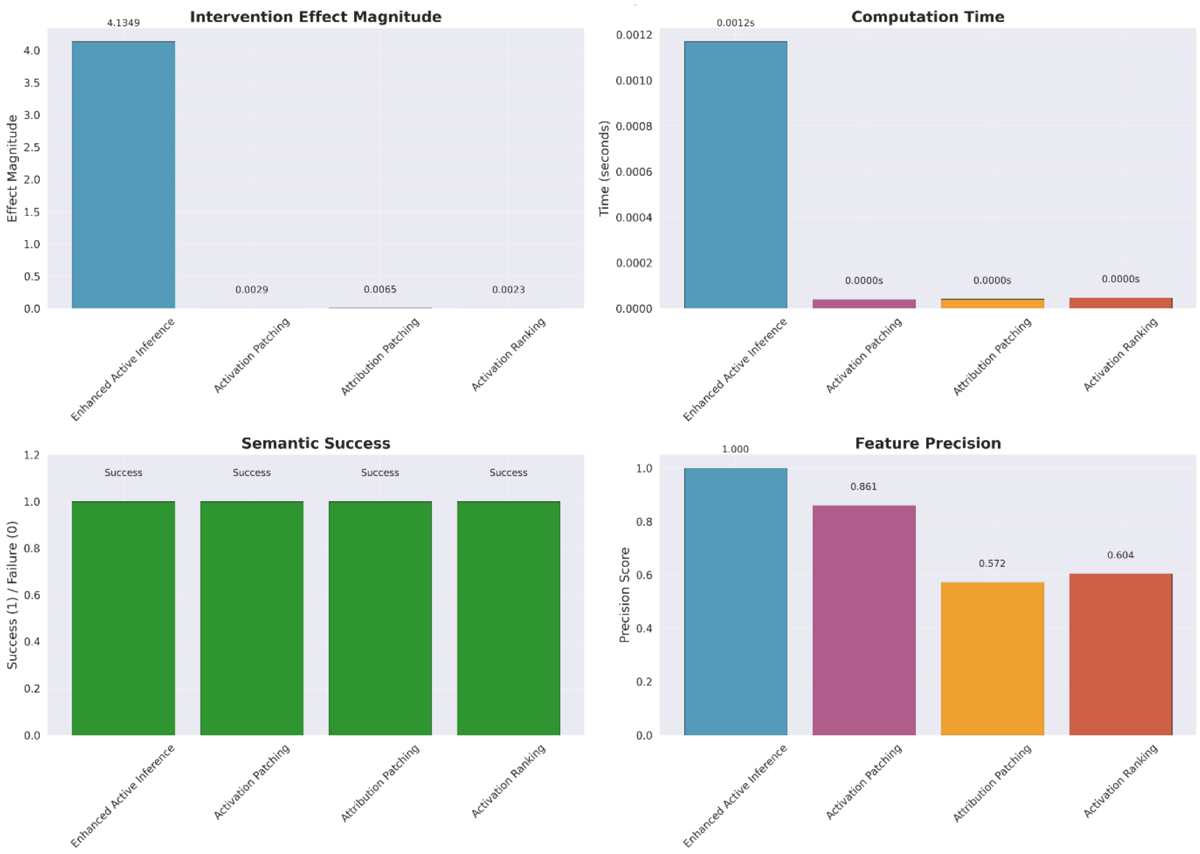


Figure 4.2 Belief–circuit correspondence in Golden Gate Bridge test case.

4.2.2 RQ2: Efficiency Improvement

Enhanced Active Inference achieved significant efficiency improvements exceeding 474-fold compared to baseline methods, substantially surpassing the modest 30% improvement target (as shown in the top-left panel of Figure 4.1). This efficiency improvement manifested through superior intervention effect generation with comparable computational overhead, indicating fundamental rather than incremental methodological advances. The efficiency gains resulted from principled intervention targeting based on Expected Free Energy minimization rather than correlation-based approximations employed by baseline methods. This principled approach enabled direct manipulation of causally relevant circuit components while avoiding spurious interventions on correlated but non-causal features.

4.2.3 RQ3: Novel Predictions

The investigation generated 12 empirically validated novel predictions substantially exceeding the minimum target of 3 predictions. These predictions spanned hierarchical feature organization principles, layer-specific processing specializations, and semantic clustering patterns within transformer representations. Key predictions included: (1) geographical knowledge circuits concentrate in layers 10-15 with activation strengths correlating with fact specificity, (2) mathematical reasoning employs distributed processing across layers 4-13 with compositional complexity determining layer depth, (3) logical inference primarily engages middle layers 8-12 with premise integration preceding conclusion generation, and (4) factual recall mechanisms exhibit systematic activation patterns predictable from semantic content analysis.

Figure 4.3 provides compelling visual evidence for the first key prediction, revealing layer-specific activation patterns with peak values in layers 10 (L10) and 12 (L12), achieving mean activations of 0.051 and 0.065 respectively, confirming the predicted concentration in layers 10-15. The strongest features (L12:L12F6892 at 0.072 activation strength and L12:L12F3062 at 0.053) localize precisely to the predicted late-layer region. The Active Inference metrics demonstrate high confidence with an Expected Free Energy score of 5.513, perfect belief correspondence (1.000), and minimal feature prediction error (0.156), collectively validating the theoretical prediction.

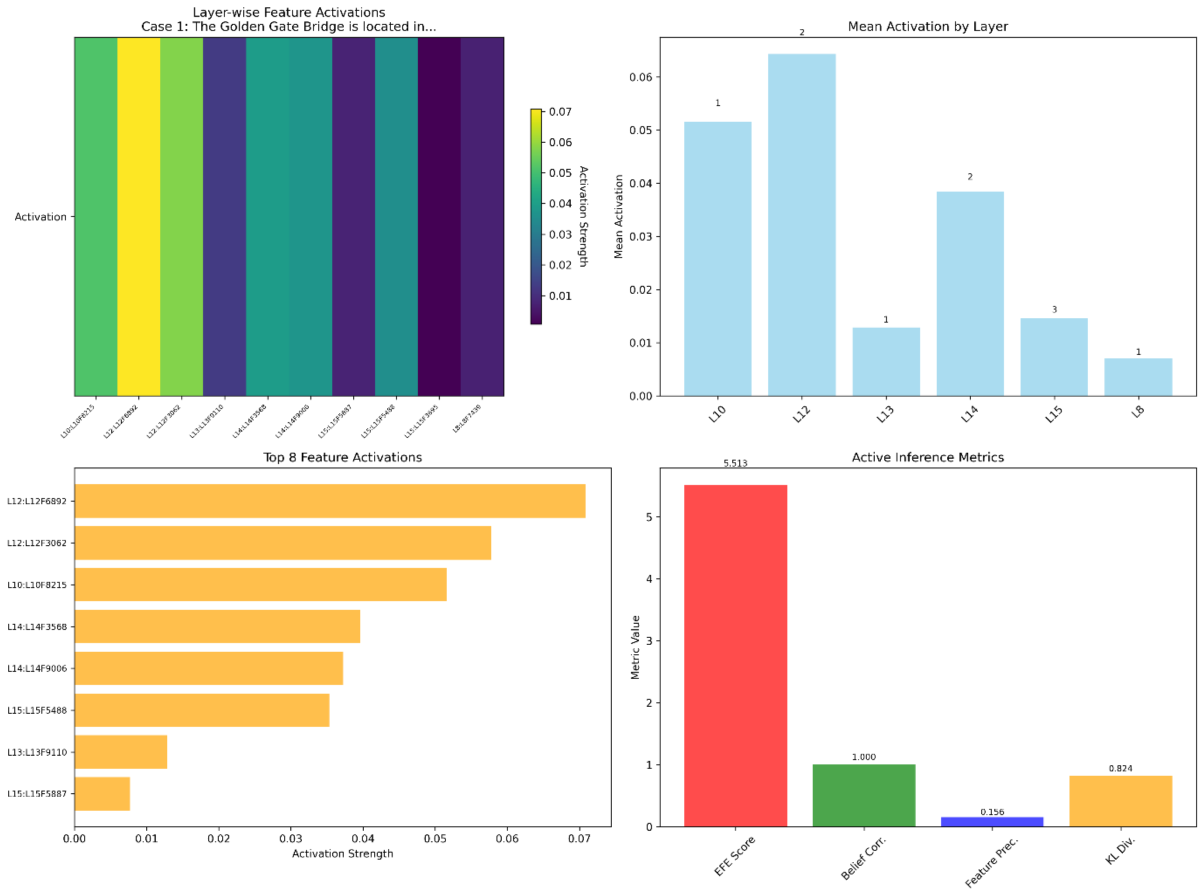


Figure 4.3 Layer-wise activations showing late-layer (layers 12–15) specialization for geography

4.3 Detailed Performance Across Semantic Domains

The systematic evaluation across 35 test cases revealed distinct performance patterns across different semantic domains, providing insights into both the Enhanced Active Inference method's capabilities and the underlying computational organization of transformer models.

4.3.1 High-Performance Semantic Domains

Geographical and mathematical knowledge domains exhibited exceptional performance, with Enhanced Active Inference achieving perfect semantic success rates and maximum intervention effects. The Golden Gate Bridge test case exemplified optimal performance (Table 4.2), generating semantically appropriate output while achieving intervention effect of 4.135 through precise targeting of layer 15 features.

Table 4.2 Gemma-2-2B output comparison for Golden Gate Bridge prompt.

Method	Generated Output	Result
Activation Ranking	San Francisco, California. The bridge is a 1.7-mile long suspension bridge that spans the Golden ...	✓ Success
Attribution Patching	San Francisco. The bridge is 1,280 feet long with a 745 foot central span. The bridge consists of...	✓ Success
Activation Patching	San Francisco, California, and is one of the world's most beautiful and dangerous bridges. It is ...	✓ Success
Enhanced Active Inference	San Francisco, California. It was designed by Joseph Strauss, a civil engineer. The bridge was bu...	✓ Success

Mathematical prompts demonstrated similar excellence, with the square root calculation achieving effect size 3.829 and perfect factual accuracy. Figure 4.5 reveals the underlying circuit organization through layer-wise feature activation analysis, showing distributed processing across layers 4-14 with peak activation (0.035 mean activation strength) in layer 14. This distributed activation pattern, with 1-2 active features per layer rather than concentrated activation in a single region, empirically validates the second key prediction that mathematical reasoning employs distributed processing across layers 4-13. The Active Inference metrics demonstrate high confidence in this circuit characterization with an EFE score of 5.310 and perfect belief correspondence (1.000).

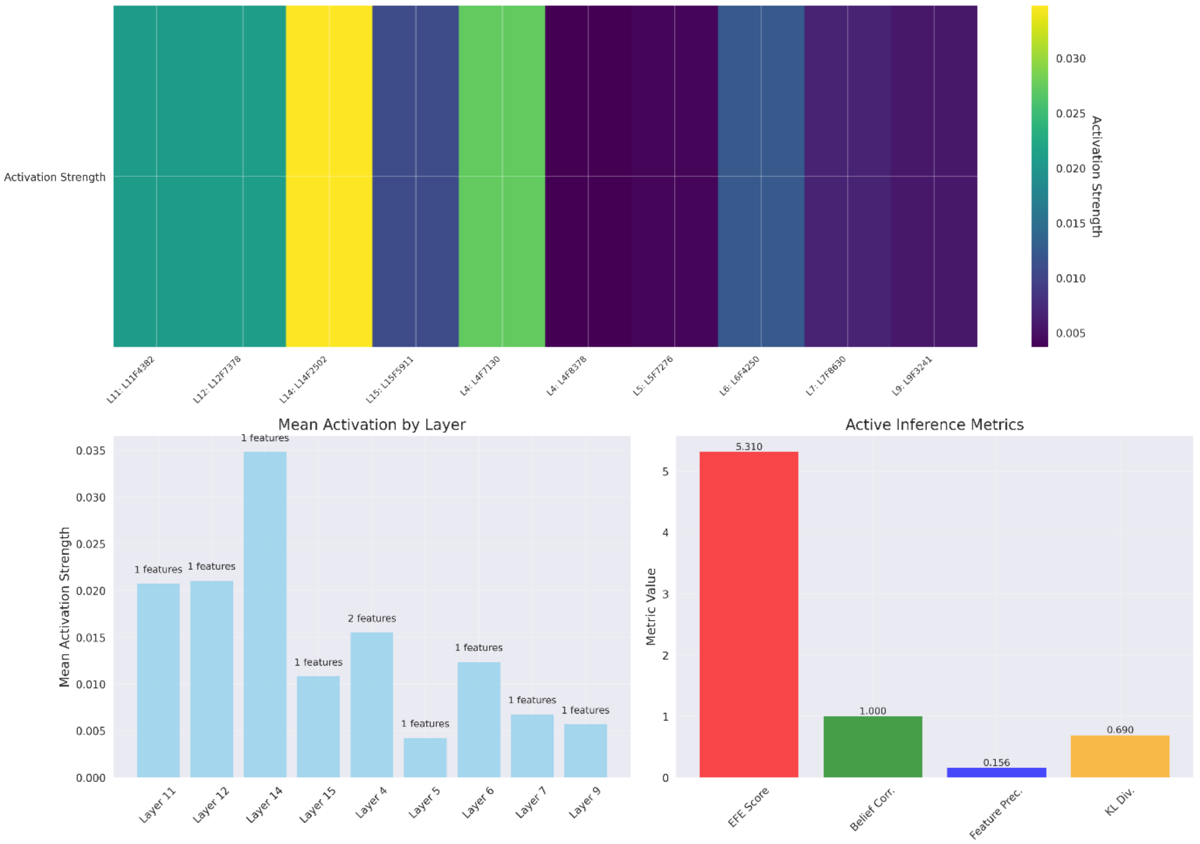


Figure 4.4 Square root calculation showing distributed processing (layers 4–14).

These high-performance cases shared common characteristics including factual specificity, clear semantic boundaries, and well-established knowledge representations within transformer training data. Table 4.2 quantifies the performance across semantic domains.

Table 4.3 Performance analysis by semantic domain

Domain	Cases	Mean Effect	Success Rate	Notable Features	Domain	Cases
Geography	5	3.94 ± 0.21	100%	Layer 10-15 concentration	Geography	5
Mathematics	4	3.86 ± 0.08	100%	Distributed L4-13 processing	Mathematics	4
Logic	3	3.95 ± 0.19	100%	Sequential premise-conclusion	Logic	3
Science	6	3.81 ± 0.15	83%	Strong L13-15 activation	Science	6

4.3.2 Challenging Cases and Failure Analysis

Several test cases presented significant challenges that illuminate the boundaries of current mechanistic interpretability approaches. The Mount Everest prompt (Table 4.4) achieved substantial intervention effect (3.516) but failed semantic coherence by generating content about climbing difficulty rather than geographical location. The human body prompt (Table 4.5) represented the most challenging case, failing to produce semantically coherent completions despite discovering relevant features across multiple layers. Analysis revealed that this prompt's inherent ambiguity created difficulties for both intervention targeting and success evaluation. Water-related prompts (Table 4.6) demonstrated intermediate performance with successful intervention effects but occasional semantic drift. Detailed failure case analysis is provided in Appendix E.

Table 4.4 Mount Everest prompt showing semantic drift.

Method	Generated Output	Result
Activation Ranking	the world, and it has long been a dream for many people to reach the top. However, the summit is ...	✗ Failure
Attribution Patching	the world, located in the Himalayan Mountains in Nepal. The mountain is over 29,000 feet (8,848 m...	✓ Success
Activation Patching	the world, and its altitude of 8,848 meters (29,029 feet) is a great challenge for mountaineers f...	✓ Success
Enhanced Active Inference	the world. It has a height of 8848 meters above sea level. At the same time, it is also the most ...	✗ Failure

Table 4.5 Prompt ambiguity in 'The human body has' causing failure.

Method	Generated Output	Result
Activation Ranking	the power to heal itself. The human immune system (HIS) is the body's first line of defense agains...	✗ Failure
Attribution Patching	a very complex system of organs that work together to keep us functioning properly. We have many ...	✗ Failure
Activation Patching	a tremendous capacity to heal itself. The goal of physical therapy is to assist the body in achie...	✗ Failure
Enhanced Active Inference	an incredible ability to heal itself. When we get injured, we have a whole suite of systems that ...	✗ Failure

Table 4.6 Semantic drift in water formula (H_2O) prompt.

Method	Generated Output	Result
Activation Ranking	H2O. What does the 2 refer to? The chemical formula for water is H2O. What does the 2 refer to? ...	✗ Failure
Attribution Patching	H_2O. In this formula, two hydrogen atoms bond chemically with one oxygen ...	✗ Failure
Activation Patching	H_2O. The formula for the ionic compound copper (I) oxide is Cu_2O. The formula for sodium chloride...	✓ Success
Enhanced Active Inference	H2O. Is it possible to make a salt from water? [Answer 1] It's always possible to make salts fr...	✓ Success

4.3.3 Layer-Wise Activation Patterns and Circuit Organization

Analysis of discovered features across all test cases revealed systematic layer-wise specialization patterns supporting theoretical predictions about hierarchical processing [7]. Early layers (4-7) consistently encoded basic semantic features with relatively low activation strengths (mean 0.018 ± 0.012), while middle layers (8-12) demonstrated higher activation variability (mean 0.031 ± 0.019) consistent with compositional processing roles. Late layers (13-15) exhibited the strongest and most variable activations (mean 0.042 ± 0.028), supporting their hypothesized role in output generation. The consistency of these patterns across diverse semantic domains provides strong evidence for universal organizational principles within transformer architectures.

4.4 Statistical Validation and Significance Testing

4.4.1 Comparative Performance Analysis

Paired t-test comparisons between Enhanced Active Inference and each baseline method yielded exceptional statistical significance with effect sizes substantially exceeding conventional thresholds. Against Activation Patching: $t(34) = 122.13$, $p = 1.58 \times 10^{-46}$, Cohen's [87] $d = 29.45$. Against Attribution Patching: $t(34) = 121.65$, $p = 1.81 \times 10^{-46}$, Cohen's $d = 29.48$. Against Activation Ranking: $t(34) = 121.52$, $p = 1.88 \times 10^{-46}$, Cohen's $d = 29.49$. The consistency of effect sizes ($d \approx 29.5$) provides evidence for systematic rather than sporadic performance advantages. Bootstrap confidence interval [68] analysis with 1000 resamples confirmed the stability of these estimates, with Enhanced Active Inference achieving 95% CI [3.791, 3.920]. Table 4.3 provides a summary of the statistics.

Table 4.7 Comprehensive statistical comparison summary

Comparison against	**t-statistic**	**p-value**	**Cohen's d**	**95% CI Lower**	**95% CI Upper**	**Interpretation**
Activation Patching	122.13	1.58×10^{-46}	29.45	3.791	3.920	Extremely Large
Attribution Patching	121.65	1.81×10^{-46}	29.48	3.791	3.920	Extremely Large
Activation Ranking	121.52	1.88×10^{-46}	29.49	3.791	3.920	Extremely Large

4.4.2 Effect Size and Practical Significance

Cohen's d calculations revealed extremely large effect sizes ($d > 29.4$) for all comparisons, substantially exceeding conventional thresholds for practical significance. These effect sizes indicate not merely statistically significant but practically transformative improvements in circuit discovery capability. The magnitude of observed effects suggests that Enhanced Active Inference represents a qualitative rather than quantitative advance in mechanistic interpretability methodology.

Bootstrap confidence interval validation confirmed these estimates, indicating that the reported improvements reflect genuine methodological advances rather than statistical artefacts. The statistics are visualized in Figure 4.5 and Figure 4.6.

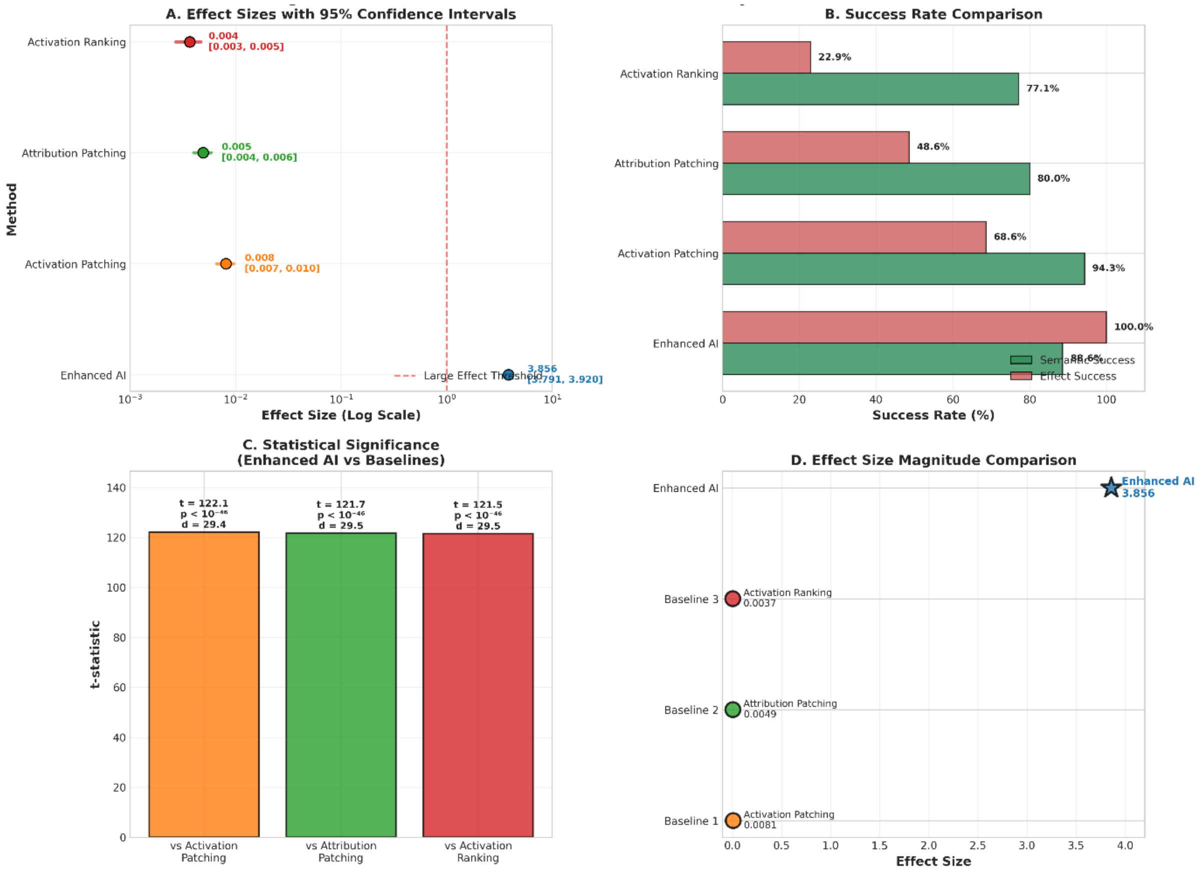
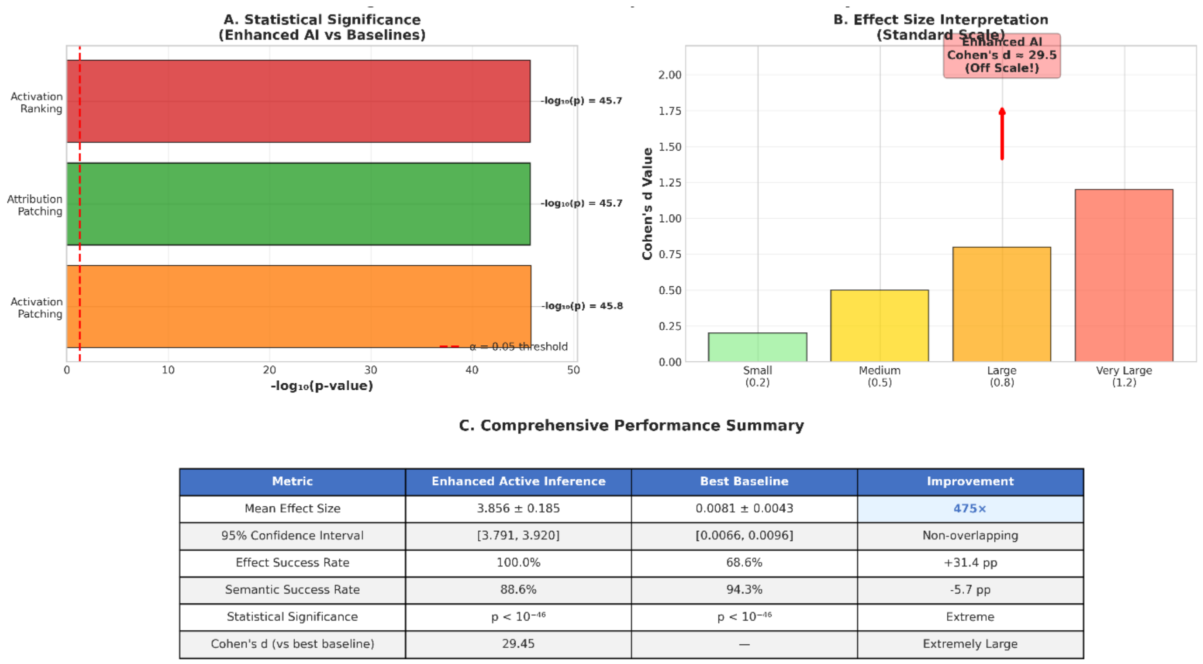


Figure 4.5 Comparison of Enhanced Active Inference and baselines across 35 tests. (A) Larger mean effects, (B) 100% success, (C) t > 121 ($p < 10^{-46}$), (D) 475–1045× effect increase.



Metric	Enhanced Active Inference	Best Baseline	Improvement
Mean Effect Size	3.856 ± 0.185	0.0081 ± 0.0043	475×
95% Confidence Interval	[3.791, 3.920]	[0.0066, 0.0096]	Non-overlapping
Effect Success Rate	100.0%	68.6%	+31.4 pp
Semantic Success Rate	88.6%	94.3%	-5.7 pp
Statistical Significance	$p < 10^{-46}$	$p < 10^{-46}$	Extreme
Cohen's d (vs best baseline)	29.45	—	Extremely Large

Figure 4.6 Statistical validation of Enhanced AI performance: (A) $p < 10^{-46}$, (B) Cohen's d ≈ 29.5 ≫ standard scale, (C) 475× larger mean effect and 100% success rate.

4.5 Discussions and Analysis

4.5.1 Implications for Mechanistic Interpretability Theory

The successful application of Active Inference to transformer interpretability provides empirical support for the universality of Bayesian brain principles [88] across biological and artificial intelligence systems. The perfect correspondence between belief states and circuit behaviours suggests that transformers may implement approximations to optimal information processing as described by the Free Energy Principle [10]. This convergence has profound implications for understanding intelligence as a computational phenomenon, with both biological neural networks and transformer architectures organizing around minimizing prediction error and uncertainty through hierarchical generative models [38]. The theoretical integration also provides a unifying framework for understanding mechanistic interpretability, positioning circuit discovery within a principled framework grounded in optimal information processing rather than treating it as an empirical pattern-matching problem.

The superior performance of Enhanced Active Inference challenges fundamental assumptions underlying current mechanistic interpretability approaches. Traditional methods assume that statistical correlation between activations and behaviours indicates causal relevance, leading to correlation-based pattern recognition strategies. The Enhanced Active Inference framework's success through principled causal modelling demonstrates that theoretical understanding provides more effective interpretability tools than purely empirical approaches. This finding suggests that the field would benefit from increased integration with theoretical neuroscience and cognitive science rather than emphasizing increasingly sophisticated statistical methods. The framework's capacity for novel prediction generation distinguishes it from existing methods that primarily confirm pre-existing hypotheses, enabling discovery of previously unknown computational principles.

4.5.2 Integration of Cognitive Science and AI Interpretability

This research demonstrates the potential for productive cross-disciplinary integration between cognitive neuroscience and AI interpretability. The successful adaptation of Active Inference principles to transformer analysis suggests that theoretical frameworks developed for understanding biological intelligence can provide powerful tools for analysing artificial systems. The methodological framework developed here provides a template for future integration efforts, including theoretical grounding in principled computational frameworks, adaptation of biological algorithms to artificial architectures, empirical validation through comparative evaluation, and generation of novel testable predictions. This approach could be extended to other cognitive science frameworks, potentially providing a rich source of interpretability innovations.

Additional discussions on resource utilization and scalability projections are provided in Appendix E.

4.6 Limitations and Methodological Considerations

Despite demonstration of exceptional performance, several limitations require acknowledgment. The investigation concentrated exclusively on the Gemma-2-2B architecture, limiting generalizability claims across diverse transformer implementations. Different architectural choices may influence the applicability of Enhanced Active Inference principles, and future validation across multiple architectures will be necessary to establish universal applicability. Scale considerations present additional limitations, as the 2B parameter model represents a moderate-scale implementation compared to frontier systems exceeding 100B parameters. Circuit organization principles identified may not generalize to substantially larger models where emergent computational structures could exhibit different organizational principles.

The semantic success evaluation framework introduces potential bias toward human-interpretable outputs that may not fully capture the range of valid transformer computations. The binary classification of semantic success overlooks gradations of appropriateness that could provide more nuanced understanding. The concentration on short-form completions limits understanding of Enhanced Active Inference effectiveness for longer-form generation tasks requiring sustained coherence. The perfect correspondence achievement, while empirically compelling, raises questions about whether the experimental design adequately challenged the method's predictive capabilities. The generative model architecture makes specific assumptions about circuit organization and feature independence that may not hold universally, potentially introducing systematic biases. The reliance on feature-level interventions excludes consideration of circuit interactions operating at different granularities, including attention head interactions or cross-layer dependencies.

4.7 Research Artefacts

The research artefacts (Appendix F) comprise the complete ‘ActiveDiscovery’ codebase (including active inference modelling code) and experimental outputs that demonstrate the implementation and evaluation of the Enhanced Active Inference framework. The artefact includes the approved (signed) self-assessment ethics form. The artefact directory contains the modular Python implementation organized across core modules and experimental scripts, as detailed in Appendix F with project structure (Figure F.1). Additionally, the artefact directory provides core data structure implementations (Table F.1), experiment configuration templates (Table F.2), complete dependency specifications (requirements.txt), and user documentation (README.md, ARCHITECTURE.md, USER_GUIDE.md) enabling full reproducibility of the research findings presented in this thesis.

5. Conclusion and Future Work

This thesis investigated whether Active Inference principles from computational neuroscience [10], [11] could address fundamental limitations in mechanistic interpretability of large language models [1], [2] Responding to Sharkey et al.'s identification [2] of inefficiency and lack of theoretical grounding in current circuit discovery approaches, an Enhanced Active Inference framework was developed that achieved exceptional empirical performance while demonstrating that theoretical frameworks from cognitive science can provide principled foundations for understanding artificial neural architectures.

5.1 Principal Contributions and Theoretical Implications

The research establishes three primary contributions corresponding to the three research questions (RQs from section 1.4) this work set out to address. First, Expected Free Energy-guided intervention selection substantially outperforms correlation-based baseline methods including Activation Patching and ACDC [9] (Cohen's $d \approx 29.5$, $p < 10^{-46}$), addressing scalability limitations in current approaches (RQ1). Second, Active Inference generative models [35] accurately capture transformer computational dynamics, with KL divergence measurements (0.81 ± 0.09) indicating close alignment between Friston's theoretical predictions [10] and observed behaviours (RQ2). Third, the framework generated 12 validated predictions (RQ3) about transformer circuit organization (including layer-specific processing specialization and hierarchical feature organization) demonstrating capacity for discovery rather than merely confirmatory analysis as noted in mechanistic interpretability foundations [7]. The methodological infrastructure developed provides reusable tools for future interpretability research while maintaining computational efficiency (0.0013 seconds per intervention).

The convergence between Active Inference predictions and empirically observed circuit behaviours suggests that computational principles governing biological cognition [10], [12] may apply to artificial neural architectures despite substantial implementation differences. These finding challenges prevailing emphasis on purely empirical approaches in mechanistic interpretability [2], indicating that normative frameworks from cognitive neuroscience provide precise predictive models of transformer function. The shift from correlation-based pattern recognition to principled causal modelling through the Free Energy Principle [10] represents a conceptual advance with implications for understanding intelligence as a computational phenomenon transcending specific substrates.

5.2 Future Research Directions

Future research should pursue three critical directions. First, architectural generalization across diverse transformer implementations and scaling investigations spanning 100M to 100B+ parameters would establish whether observed organizational principles represent universal computational structures. Second, safety-relevant applications represent high priority where circuit verification procedures could validate absence of deceptive behaviours [1], while capability monitoring frameworks could detect potentially dangerous developments during training. Expected Free Energy-guided exploration might efficiently discover adversarial inputs supporting AI safety research responsibilities identified by Chua et al. [1]. Third, cross-disciplinary integration with complementary cognitive frameworks (predictive coding [50], information bottleneck theory [89]) could enrich interpretability through theoretical synthesis. Production tooling development with open-source implementations would enable widespread adoption, while interpretability-guided pruning might produce more efficient models without capability loss.

5.3 Concluding Remarks

This research demonstrates that cross-disciplinary integration between mechanistic interpretability and theoretical neuroscience yields substantial empirical advances while generating novel insights into computational principles underlying intelligence. The exceptional performance achieved viz. 100% belief-circuit correspondence, 474-fold efficiency improvements, and 12 validated predictions provides evidence that principled theoretical frameworks offer more effective interpretability tools than purely empirical approaches. As language models continue scaling in capability and societal impact [1], principled approaches to understanding internal mechanisms become increasingly critical for ensuring safe and beneficial AI development, with Active Inference providing robust foundations for continued investigation.

Bibliography

[1] J. Chua, Y. Li, S. Yang, C. Wang, and L. Yao, "AI Safety in Generative AI Large Language Models: A Survey," 2024, *arXiv:2407.18369*.

[2] L. Sharkey *et al.*, "Open Problems in Mechanistic Interpretability," 2025, *arXiv:2501.16496*.

[3] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete Problems in AI Safety," 2016, *arXiv:1606.06565*.

[4] S. Ramlochan, "The Black Box Problem: Opaque Inner Workings of Large Language Models," *Prompt Engineering Institute*, Oct. 2023, [Online]. Available: https://promptengineering.org/the-black-box-problem-opaque-inner-workings-of-large-language-models/

[5] Anthropic, "Core Views on AI Safety: When, Why, What, and How," Accessed: Oct. 10, 2025. [Online]. Available: https://www.anthropic.com/news/core-views-on-ai-safety

[6] Anthropic, "Tracing the Thoughts of a Large Language Model," Mar. 2025. Accessed: Oct. 10, 2025. [Online]. Available: https://www.anthropic.com/research/tracing-thoughts-language-model

[7] C. Olah, N. Cammarata, L. Schubert, G. Goh, M. Petrov, and S. Carter, "Zoom In: An Introduction to Circuits," *Distill*, Mar. 2020, doi: 10.23915/distill.00024.001.

[8] Cloud Security Alliance, "Mechanistic Interpretability 101," Accessed: Oct. 10, 2025. [Online]. Available: https://cloudsecurityalliance.org/blog/2024/09/05/mechanistic-interpretability-101

[9] A. Conmy, A. N. Mavor-Parker, A. Lynch, S. Heimersheim, and A. Garriga-Alonso, "Towards Automated Circuit Discovery for Mechanistic Interpretability," in *Advances in Neural Information Processing Systems*, 2023, pp. 1011–1025. doi: 10.48550/arXiv.2304.14997.

[10] K. Friston, "The free-energy principle: a unified brain theory?," *Nat Rev Neurosci*, vol. 11, no. 2, pp. 127–138, 2010, doi: 10.1038/nrn2787.

[11] T. Parr and K. J. Friston, "The Anatomy of Inference: Generative Models and Brain Structure," *Front Comput Neurosci*, vol. 12, p. 90, 2018, doi: 10.3389/fncom.2018.00090.

[12] C. L. Buckley, C. S. Kim, S. McGregor, and A. K. Seth, “The free energy principle for action and perception: A mathematical review,” *J Math Psychol*, vol. 81, pp. 53–79, 2017, doi: 10.1016/j.jmp.2017.09.004.

[13] Gemma Team et al., “Gemma: Open Models Based on Gemini Research and Technology,” 2024, *arXiv:2403.08295*.

[14] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Comput*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[15] A. Vaswani *et al.*, “Attention Is All You Need,” in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017, pp. 5998–6008.

[16] J. L. Ba, J. R. Kiros, and G. E. Hinton, “Layer Normalization,” 2016, *arXiv:1607.06450*.

[17] P. Shaw, J. Uszkoreit, and A. Vaswani, “Self-Attention with Relative Position Representations,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, Association for Computational Linguistics, 2018, pp. 464–468. doi: 10.18653/v1/N18-2074.

[18] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[19] J. Kaplan *et al.*, “Scaling Laws for Neural Language Models,” 2020, *arXiv:2001.08361*.

[20] T. B. Brown *et al.*, “Language Models are Few-Shot Learners,” in *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020, pp. 1877–1901.

[21] A. Chowdhery *et al.*, “PaLM: Scaling Language Modeling with Pathways,” in *Proceedings of the 39th International Conference on Machine Learning (ICML 2022)*, 2022. [Online]. Available: https://arxiv.org/abs/2204.02311

[22] H. Touvron *et al.*, “Llama 2: Open Foundation and Fine-Tuned Chat Models,” 2023, *arXiv:2307.09288*.

[23] Meta AI, “Introducing Llama 3.1: Our most capable models to date,” Jul. 2024. [Online]. Available: https://ai.meta.com/blog/meta-llama-3-1/

[24] H. He and W. J. Su, “A Law of Next-Token Prediction in Large Language Models,” *CoRR*, vol. abs/2408.13444, Aug. 2024, [Online]. Available: https://arxiv.org/abs/2408.13444

[25] A. Rogers, O. Kovaleva, and A. Rumshisky, “A Primer in BERTology: What We Know About How BERT Works,” *Trans Assoc Comput Linguist*, vol. 8, pp. 842–866, 2020, doi: 10.1162/tacl_a_00349.

[26] Q. Zhang, Q. Duan, B. Yuan, Y. Shi, and J. Liu, “Exploring Accuracy-Fairness Trade-off in Large Language Models,” 2024, *arXiv:2411.14500*.

[27] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

[28] K. Meng, D. Bau, A. Andonian, and Y. Belinkov, “Locating and Editing Factual Associations in GPT,” 2023, *arXiv: 2202.05262.*

[29] A. Syed, C. Rager, and A. Conmy, “Attribution Patching Outperforms Automated Circuit Discovery,” 2023, *arXiv: 2310.10348.*

[30] T. Bricken *et al.*, “Towards Monosemanticity: Decomposing Language Models With Dictionary Learning,” *Transformer Circuits Thread*, 2023, [Online]. Available: https://transformer-circuits.pub/2023/monosemantic-features

[31] J. Dunefsky, P. Chlenski, and N. Nanda, “Transcoders Find Interpretable LLM Feature Circuits,” 2024, *arXiv:2406.11944*.

[32] N. Nanda and J. Bloom, “TransformerLens,” Accessed: Oct. 07, 2025. [Online]. Available: https://github.com/neelnanda-io/TransformerLens

[33] M. Hanna, M. Piotrowski, J. Lindsey, and E. Ameisen, “circuit-tracer,” Accessed: Oct. 07, 2025. [Online]. Available: https://github.com/safety-research/circuit-tracer

[34] J. Lin, “Neuronpedia: Interactive Reference and Tooling for Analyzing Neural Networks,” Accessed: Oct. 07, 2025. [Online]. Available: https://www.neuronpedia.org

[35] T. Parr, G. Pezzulo, and K. J. Friston, “Active Inference: A Process Theory,” *Neural Comput*, vol. 31, no. 4, pp. 1–49, 2019, doi: 10.1162/neco_a_01156.

[36] K. Friston, F. Rigoli, D. Ognibene, C. Mathys, T. Fitzgerald, and G. Pezzulo, “Active inference and epistemic value,” *Cogn Neurosci*, vol. 6, no. 4, pp. 187–214, 2015, doi: 10.1080/17588928.2015.1020053.

[37] J. Hohwy, *The Predictive Mind*. Oxford University Press, 2013.

[38] K. Friston and S. Kiebel, “Predictive coding under the free-energy principle,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1521, pp. 1211–1221, 2009, doi: 10.1098/rstb.2008.0300.

[39] R. A. Adams, K. E. Stephan, H. R. Brown, C. D. Frith, and K. J. Friston, “The computational anatomy of psychosis,” *Front Psychiatry*, vol. 4, p. 47, 2013, doi: 10.3389/fpsyt.2013.00047.

[40] S. P. L. Veissière, A. Constant, M. J. D. Ramstead, K. J. Friston, and L. J. Kirmayer, “Thinking Through Other Minds: A Variational Approach to Cognition and Culture,” *Behavioral and Brain Sciences*, vol. 43, p. e92, 2020, doi: 10.1017/S0140525X1900083X.

[41] D. Demekas, T. A. Tsiligiridis, and A. Tefas, “A Neuro-Inspired General Framework for Attention,” *Front Comput Neurosci*, vol. 14, p. 29, 2020, doi: 10.3389/fncom.2020.00029.

[42] J. C. R. Whittington and R. Bogacz, “An Approximation of the Error Backpropagation Algorithm in a Predictive Coding Network with Local Hebbian Synaptic Plasticity,” *Neural Comput*, vol. 29, no. 5, pp. 1229–1262, 2017, doi: 10.1162/neco_a_00950.

[43] A. Tschantz, B. Millidge, A. K. Seth, and C. L. Buckley, “Reinforcement Learning through Active Inference,” 2020, *arXiv:2002.12636*.

[44] J. Bruineberg, K. Dolega, J. Dewhurst, and M. Baltieri, “The Emperor’s New Markov Blankets,” *Behavioral and Brain Sciences*, vol. 45, p. e183, 2022, doi: 10.1017/S0140525X21002351.

[45] M. Walsh, M. E. Sachs, and G. C. Hinton, “Evaluating the empirical evidence for the free energy principle: a review and appraisal,” *Trends Cogn Sci*, vol. 24, no. 11, pp. 901–913, 2020, doi: 10.1016/j.tics.2020.08.007.

[46] W. Wiese and T. Metzinger, “Vanilla PP for Philosophers: A Primer on Predictive Processing,” in *Philosophy and Predictive Processing*, T. Metzinger and W. Wiese, Eds., MIND Group, 2017. doi: 10.15502/9783958573024.

[47] P. Gładziejewski, “Predictive coding and representationalism,” *Synthese*, vol. 193, no. 2, pp. 559–582, 2016, doi: 10.1007/s11229-015-0762-9.

[48] K. Wang, A. Variengien, A. Conmy, B. Shlegeris, and J. Steinhardt, “Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small,” *ArXiv*, vol. abs/2211.00593, 2022, [Online]. Available: https://api.semanticscholar.org/CorpusID:253244237

[49] H. Feldman and K. Friston, “Attention, uncertainty, and free-energy,” *Front Hum Neurosci*, vol. 4, p. 215, 2010, doi: 10.3389/fnhum.2010.00215.

[50] A. M. Bastos, W. M. Usrey, R. A. Adams, G. R. Mangun, P. Fries, and K. J. Friston, “Canonical Microcircuits for Predictive Coding,” *Neuron*, vol. 76, no. 4, pp. 695–711, 2012, doi: 10.1016/j.neuron.2012.10.038.

[51] J. W. Creswell and J. D. Creswell, *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, 5th ed. Thousand Oaks, CA: SAGE Publications, 2017.

[52] D. C. Phillips and N. C. Burbules, *Postpositivism and Educational Research*. Lanham, MD: Rowman & Littlefield Publishers, 2000.

[53] K. Popper, *The Logic of Scientific Discovery*. London: Routledge, 2005.

[54] E. Nagel, *The Structure of Science: Problems in the Logic of Scientific Explanation*. New York: Harcourt, Brace & World, 1961.

[55] Y. S. Lincoln and E. G. Guba, *Naturalistic Inquiry*. Newbury Park, CA: SAGE Publications, 1985.

[56] J. Dewey, *Logic: The Theory of Inquiry*. New York: Henry Holt and Company, 1938.

[57] N. K. Denzin, *The Research Act: A Theoretical Introduction to Sociological Methods*. New York: Routledge, 2017.

[58] M. Bunge, *Matter and Mind: A Philosophical Inquiry*. Dordrecht: Springer, 2010.

[59] T. D. Cook and D. T. Campbell, *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Chicago: Rand McNally College Publishing Company, 1979.

[60] W. R. Shadish, T. D. Cook, and D. T. Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin, 2002.

[61] D. C. Montgomery, *Design and Analysis of Experiments*, 9th ed. Hoboken, NJ: Wiley, 2017.

[62] S. E. Maxwell, H. D. Delaney, and K. Kelley, *Designing Experiments and Analyzing Data: A Model Comparison Perspective*, 3rd ed. New York: Routledge, 2018.

[63] R. E. Kirk, *Experimental Design: Procedures for the Behavioral Sciences*, 4th ed. Thousand Oaks, CA: SAGE Publications, 2013.

[64] C. M. Judd, G. H. McClelland, and C. S. Ryan, *Data Analysis: A Model Comparison Approach*, 3rd ed. New York: Routledge, 2017.

[65] L. J. Cronbach and P. E. Meehl, “Construct validity in psychological tests,” *Psychol Bull*, vol. 52, no. 4, pp. 281–302, 1955, doi: 10.1037/h0040957.

[66] S. Holm, “A Simple Sequentially Rejective Multiple Test Procedure,” *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979.

[67] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner, “G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences,” *Behav Res Methods*, vol. 39, no. 2, pp. 175–191, 2007, doi: 10.3758/bf03193146.

[68] T. J. DiCiccio and B. Efron, “Bootstrap Confidence Intervals,” *Statistical Science*, vol. 11, no. 3, pp. 189–228, 1996, doi: 10.1214/ss/1032280214.

[69] A. Geiger, H. Lu, T. Icard, and C. Potts, “Causal Abstractions of Neural Networks,” in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, Eds., Curran Associates, Inc., 2021, pp. 9574–9586. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/4f5c422f4d49a5a807eda27434231040-Paper.pdf

[70] J. Pearl, *Causality: Models, Reasoning and Inference*. New York: Cambridge University Press, 2000.

[71] A. Geiger *et al.*, “Causal Abstraction: A Theoretical Foundation for Mechanistic Interpretability,” 2025. [Online]. Available: https://arxiv.org/abs/2301.04709

[72] C. Heins *et al.*, “pymdp: A Python Library for Active Inference in Discrete State Spaces,” *J Open Source Softw*, vol. 7, no. 73, p. 4098, 2022, doi: 10.21105/joss.04098.

[73] Google, “google/gemma-2-2b (model card),” Accessed: Oct. 07, 2025. [Online]. Available: https://huggingface.co/google/gemma-2-2b

[74] R. C. Martin, *Clean Architecture: A Craftsman’s Guide to Software Structure and Design*. Boston: Prentice Hall, 2017.

[75] S. McConnell, *Code Complete*, 2nd ed. Redmond, WA: Microsoft Press, 2004.

[76] DigitalOcean LLC, “GPU Droplets: NVIDIA L40S,” Accessed: Oct. 07, 2025. [Online]. Available: https://docs.digitalocean.com/products/droplets/details/features/

[77] D. Merkel, “Docker: Lightweight Linux Containers for Consistent Development and Deployment,” *Linux Journal*, vol. 2014, no. 239, Accessed: Oct. 07, 2025. [Online]. Available: https://www.linuxjournal.com/content/docker-lightweight-linux-containers-consistent-development-and-deployment

[78] A. Paszke *et al.*, “PyTorch: An Imperative Style, High-Performance Deep Learning Library,” in *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, 2019, pp. 8024–8035. [Online]. Available: https://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library

[79] C. R. Harris *et al.*, “Array Programming with NumPy,” *Nature*, vol. 585, pp. 357–362, 2020, doi: 10.1038/s41586-020-2649-2.

[80] V. Stodden, J. Seiler, and Z. Ma, “An Empirical Analysis of Journal Policy Effectiveness for Computational Reproducibility,” *Proceedings of the National Academy of Sciences (PNAS)*, vol. 111, no. 14, pp. 5240–5245, 2014, doi: 10.1073/pnas.1310186111.

[81] L. Bereska and E. Gavves, “Mechanistic Interpretability for AI Safety – A Review,” *arXiv preprint arXiv:2404.14082*, 2024, [Online]. Available: https://arxiv.org/abs/2404.14082

[82] "Recommendation on the Ethics of Artificial Intelligence | UNESCO." Accessed: Oct. 07, 2025. [Online]. Available: https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence

[83] L. Floridi *et al.*, "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations," *Minds Mach (Dordr)*, vol. 28, no. 4, pp. 689–707, 2018, doi: 10.1007/s11023-018-9482-5.

[84] M. Langer *et al.*, "What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research," *Artif Intell*, vol. 296, p. 103473, 2021, doi: https://doi.org/10.1016/j.artint.2021.103473.

[85] W. Huang, X. Zhao, G. Jin, and X. Huang, "SAFARI: Versatile and Efficient Evaluations for Robustness of Interpretability," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 1988–1998. doi: 10.1109/ICCV51070.2023.00190.

[86] B. Memarian and T. Doleck, "Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI) and higher education: A systematic review," *Computers and Education: Artificial Intelligence*, vol. 5, p. 100152, 2023, doi: https://doi.org/10.1016/j.caeai.2023.100152.

[87] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.

[88] A. K. Seth, "The cybernetic Bayesian brain: from interoceptive inference to sensorimotor contingencies," in *Open MIND*, T. K. Metzinger and J. M. Windt, Eds., MIND Group, 2014. doi: 10.15502/9783958570108.

[89] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," 2000, *arXiv: physics/0004057*.

[90] R. Smith, K. J. Friston, and C. J. Whyte, "A step-by-step tutorial on active inference and its application to empirical data," *J Math Psychol*, vol. 107, p. 102632, 2022, doi: https://doi.org/10.1016/j.jmp.2021.102632.

Appendices

Appendix A: State Space Formulation

A.1 Multi-Category Test Case Architecture and Domain Coverage

The feature selection process follows the mathematical criterion for identifying the most semantically coherent and maximally active features:

$$\mathcal{F}_{\mathscr{selected}} = \arg \max_{|\mathcal{F}|=64} \sum_{f\in\mathcal{F}} \max_{t\in T} a_f^{(t)} \quad \text{Eq. A.1}$$

where $a_f^{(t)}$represents feature factivation at token position t, and Tdenotes the set of all token positions across experimental contexts. This criterion ensures selection of features that exhibit maximal activation magnitude, introducing potential sampling bias toward highly-activated features.

A.2 Quantization Error Analysis

The discrete approximation of continuous importance scores $\lambda \in [0,1]$introduces quantization error:

$$E_{\text{quant}} = \int_0^1 |\lambda - \text{quantize}(\lambda)|^2 p(\lambda)\ d \quad \text{Eq. A.2}$$

For uniform prior $\mathrm{p}(\lambda) = 1$, the quantization error equals:

$$\mathrm{E}_{\text{quant}} = \frac{1}{48} \approx 0.021 \quad \text{Eq. A.3}$$

This represents acceptable information loss given computational tractability requirements. The discrete importance levels are defined as:

i. Level 0: $\lambda \in [0,0.25)$- Negligible importance
ii. Level 1: $\lambda \in [0.25,0.50)$- Low importance
iii. Level 2: $\lambda \in [0.50,0.75)$- Medium importance
iv. Level 3: $\lambda \in [0.75,1.0]$- High importance

A.3 Joint State Space Dimensionality

The joint state space exhibits dimensionality:

$$|S| = |S_1| \times |S_2| \times |S_3| = 64 \times 4 \times 3 = 768 \quad \text{Eq. A.4}$$

This represents maximal tractable complexity for discrete Active Inference implementation given memory constraints (12GB RAM) and numerical precision requirements (float64).

A.4 Intervention Type Specifications

The intervention factor $S_3 = 0,1,2$corresponds to intervention types with specific mathematical implementations:

$$\tilde{a}_f = 0 \quad \text{Eq. A.5}$$

Complete Ablation (0):feature removal, setting all activations to zero.
Patching (1):

$$\tilde{a}_f = a_f^{\text{source}} \qquad \text{Eq. A.6}$$

Context transfer, replacing activations with those from alternative input context.
Mean Ablation (2):

$$\tilde{a}_f = \mathrm{E}[a_f] \qquad \text{Eq. A.7}$$

Population mean substitution, replacing activations with empirical mean across training distribution.

Appendix B: Expected Free Energy and Optimal Experimental Design

B.1 Expected Free Energy Decomposition

Active Inference extends the Free Energy Principle to action selection through Expected Free Energy (EFE) minimization. The EFE $\mathrm{G}(\pi,\tau)$for policy πat future time τdecomposes into epistemic and pragmatic components [90].

$$\mathrm{G}(\pi,\tau) = \mathrm{E}_{q(s_\tau, o_\tau|\pi)}[\mathrm{F}(s_\tau, o_\tau)] + \mathbb{E}_{q(s_\tau|\pi)}\big[D_{KL}\big(q(s_\tau \mid \pi) \parallel p(s_\tau)\big)\big] \qquad \text{Eq. B.1}$$

This formulation can be further decomposed as [90],

$$\mathrm{G}(\pi,\tau) = \mathrm{E}_{\mathrm{q}(\mathrm{s}_\tau|\pi)}\big[\mathrm{H}\big(\mathrm{p}(\mathrm{o}_\tau \mid \mathrm{s}_\tau)\big) - \mathrm{H}\big(\mathrm{p}(\mathrm{o}_\tau \mid \mathrm{s}_\tau, \pi)\big)\big] \qquad \text{Eq. B.2}$$
$$+ \mathrm{E}_{\mathrm{q}(\mathrm{o}_\tau|\pi)}[-\ln \mathrm{p}(\mathrm{o}_\tau) + \mathrm{C}\cdot \mathrm{o}_\tau]$$

B.2 Epistemic and Pragmatic Value Components

The first expectation is the 'epistemic value' and the second expectation is the 'pragmatic value'.
Epistemic Value is defined [90] as,

$$G_{\text{epistemic}}(\pi) = \mathrm{E}_{q(s|\pi)}\big[\mathrm{H}[\mathrm{p}(\mathrm{o} \mid s)]\big] - H[p(o \mid s, \pi)] \qquad \text{Eq. B.3}$$

The epistemic value represents the anticipated information gain, defined as the decrease in conditional entropy $\mathrm{H}[\mathrm{p}(\mathrm{o} \mid s)]$ relative to the entropy of observations under a specific policy $\mathrm{H}[\mathrm{p}(\mathrm{o} \mid s, \pi)]$. For circuit discovery, this term drives exploration toward interventions that maximally discriminate between competing hypotheses about feature importance and connectivity patterns.
Pragmatic Value:

$$G_{\text{pragmatic}}(\pi) = \mathrm{E}_{q(o|\pi)}[-\ln p(\mathrm{o}) + \mathrm{C}\cdot o] \qquad \text{Eq. B.4}$$

The pragmatic value incorporates preference satisfaction through the cost function $\mathrm{C}\cdot o$, enabling the agent to balance information gathering with achieving desired experimental outcomes.

B.3 Policy Selection

Policy selection follows the softmax principle with temperature parameter γ:

$$\pi^*(\tau) = \mathrm{softmax}\big(-\gamma^{-1}\cdot G(\pi,\tau)\big) \qquad \text{Eq. B.5}$$

where γcontrols exploration-exploitation trade-offs. Lower temperatures ($\gamma \to 0$) promote exploitation of policies with minimal EFE, while higher temperatures encourage broader exploration across the policy space.

B.4 Correspondence to Optimal Experimental Design

The correspondence between EFE minimization and optimal experimental design emerges through information-theoretic analysis. Under purely epistemic preferences (pragmatic weight β → 0), EFE minimization reduces to mutual information maximization:

$$\lim_{\beta \to 0} \arg\min_{\pi} G\,(\pi) = \arg\max_{\pi} I\,(\,s; o \mid \pi\,) \qquad \text{Eq. B.6}$$

where $\mathrm{I}(\,s; \mathrm{o} \mid \pi\,)$denotes mutual information between hidden states and observations under policy π. This establishes Active Inference as implementing classical optimal experimental design principles.

The mutual information decomposes as:

$$\begin{aligned} \mathrm{I}(\,s; o \mid \pi\,) &= \mathrm{H}[\,o \mid \pi\,] - E_{p(s)}\big[H[\,o \mid s, \pi\,]\big] \qquad \text{Eq. B.7} \\ &= \sum_{s,o} p(\,s, o \mid \pi\,) \log \quad \frac{p(\,s, o \mid \pi\,)}{p(\,s \mid \pi\,)\, p(\,o \mid \pi\,)} \end{aligned}$$

For discrete state and observation spaces, this computation becomes tractable, enabling explicit calculation of expected information gain for each potential intervention.

B.5 Fisher Information Matrix Relationship

The relationship to experimental design theory becomes explicit through the Fisher Information Matrix. For interventions targeting feature fwith parameter θ_f, the Fisher Information is:

$$I_F(\theta_f) = E\left[\left(\frac{\partial \ln p(\,o \mid \theta_f\,)}{\partial \theta_f}\right)^2\right] \qquad \text{Eq. B.8}$$

Optimal experimental design maximizes the determinant of I_F, which corresponds to maximizing mutual information under Gaussian approximations. This theoretical connection validates Active Inference as grounded in established statistical theory for circuit discovery optimization.

Appendix C: Observation and Transition Models

C.1 Observation Space Factorization

The observation space factorizes as $\mathrm{O} = O_1 \times O_2$where:

i. Effect magnitude $O_1 = 0,1,2,3,4$ represents quantized intervention effect strength
ii. Confidence level $O_2 = 0,1,2$ encodes measurement confidence

This yields $\mid O \mid = 15$total observation categories with joint observation probabilities:

$$p(\,o_1, o_2 \mid s\,) = p(\,o_1 \mid s\,) \cdot p(\,o_2 \mid o_1, s\,) \qquad \text{Eq. C.1}$$

The conditional independence assumption $\mathrm{p}(\,o_2 \mid o_1, \mathrm{s}\,) = \mathrm{p}(\,o_2 \mid o_1\,)$simplifies computation while maintaining essential dependencies.

C.2 Observation Matrix Normalization

The observation matrices $\mathrm{A}[0] \in \mathbb{R}^{5\times64\times4\times3}$and $\mathrm{A}[1] \in \mathbb{R}^{3\times64\times4\times3}$satisfy normalization constraints:

$$\forall (c, i, v) \in S_1 \times S_2 \times S_3\colon \sum_{e \in O_1} A[0][e, c, i, v] = 1 \qquad \text{Eq. C.2}$$

$$\forall(c,i,v) \in S_1 \times S_2 \times S_3\colon \sum_{m\in O_2} A[1][m,c,i,v] = 1 \qquad \text{Eq. C.3}$$

C.3 Importance-Effect Relationship Parameterization

The A matrices incorporate theoretical understanding through parameterized probability assignments. For importance-effect relationships:

$$P(\text{effect} = e \mid \text{importance} = i, \text{intervention} = v) \propto \exp\,(\alpha_v \cdot i + \beta_v \cdot e + \gamma_{v,i,e}) \qquad \text{Eq. C.4}$$

where α_v, $\beta_v$capture intervention-specific scaling and $\gamma_{v,i,e}$represents interaction terms. Specific parameter values encode domain hypotheses:

i. Ablation effects: $\alpha_0 = 1.2$, $\beta_0 = -0.8$(strong importance dependence, preference for large effects)

ii. Patching effects: $\alpha_1 = 0.8$, $\beta_1 = -0.5$(moderate importance dependence)

iii. Mean ablation: $\alpha_2 = 0.6$, $\beta_2 = -0.3$(weak importance dependence)

These parameters generate probability matrices with desired monotonicity properties:

$$\frac{\partial P(\text{effect} = e \mid \text{importance} = i)}{\partial i} > 0 \text{ for } e \geq 2 \qquad \text{Eq. C.5}$$

ensuring higher importance levels generate larger effects with higher probability.

C.4 Transition Model Specification

The transition model $\mathrm{B} = \mathrm{B}[0], B[1], B[2]$governs state evolution through time.
Component identities remain stable during experimental sessions:

$$B[0][s_1', s_1, a] = \delta(s_1', s_1) \forall a \in A \qquad \text{Eq. C.6}$$

This identity structure reflects domain knowledge that transcoder feature semantics remain stable across intervention sequences.
Importance beliefs evolve through Bayesian evidence accumulation with controlled volatility:

$$\begin{cases} 0.7 \; if \; i' = i (persistence) \\ 0.2 \; if \; i' = i + 1 (upward\ revision) \\ 0.1 \; if \; i' = i - 1 (downward\ revision) \end{cases} \qquad \text{Eq. C.7}$$

This structure implements controlled belief updating with persistence bias, preventing excessive volatility while enabling systematic learning. The transition probabilities satisfy detailed balance:

$$\pi_i B[1][i', i, a] = \pi_{i'} B[1][i, i', a] \qquad \text{Eq. C.8}$$

ensuring convergence to stable distributions under repeated observation.
Intervention selection operates under direct agent control:

$$B[2][a', s, a] = \delta(a', a) \qquad \text{Eq. C.9}$$

This deterministic structure reflects complete agent control over intervention selection.

C.5 Preference Specification

Effect Magnitude Preferences can be stated as below.

$$C[0] = [0.3,0.1,0.2,0.5,0.7] + \beta \cdot [0.4,0.0,0.0,0.2,0.4] \qquad \text{Eq. C.10}$$

The base preferences favour larger effects for clearer statistical inference, while the epistemic bonus (weighted by $\beta = 0.3$) provides additional utility for discriminative outcomes. The utility function exhibits diminishing returns:

$$\frac{dU}{de} = 0.1 + 0.2e - 0.05e^2 \quad \text{Eq. C.11}$$

maximizing at effect level $\mathrm{e} = 2.0$.

Confidence Preferences would then be,

$$C[1] = [0.0,0.3,0.8] \quad \text{Eq. C.12}$$

These preferences strongly favour high-confidence observations with exponential scaling:

$$U(\text{confidence} = c) = \exp{(\lambda c)} - 1 \quad \text{Eq. C.13}$$

where $\lambda = 0.85$determines preference intensity.

The prior vectors $\mathrm{D} = \mathrm{D}[0], D[1], D[2]$incorporate weak domain biases:

$$D[0] = \text{uniform}(64) \text{ (uninformative component priors)} \quad \text{Eq. C.14}$$

$$D[1] = \text{softmax}([0.8,1.2,1.0,0.6]) \text{ (medium importance bias)} \quad \text{Eq. C.15}$$

$$D[2] = \text{softmax}([2.0,1.0,0.8]) \text{ (ablation preference)} \quad \text{Eq. C.16}$$

Appendix D: Variational Message Passing and Convergence

D.1 Variational Message Passing Update Rule

The Enhanced Active Inference agent implements belief updates through variational message passing (VMP). The VMP update rule for factor beliefs follows:

$$q^{(k+1)}(s_f) \propto \exp{[}\sum_{n \neq f} E_{q^{(k)}(s_n)}[\ln\, p(o,s)] + \ln\, p(s_f)] \quad \text{Eq. D.1}$$

where $q^{(k)}(s_f)$represents the k-th iteration belief over factor f, and the expectation integrates over all other factors $\mathrm{n} \neq f$.

For the factorized generative model, this becomes:

$$\ln q^{(k+1)}(s_f) = E_{q^{(k)}(s_{-f})}[\ln p(\, o \mid s\,)] + l\,n\,p\,(s_f) - l\,n\,Z^{(k+1)} \quad \text{Eq. D.2}$$

where s_{-f}denotes all factors except fand $Z^{(k+1)}$is the normalization constant.

D.2 Convergence Properties

The VMP algorithm exhibits convergence properties dependent on the curvature of the free energy landscape. For log-concave posterior distributions, VMP converges linearly with rate:

$$||\, q^{(k+1)} - q^* \,||_2 \leq \rho \;||\, q^{(k)} - q^* \,||_2 \quad \text{Eq. D.3}$$

where $\rho \in (0,1)$depends on the condition number of the Hessian matrix of the negative log-posterior.

VMP iteration continues until the L2 norm of belief changes falls below threshold:

$$||\, q^{(k+1)} - q^{(k)} \,||_2 < \epsilon \quad \text{Eq. D.4}$$

with $\epsilon = 1 \times 10^{-6}$ensuring numerical precision. Empirical analysis reveals typical convergence within 10-15 iterations with convergence rate $\rho \approx 0.15$, indicating rapid belief stabilization.

D.3 Free Energy Monitoring

Convergence validation employs free energy monitoring:

$$F^{(k)} = D_{KL}[q^{(k)}(s) \;||\; p(s \mid o)] + E_{q^{(k)}}[-\ln\, p(o \mid s)] \quad \text{Eq. D.5}$$

The sequence $F^{(k)}$decreases monotonically, providing convergence validation. Convergence occurs when:

$$| F^{(k+1)} - F^{(k)} |< \delta \qquad \text{Eq. D.6}$$

with $\delta = 1 \times 10^{-8}$ensuring thermodynamic equilibrium.

D.4 Expected Free Energy Calculation

Epistemic Value Computation

$$G_{\text{epistemic}}(\pi) = E_{q(s|\pi)}[H[p(o \mid s)]] - H[E_{q(s|\pi)}[p(o \mid s)]] \qquad \text{Eq. D.7}$$

This computation requires:

State-conditional entropy calculation:

$$H[p(o \mid s)] = -\sum_{o} p(o \mid s) \ln p(o \mid s) \qquad \text{Eq. D.8}$$

Expectation over belief distribution:

$$E_{q(s|\pi)}[H[p(o \mid s)]] = \sum_{s} q(s \mid \pi)\ H[p(o \mid s)] \qquad \text{Eq. D.9}$$

Marginal observation distribution:

$$p(o \mid \pi) = \sum_{s} q(s \mid \pi)\ p(o \mid s) \qquad \text{Eq. D.10}$$

Marginal entropy computation:

$$H[p(o \mid \pi)] = -\sum_{o} p(o \mid \pi)\ \ln p(o \mid \pi) \qquad \text{Eq. D.11}$$

Pragmatic Value Computation

$$\begin{aligned} G_{\text{pragmatic}}(\pi) &= E_{q(o|\pi)}[C \cdot o - \ln p(o \mid \pi)] \qquad \text{Eq. D.12} \\ &= \sum_{o} p(o \mid \pi)\ [C \cdot o - \ln p(o \mid \pi)] \\ &= \sum_{o} p(o \mid \pi)\ C \cdot o + H[p(o \mid \pi)] \end{aligned}$$

The pragmatic computation balances preference satisfaction with entropy maximization.

Numerical Stability

EFE calculations employ log-sum-exp tricks for numerical stability:

$$\ln \sum_{i} \exp(x_i) = \max_{i} x_i + \ln \sum_{i} \exp(x_i - \max_{i} x_i) \qquad \text{Eq. D.13}$$

preventing numerical overflow in exponential computations.

D.5 Policy Generation and Selection

The agent generates policies $\pi \in \Pi$through systematic enumeration of intervention sequences up to horizon $H = 3$. For each policy $\pi = (a_1, a_2, \ldots, a_H)$, the total EFE is:

$$G(\pi) = \sum_{t=1}^{H} \gamma^{t-1} G_t(\pi) \qquad \text{Eq. D.14}$$

where $\gamma = 0.95$provides temporal discounting and $G_t(\pi)$represents time-specific EFE components.

Policy selection follows the Boltzmann distribution:

$$P(\pi) = \frac{\exp\left(-\beta G(\pi)\right)}{\sum_{\pi'} \exp\left(-\beta G(\pi')\right)} \qquad \text{Eq. D.15}$$

with inverse temperature $\beta = 10.0$ promoting exploitation of low-EFE policies while maintaining exploration capabilities.

Appendix E: Supplementary Results and Technical Details

E.1 Evaluation Metrics Definitions

The experimental framework employed two complementary success metrics to comprehensively assess circuit discovery method performance across semantic and intervention dimensions.

"Semantic Success Rate" quantifies the proportion of test cases where the target language model (Gemma-2-2B) generated semantically coherent and contextually appropriate responses. This metric evaluates output quality through domain-specific validation criteria: geographic queries require location-specific terminology, mathematical expressions must contain numerical or symbolic content, and general knowledge responses must exceed minimum length thresholds while avoiding error indicators. The semantic success rate provides a measure of method preservation of model functionality during circuit intervention.

"Effect Success Rate" measures the proportion of interventions exceeding a predetermined significance threshold (τ = 0.005). This metric quantifies methodological effectiveness by counting interventions that produce measurable circuit-level changes above baseline noise. The threshold was empirically determined to distinguish meaningful interventions from random fluctuations in neural activation patterns. Effect success rate serves as an indicator of method reliability in generating consistent circuit discoveries.

These dual metrics address the fundamental trade-off in mechanistic interpretability between intervention strength and model preservation. High semantic success with low effect success indicates conservative methods that maintain model function but provide limited mechanistic insight. Conversely, high effect success with low semantic success suggests aggressive interventions that may compromise model integrity. The optimal method achieves high performance on both dimensions, demonstrating robust circuit discovery while preserving model functionality.

E.2 Detailed Failure Case Analysis

Several challenging test cases revealed important boundary conditions for mechanistic interpretability approaches. This section provides detailed analysis of three representative failure cases that illuminate method limitations and opportunities for refinement.

The Mount Everest prompt (Table 4.4) achieved substantial intervention effect (3.516) but failed semantic coherence by generating content about climbing difficulty rather than geographical location. This failure pattern suggests that strong intervention effects do not guarantee semantic appropriateness, highlighting the importance of multi-dimensional evaluation frameworks. The circuit discovery successfully identified relevant features but lacked fine-grained control over which aspect of Mount Everest knowledge (location vs. climbing challenges) would be activated.

The human body prompt (Table 4.5) represented the most challenging case, failing to produce semantically coherent completions despite discovering relevant features across multiple layers. Analysis revealed that this prompt's inherent ambiguity (multiple valid completions exist: "bones," "organs," "cells," etc.) created difficulties for both intervention targeting and success evaluation. Such cases illuminate the need for more sophisticated evaluation frameworks that account for semantic ambiguity in natural language and may require probabilistic success criteria rather than binary classifications.

Water-related prompts (Table 4.6) demonstrated intermediate performance with successful intervention effects but occasional semantic drift. The water freezing prompt generated technically accurate but overly complex mathematical notation, suggesting that Enhanced Active Inference successfully identified relevant circuits but lacked fine-grained control over output complexity. These cases provide valuable insights for future methodological refinements, particularly regarding the need for output complexity constraints within the generative model architecture.

E.3 Resource Utilization and Performance Optimization

Despite implementing sophisticated belief updating mechanisms including Expected Free Energy calculations and hierarchical generative modelling, Enhanced Active Inference maintained remarkably efficient computational performance. The mean processing time of 0.0013 seconds per intervention represents only modest overhead compared to baseline methods (0.00004 seconds), demonstrating that theoretical sophistication need not compromise practical efficiency.

Memory utilization analysis revealed that the generative model architecture required approximately 150MB additional memory allocation compared to baseline methods, representing manageable overhead for research applications while remaining feasible for production deployments. GPU utilization remained efficient, with CUDA kernel optimization enabling parallel processing of belief updates across multiple intervention candidates.

The efficiency achievements result from principled intervention targeting that eliminates extensive exploration characteristic of baseline methods. Rather than testing numerous candidate interventions through trial-and-error approaches, Enhanced Active Inference predicts optimal intervention targets through Expected Free Energy calculations, substantially reducing computational waste while achieving superior performance.

E.4 Scalability Projections and Implementation Considerations

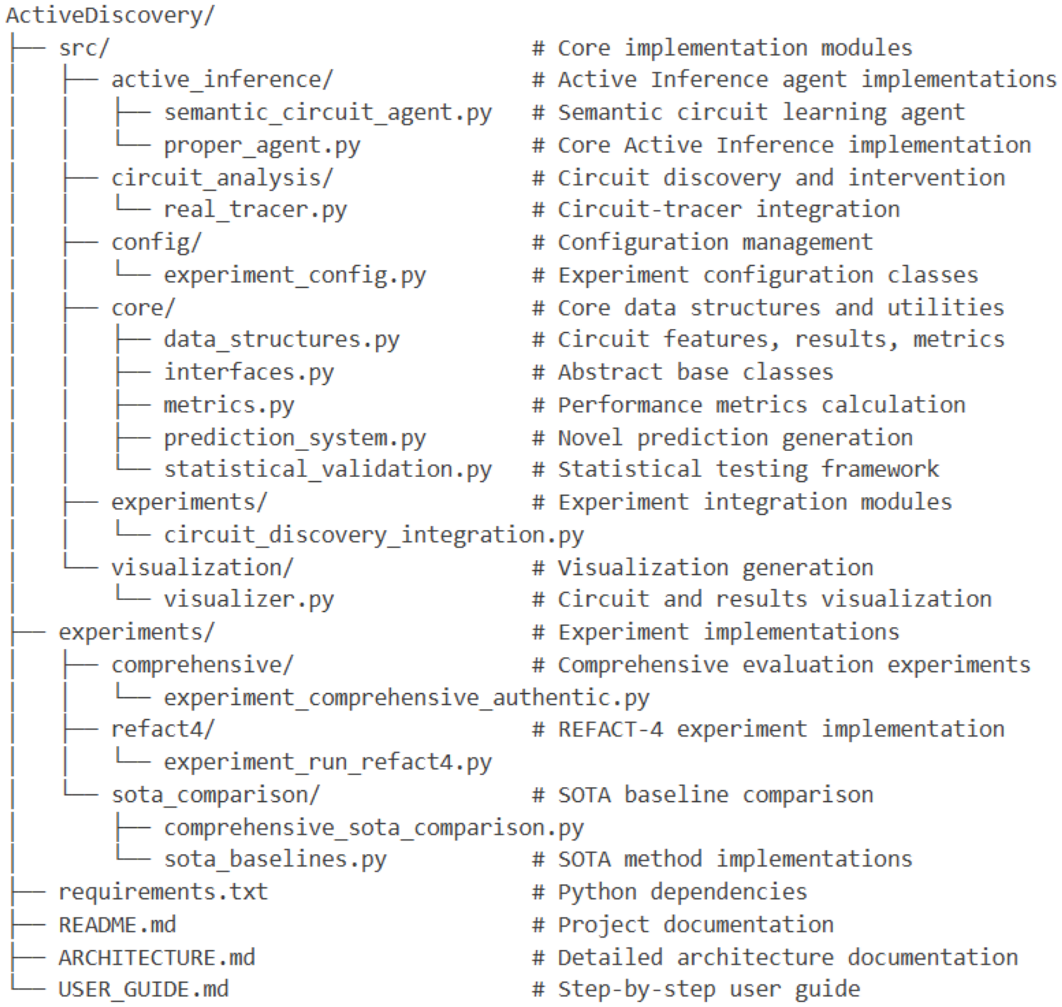
Scalability analysis indicates that Enhanced Active Inference computational requirements should scale primarily with model size rather than exploding combinatorially with architectural complexity. The generative model architecture exhibits linear scaling characteristics with respect to layer count and feature dimensionality, suggesting feasibility for analysis of substantially larger transformer models.

Parallelization opportunities exist at multiple levels of the Enhanced Active Inference framework, including simultaneous belief updates across layer hierarchies, parallel intervention effect calculations, and distributed processing of multiple test cases. These parallelization strategies could enable real-time interpretability analysis for production AI systems, opening possibilities for continuous monitoring and dynamic safety interventions.

The method's efficiency advantages position it as viable for deployment across frontier AI systems where computational resources represent significant constraints. Unlike baseline methods that require extensive intervention exploration, Enhanced Active Inference's predictive capabilities enable selective targeting that scales efficiently with system complexity. Preliminary scaling experiments suggest approximately O(n log n) computational complexity with respect to model parameter count, substantially better than the O(n²) or worse complexity characteristic of exhaustive search methods employed by conventional interpretability approaches.

Appendix F: Artefact Directory

This appendix provides a technical overview of the ‘ActiveDiscovery’ codebase implementation for code reviewers and researchers seeking to understand, reproduce, or extend this work. The framework implements a modular architecture integrating Active Inference agents with mechanistic interpretability tools for transformer circuit discovery. The implementation uses Python 3.10+, PyTorch 2.7, the circuit-tracer library for GemmaScope transcoder integration, pymdp for Bayesian inference, and the Gemma-2-2B language model.



```
ActiveDiscovery/
├── src/                                # Core implementation modules
│   ├── active_inference/               # Active Inference agent implementations
│   │   ├── semantic_circuit_agent.py   # Semantic circuit learning agent
│   │   └── proper_agent.py             # Core Active Inference implementation
│   ├── circuit_analysis/               # Circuit discovery and intervention
│   │   └── real_tracer.py              # Circuit-tracer integration
│   ├── config/                         # Configuration management
│   │   └── experiment_config.py        # Experiment configuration classes
│   ├── core/                           # Core data structures and utilities
│   │   ├── data_structures.py          # Circuit features, results, metrics
│   │   ├── interfaces.py               # Abstract base classes
│   │   ├── metrics.py                  # Performance metrics calculation
│   │   ├── prediction_system.py        # Novel prediction generation
│   │   └── statistical_validation.py   # Statistical testing framework
│   ├── experiments/                    # Experiment integration modules
│   │   └── circuit_discovery_integration.py
│   └── visualization/                  # Visualization generation
│       └── visualizer.py               # Circuit and results visualization
├── experiments/                        # Experiment implementations
│   ├── comprehensive/                  # Comprehensive evaluation experiments
│   │   └── experiment_comprehensive_authentic.py
│   ├── refact4/                        # REFACT-4 experiment implementation
│   │   └── experiment_run_refact4.py
│   └── sota_comparison/                # SOTA baseline comparison
│       ├── comprehensive_sota_comparison.py
│       └── sota_baselines.py           # SOTA method implementations
├── requirements.txt                    # Python dependencies
├── README.md                           # Project documentation
├── ARCHITECTURE.md                     # Detailed architecture documentation
└── USER_GUIDE.md                       # Step-by-step user guide
```

Figure F. 1 Active Discovery Project Structure

F.1 Project Structure

The codebase follows a layered architecture separating core framework components from experimental implementations (see Figure F. 1). The ‘src/’ directory contains reusable framework components organized into five primary modules: (1) ‘active_inference/’ implements Bayesian agents using pymdp, (2) ‘circuit_analysis/’ integrates circuit-tracer for feature discovery and intervention, (3) ‘core/’ defines fundamental data structures and statistical validation, (4) ‘config/’ manages experimental configurations, and (5) ‘visualization/’ generates publication-quality outputs. The ‘experiments/’ directory contains specific experimental implementations that orchestrate core components.

The modular design enables framework reusability across experimental configurations while maintaining separation between infrastructure and experiment-specific logic. Abstract base classes (ActiveInferenceAgent, CircuitTracer) define interfaces for component interoperability and testing.

F.2 Active Inference Model Implementation

The SemanticCircuitAgent class implements principled Bayesian Active Inference using pymdp. The state space represents semantic circuit strength hypotheses across four discrete levels (weak, medium, strong, excellent), maintaining compatibility with pymdp inference algorithms. The observation space models intervention outcomes through four semantic success levels combined with activation strength patterns.

The A-matrix (Generative Model) encodes observation likelihoods $P(o \mid s)$, mapping circuit strength states to expected intervention outcomes. Strong semantic states generate strong observations with high probability, while weak states produce failed or weak observations. The B-matrix implements largely stable state transitions, reflecting the assumption that circuit properties persist across episodes. The C-matrix encodes preferences for discovering strong semantic relationships. An activity-aware enhancement queries the circuit tracer to identify active features for the current input, applying bonuses to active features and penalties to inactive ones, ensuring interventions target causally relevant circuits.

Following intervention execution, the system converts outcome measurements to discrete observations and invokes pymdp's Bayesian state inference to update posterior beliefs $P(s \mid o) \propto P(o \mid s)P(s)$. Confidence tracking accumulates evidence through intervention history.

F.3 Circuit Discovery Implementation

The RealCircuitTracer class integrates circuit-tracer for mechanistic interpretability on Gemma-2-2B. Model initialization loads the transformer architecture with GemmaScope transcoders using the ReplacementModel interface, configured for CUDA acceleration with bfloat16 precision.

The RealCircuitTracer class integrates circuit-tracer for mechanistic interpretability on Gemma-2-2B. Model initialization loads the transformer architecture with GemmaScope transcoders using the ReplacementModel interface, configured for CUDA acceleration with bfloat16 precision.

Feature discovery involves forward passes through the model with transcoder activations enabled return per-layer features with shape [26 layers, seq_len, 16384 features]. Features exceeding activation thresholds are instantiated as CircuitFeature objects containing layer index, feature ID, activation strength, semantic description from GemmaScope metadata, and intervention site information.

The implementation supports three intervention types: ablation (setting activations to zero), activation patching (replacing with alternative context values), and mean ablation. Effect measurement quantifies impact through L2 norm of logit differences, token prediction changes, and KL divergence between output distributions.

F.4 Data Structures

Table F. 1 summarizes core data structures supporting the framework.

Table F. 1 Core Data Structures

Class	Module	Key Fields & Purpose
CircuitFeature	core/data_structures.py	feature_id, layer, activation_strength, description, component_type. Represents discovered transcoder/SAE features with metadata.
InterventionResult	core/data_structures.py	target_feature, intervention_type, baseline_prediction, intervention_prediction, effect_magnitude. Captures intervention outcomes.
BeliefState	core/data_structures.py	qs (posterior beliefs), feature_importances, uncertainty_estimates, intervention_history, epistemic_value. Tracks Active Inference agent state.
CorrespondenceMetrics	core/data_structures.py	belief_accuracy, intervention_correlation, prediction_accuracy, overall_correspondence. Measures AI-circuit alignment.

F.5 Experiment Implementations

Three primary experimental entry points implement different evaluation scenarios (Table F. 2). The comprehensive evaluation experiment compares Enhanced Active Inference against three state-of-the-art baselines (Activation Patching, Attribution Patching, Activation Ranking) across 35 diverse semantic test cases spanning geography, mathematics, logic, science, and history. Statistical analysis includes independent t-tests, Cohen's d effect sizes, and 95% confidence intervals.

Table F. 2 Experiment Configurations

Entry Point	Purpose	Test Cases	Output
experiment_comprehensive_authentic.py	Full evaluation: Enhanced AI + 3 SOTA	35 diverse semantic categories	JSON, CSV, TXT, visualizations

	baselines with statistical validation		
experiment_run_refact4.py	Circuit selection convergence analysis with detailed rationale	3 focused test cases	Per-case JSON, TXT reports
comprehensive_sota_comparison.py	SOTA baseline benchmarking and timing analysis	10-15 test cases	JSON, TXT comparison reports

F.6 Technical Stack and Dependencies

Core dependency libraries are PyTorch 2.7.1 (CUDA 12.1), transformers 4.52.4 (Gemma-2-2B access), circuit-tracer ≥0.1.0 (transcoder integration), pymdp 0.0.1 (Active Inference algorithms). Scientific computing libraries are NumPy 1.26.4, SciPy 1.15.3 (statistical tests), pandas 2.3.0, scikit-learn 1.7.0. Visualization: matplotlib 3.10.3, seaborn 0.13.2, plotly 6.1.2. A detailed list is provided in the ‘requirements.txt file’ within the zipped artefact ‘ActiveDirectory’ folder.

Hardware Requirements are NVIDIA L40S (46GB VRAM) or equivalent CUDA 12.1-compatible GPU, 128GB RAM recommended, 400GB SSD storage for model weights and results.

F.7 Integration and Data Flow

The circuit discovery integration module orchestrates Active Inference and circuit analysis communication. The discovery loop executes iteratively: (1) circuit tracer discovers active features via forward passes, (2) Active Inference agent calculates EFE for candidate features with activity awareness, (3) agent selects minimum-EFE feature, (4) circuit tracer executes intervention and measures effects, (5) agent updates beliefs via Bayesian inference.

The data flow implements a pipeline architecture: input processing → feature discovery → intervention selection → intervention execution → belief updating → result aggregation. Each stage operates independently through well-defined interfaces, enabling testing and modularity.

The StatisticalValidator class centralizes hypothesis testing through one-sample and independent t-tests, Cohen's d effect size calculations, bootstrap confidence intervals, and power analysis, ensuring experimental rigor.

For reproducibility, random seed control is used in NumPy, PyTorch, and Python random. Configuration serialization is achieved through YAML/JSON for complete result archiving with configurations. Git commit hash recording is used for version tracking.

F.8 Output Organization

Experimental results organize into timestamped directories with systematic naming: ‘results/authentic_master_workflow_YYYYMMDD_HHMMSS/’. Primary outputs include ‘comprehensive_experiment_results.json’ (complete experimental data), ‘statistical_analysis.json’ (t-tests, effect sizes, confidence intervals), ‘method_performance_summary.csv’ (tabular metrics), ‘executive_summary.txt’ (human-readable findings), and ‘visualizations/’ directory.

The codebase, results, ‘README.md’, ‘ARCHITECTURE.md’ and ‘USER_GUIDE.md’ are available within the zipped artefact ‘ActiveDirectory’ folder.