Literature Review

Table of Contents

Overview

Neuro-symbolic AI integrates the robustness of symbolic reasoning with the perceptual learning capabilities of neural networks, providing a pathway to overcome the limitations of purely statistical models like LLMs in tasks that require generalization, reasoning, and traceability. The reviewed literature spans theoretical frameworks, empirical evaluations, and practical applications that align symbolic architectures with transformer-based LLMs.

Neuro-Symbolic AI Literature

- Hamilton et al. (2022)[1] systematically assess the evolution and bottlenecks of neuro-symbolic systems in NLP. They highlight challenges like symbolic scalability and LLM limitations in logical inference.
- Belle (2024)[2] explores the philosophical and mathematical foundations of logic-driven AI, emphasizing that formal logic is essential for interpretability and verifiable reasoning in LLMs. The work draws connections between early expert systems and contemporary neuro-symbolic learning.
- Oltramari (2023, 2024)[3] proposes frameworks that integrate symbolic memory structures with neural reasoning agents. The 2023 paper introduces modular systems aligned with human cognitive models, while the 2024 follow-up provides early results on embedding procedural rules into LLM-like architectures.

Neuro-Symbolic AI Literature

- Colelough & Regli (2025)[4] provide a meta-review of neuro-symbolic architectures across multiple domains, including vision, language, and reasoning. They catalog over 60 architectures and categorize them based on their integration depth and cognitive fidelity.
- Roy et al. (2024)[5] propose a Common Model of Cognition-inspired wrapper for foundation models, demonstrating that it improves factual consistency in multi-step reasoning tasks. The framework introduces declarative memory and goal-oriented action planning.

Neuro-Symbolic AI Literature

- Hersche et al. (2024)[6] compare pure LLMs to neuro-symbolic systems on abstract tasks such as Raven's Matrices. Their findings show that symbolic hybrids outperform LLMs in systematic generalization and structure recognition, especially under constrained input.
- Waterworth (2024) presents meta-cognitive LLM agents that integrate symbolic semantic graphs to enhance retrieval and user alignment. The study includes ablation tests showing improved accuracy and reduced hallucination rates when feedback-driven correction is enabled.

LLM Hallucinations

- Cleti and Jano [15] categorise LLM hallucinations into 4 categories : intrinsic hallucinations, extrinsic hallucinations, amalgamated hallucinations and non-factual hallucinations ; with sub-types including input-conflicting, context-conflicting, logical and semantic hallucinations.

LLM Hallucinations

Hallucination Type	**Definition**	**Underlying Mechanism/Cause**	**Example**
Intrinsic Hallucinations	Outputs inconsistent with the model's internal knowledge (training data or general world knowledge).	Failure to accurately retrieve or represent information from parametric memory.	Changing "boy" to "girl" or "creek" to "river" during translation.
Extrinsic Hallucinations	Outputs inconsistent with provided external context or world knowledge, even if consistent with training data.	Misinterpretation or failure to incorporate given context or prompt correctly; generating novel content to fill knowledge gaps.	Model output conflicting with external factual information.
Amalgamated Hallucinations	Incorrect combination of multiple facts or conditions presented in a prompt.	Failure to properly integrate disparate pieces of information, resulting in a blended erroneous output.	Merging several conditions from a prompt into a single, incorrect statement.
Non-Factual Hallucinations	Broad category for content contradicting established facts or world knowledge.	Gaps in subject-specific knowledge; failure in information extraction.	"Fact-conflicting hallucinations" (inaccurate statements).
Sub-type: Input-Conflicting	Output deviates from user input without being factually false.	Model misinterprets user intent.	User asks for "truck," model responds with "car."
Sub-type: Context-Conflicting	Output is inconsistent with previous outputs in a multi-turn conversation.	Semantic drift or difficulty maintaining consistency over turns.	Model states ocean is 139M sq miles, then 140M sq miles in subsequent turn.
Sub-type: Logical Hallucinations	Failure in logical reasoning.	Probabilistic models and pattern matching lacking true cognitive reasoning.	Generating coherent but illogical sequences.
Sub-type: Semantic Hallucinations	Semantic drift over multiple turns.	Model's understanding of user intention changes.	Gradual shift in topic or meaning over a conversation.

LLM Hallucinations

- The authors, in their paper, mentioned the underlying mechanisms of hallucinations :
  - Knowledge Overshadowing : This phenomenon occurs when certain aspects of a prompt dominate the model's attention, leading to hallucinations.
  - Insufficient Knowledge Representation : Hallucinations can arise from deficiencies in the lower layers of the model's neural network where the model generates unsupported information due to gaps in its subject-specific knowledge.
  - Failure in Information Extraction : This refers to the model's inability to accurately extract relevant attributes or details.
  - Contextual Misalignment : Hallucinations can occur when the model generates outputs misaligned with the provided context.
  - Semantic Entropy : This concept has been proposed to detect hallucinations caused by knowledge gaps, indicating when the model's output deviates from expected factual content.

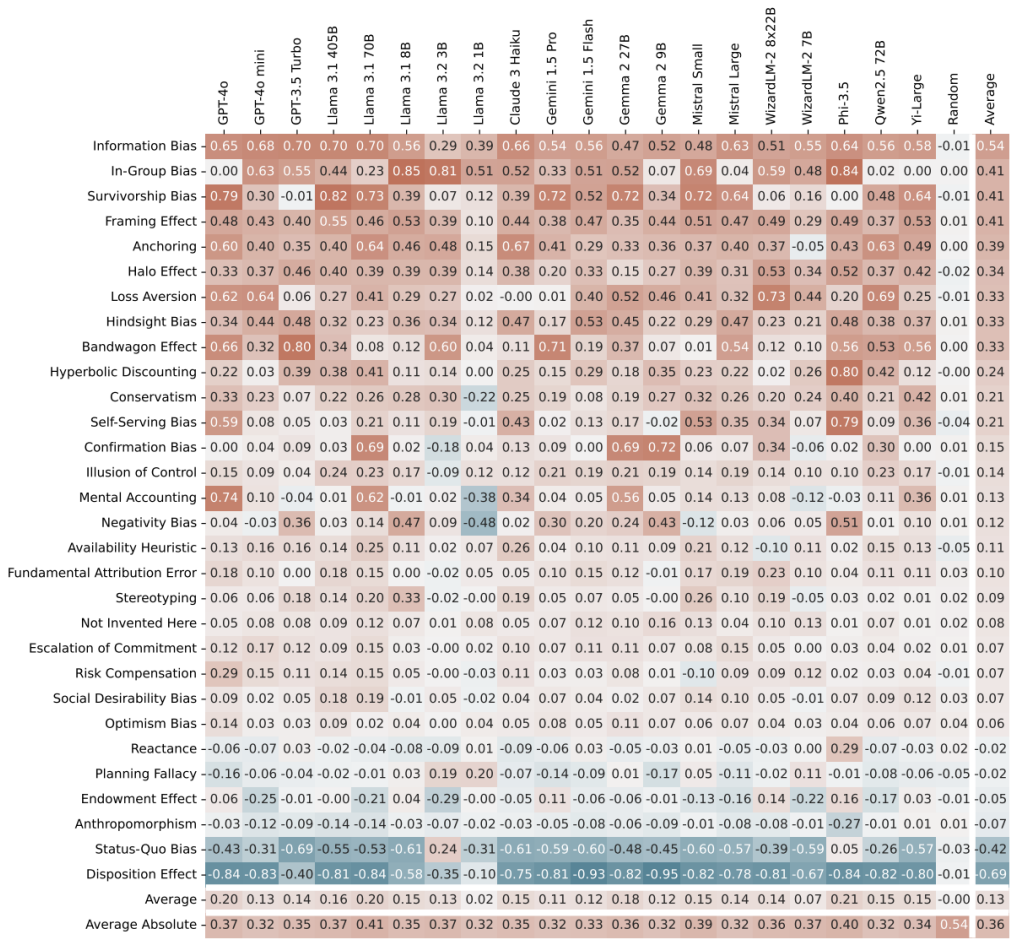
LLM Hallucinations

- The authors also inspected the factors influencing hallucinations :
  - Training Data Quality and Diversity : Imbalances or biases in the training data are critical factors, particularly in scenarios that are under-represented during training, leading to hallucinations.
  - Model Architecture and Size : While larger models generally possess more comprehensive knowledge, they may also exhibit a higher propensity for hallucination due to increased complexity.
  - Task Complexity : More complex or ambiguous queries posed to the LLM are more likely to provoke hallucinated responses.
  - Dynamic Nature of Language Model Reasoning : Recent studies suggest that the reasoning processes within LLMs are dynamic, contributing to the complexity of understanding and mitigating hallucinations.

Cognitive Biases in LLMs

- Echterhoff et al. [16] demonstrated that LLMs, despite lacking human cognitive structures, exhibit biases that functionally resemble these of human cognitive patterns. They proposed BiasBuster, a framework that encapsulates quantitative evaluation and automatic mitigation procedures for human-like cognitive bias.
- Malberg et al. [17] evaluated 30 cognitive biases in 20 LLMs under various decision-making scenarios ; contributing a general-purpose test framework for large-scale generation of tests for LLMs, and a benchmark dataset with 30,000 tests for detecting cognitive biases in LLMs.

Cognitive Biases in LLMs



	GPT-4o	GPT-4o mini	GPT-3.5 Turbo	Llama 3.1 405B	Llama 3.1 70B	Llama 3.1 8B	Llama 3.2 3B	Llama 3.2 1B	Claude 3 Haiku	Gemini 1.5 Pro	Gemini 1.5 Flash	Gemma 2 27B	Gemma 2 9B	Mistral Small	Mistral Large	WizardLM-2 8x22B	WizardLM-2 7B	Phi-3.5	Qwen2.5 72B	Yi-Large	Random	Average
Information Bias	0.65	0.68	0.70	0.70	0.70	0.56	0.29	0.39	0.66	0.54	0.56	0.47	0.52	0.48	0.63	0.51	0.55	0.64	0.56	0.58	-0.01	0.54
In-Group Bias	0.00	0.63	0.55	0.44	0.23	0.85	0.81	0.51	0.52	0.33	0.51	0.52	0.07	0.69	0.04	0.59	0.48	0.84	0.02	0.00	0.00	0.41
Survivorship Bias	0.79	0.30	-0.01	0.82	0.73	0.39	0.07	0.12	0.39	0.72	0.52	0.72	0.34	0.72	0.64	0.06	0.16	0.00	0.48	0.64	-0.01	0.41
Framing Effect	0.48	0.43	0.40	0.55	0.46	0.53	0.39	0.10	0.44	0.38	0.47	0.35	0.44	0.51	0.47	0.49	0.29	0.49	0.37	0.53	0.01	0.41
Anchoring	0.60	0.40	0.35	0.40	0.64	0.46	0.48	0.15	0.67	0.41	0.29	0.33	0.36	0.37	0.40	0.37	-0.05	0.43	0.63	0.49	0.00	0.39
Halo Effect	0.33	0.37	0.46	0.40	0.39	0.39	0.39	0.14	0.38	0.20	0.33	0.15	0.27	0.39	0.31	0.53	0.34	0.52	0.37	0.42	-0.02	0.34
Loss Aversion	0.62	0.64	0.06	0.27	0.41	0.29	0.27	0.02	-0.00	0.01	0.40	0.52	0.46	0.41	0.32	0.73	0.44	0.20	0.69	0.25	-0.01	0.33
Hindsight Bias	0.34	0.44	0.48	0.32	0.23	0.36	0.34	0.12	0.47	0.17	0.53	0.45	0.22	0.29	0.47	0.23	0.21	0.48	0.38	0.37	0.01	0.33
Bandwagon Effect	0.66	0.32	0.80	0.34	0.08	0.12	0.60	0.04	0.11	0.71	0.19	0.37	0.07	0.01	0.54	0.12	0.10	0.56	0.53	0.56	0.00	0.33
Hyperbolic Discounting	0.22	0.03	0.39	0.38	0.41	0.11	0.14	0.00	0.25	0.15	0.29	0.18	0.35	0.23	0.22	0.02	0.26	0.80	0.42	0.12	-0.00	0.24
Conservatism	0.33	0.23	0.07	0.22	0.26	0.28	0.30	-0.22	0.25	0.19	0.08	0.19	0.27	0.32	0.26	0.20	0.24	0.40	0.21	0.42	0.01	0.21
Self-Serving Bias	0.59	0.08	0.05	0.03	0.21	0.11	0.19	-0.01	0.43	0.02	0.13	0.17	-0.02	0.53	0.35	0.34	0.07	0.79	0.09	0.36	-0.04	0.21
Confirmation Bias	0.00	0.04	0.09	0.03	0.69	0.02	-0.18	0.04	0.13	0.09	0.00	0.69	0.72	0.06	0.07	0.34	-0.06	0.02	0.30	0.00	0.01	0.15
Illusion of Control	0.15	0.09	0.04	0.24	0.23	0.17	-0.09	0.12	0.12	0.21	0.19	0.21	0.19	0.14	0.19	0.14	0.10	0.10	0.23	0.17	-0.01	0.14
Mental Accounting	0.74	0.10	-0.04	0.01	0.62	-0.01	0.02	-0.38	0.34	0.04	0.05	0.56	0.05	0.14	0.13	0.08	-0.12	-0.03	0.11	0.36	0.01	0.13
Negativity Bias	0.04	-0.03	0.36	0.03	0.14	0.47	0.09	-0.48	0.02	0.30	0.20	0.24	0.43	-0.12	0.03	0.06	0.05	0.51	0.01	0.10	0.01	0.12
Availability Heuristic	0.13	0.16	0.16	0.14	0.25	0.11	0.02	0.07	0.26	0.04	0.10	0.11	0.09	0.21	0.12	-0.10	0.11	0.02	0.15	0.13	-0.05	0.11
Fundamental Attribution Error	0.18	0.10	0.00	0.18	0.15	0.00	-0.02	0.05	0.05	0.10	0.15	0.12	-0.01	0.17	0.19	0.23	0.10	0.04	0.11	0.11	0.03	0.10
Stereotyping	0.06	0.06	0.18	0.14	0.20	0.33	-0.02	-0.00	0.19	0.05	0.07	0.05	-0.00	0.26	0.10	0.19	-0.05	0.03	0.02	0.01	0.02	0.09
Not Invented Here	0.05	0.08	0.08	0.09	0.12	0.07	0.01	0.08	0.05	0.07	0.12	0.10	0.16	0.13	0.04	0.10	0.13	0.01	0.07	0.01	0.02	0.08
Escalation of Commitment	0.12	0.17	0.12	0.09	0.15	0.03	-0.00	0.02	0.10	0.07	0.11	0.11	0.07	0.08	0.15	0.05	0.00	0.03	0.04	0.02	0.01	0.07
Risk Compensation	0.29	0.15	0.11	0.14	0.15	0.05	-0.00	-0.03	0.11	0.03	0.03	0.08	0.01	-0.10	0.09	0.09	0.12	0.02	0.03	0.04	-0.01	0.07
Social Desirability Bias	0.09	0.02	0.05	0.18	0.19	-0.01	0.05	-0.02	0.04	0.07	0.04	0.02	0.07	0.14	0.10	0.05	-0.01	0.07	0.09	0.12	0.03	0.07
Optimism Bias	0.14	0.03	0.03	0.09	0.02	0.04	0.00	0.04	0.05	0.08	0.05	0.11	0.07	0.06	0.07	0.04	0.03	0.04	0.06	0.07	0.04	0.06
Reactance	-0.06	-0.07	0.03	-0.02	-0.04	-0.08	-0.09	0.01	-0.09	-0.06	0.03	-0.05	-0.03	0.01	-0.05	-0.03	0.00	0.29	-0.07	-0.03	0.02	-0.02
Planning Fallacy	-0.16	-0.06	-0.04	-0.02	-0.01	0.03	0.19	0.20	-0.07	-0.14	-0.09	0.01	-0.17	0.05	-0.11	-0.02	0.11	-0.01	-0.08	-0.06	-0.05	-0.02
Endowment Effect	0.06	-0.25	-0.01	-0.00	-0.21	0.04	-0.29	-0.00	-0.05	0.11	-0.06	-0.06	-0.01	-0.13	-0.16	0.14	-0.22	0.16	-0.17	0.03	-0.01	-0.05
Anthropomorphism	-0.03	-0.12	-0.09	-0.14	-0.14	-0.03	-0.07	-0.02	-0.03	-0.05	-0.08	-0.06	-0.09	-0.01	-0.08	-0.08	-0.01	-0.27	-0.01	0.01	0.01	-0.07
Status-Quo Bias	-0.43	-0.31	-0.69	-0.55	-0.53	-0.61	0.24	-0.31	-0.61	-0.59	-0.60	-0.48	-0.45	-0.60	-0.57	-0.39	-0.59	0.05	-0.26	-0.57	-0.03	-0.42
Disposition Effect	-0.84	-0.83	-0.40	-0.81	-0.84	-0.58	-0.35	-0.10	-0.75	-0.81	-0.93	-0.82	-0.95	-0.82	-0.78	-0.81	-0.67	-0.84	-0.82	-0.80	-0.01	-0.69
Average	0.20	0.13	0.14	0.16	0.20	0.15	0.13	0.02	0.15	0.11	0.12	0.18	0.12	0.15	0.14	0.14	0.07	0.21	0.15	0.15	-0.00	0.13
Average Absolute	0.37	0.32	0.35	0.37	0.41	0.35	0.37	0.32	0.35	0.32	0.33	0.36	0.32	0.39	0.32	0.36	0.37	0.40	0.32	0.34	0.54	0.36

Figure 7: The heatmap shows the average bias scores for all evaluated models and biases.

Cognitive Biases in LLMs

- Lyu et al. [18] proposed a cognitive debiasing approach, selfadaptive cognitive debiasing (SACD), that enhances the reliability of LLMs by iteratively refining prompts for cognitive biases. They focus on availability bias, bandwagon bias and loss aversion and test their framework on different domains such as law and healthcare.
- Malberg et al. [17] evaluated 30 cognitive biases in 20 LLMs under various decision-making scenarios ; contributing a general-purpose test framework for large-scale generation of tests for LLMs, and a benchmark dataset with 30,000 tests for detecting cognitive biases in LLMs.

Cognitive Phenomena : Confabulation

- Definition : The generation of fluent but fabricated information that is not grounded in the input context. It mirrors the human phenomenon of filling memory gaps with plausible, but false, details.
- Mechanisms : In LLMs, confabulation arises when the model uses high-probability patterns to answer questions lacking sufficient grounding. This can be driven by training distribution skew, poor alignment, or token-level optimization.

Cognitive Phenomena : Confabulation

- Sui et al. (2024)[7] redefine hallucinations as confabulationsâfluent but fabricated responsesâand argue that such outputs may aid creative or narrative generation. They introduce an evaluation corpus annotated for "plausibility vs. grounding."
- Smith et al. (2023) establish neurological analogies to LLM hallucinations, proposing that confabulation arises from generative fluency without verification layers.
- Rawte et al. (2023)[8] create an extended hallucination taxonomy, validated on QA and summarization datasets.
- Liu et al. (2023) empirically measure internal model uncertainty and show misalignment between high-confidence tokens and actual truthfulness.

Cognitive Phenomena : Recency Effect

- Definition : A form of serial position bias where models overweight the final segments of input, leading to misinterpretation in multi-hop or long-context reasoning.
- Mechanisms : Transformer attention inherently favours recent tokens in large contexts, particularly under position-agnostic decoding.

Cognitive Phenomena : Recency Effect

- Guo & Vosoughi (2024)[9] validate recency bias in transformer attention across various datasets. They show task accuracy declines when vital context appears early in the input.
- Peysakhovich & Lerer (2023) propose "attention sorting" to reshuffle input to counteract recency. Their method improves retrieval-augmented generation (RAG) on long-input tasks.
- Horowitz & Plonsky (2025) : contrast human and LLM recency patterns, revealing architectural artifacts in transformer-based inference.

Cognitive Phenomena : Availability Heuristic

- Definition : The tendency to default to frequent or memorable patterns, leading to repetition of common knowledge even when contextually irrelevant.
- Mechanisms : LLM token probabilities are skewed toward frequently seen completions. Without constraint, this results in overgeneration of âsafeâ or âpopularâ phrases.

Cognitive Phenomena : Availability Heuristic

- Suri et al. (2023), Ross et al. (2024)[10, 11] demonstrate that GPT-3.5 favors familiar token sequences under uncertainty, replicating human availability bias ; and model these tendencies within utility theory frameworks, aligning LLM behavior with human cognitive ease.
- Echterhoff et al. (2024) operationalize this in the BiasBuster dataset, showing statistically significant overgeneration of popular phrases.

Cognitive Phenomena : Source Amnesia

- Definition : The failure to correctly attribute information to a source, even when the content is partially accurate.
- Mechanisms : Neural models decouple content from source during pretraining. In decoding, they recombine elements plausibly without preserving origin.

Cognitive Phenomena : Source Amnesia

- Khalifa et al. (2024)[12] show that adding source identifiers during training enables source attribution in output.
- Pavlick (2023)[13] conceptualizes grounding failures as a symbolic absence, not just model error.

Cognitive Phenomena : Loss Aversion

- Definition : A behavioral pattern where LLMs prefer producing any answerâeven incorrectârather than abstaining or expressing uncertainty.
- Mechanisms : Fine-tuning incentives prioritize answerability over null prediction, especially on single-reference tasks.

Cognitive Phenomena : Loss Aversion

- Jia et al. (2024), Ross et al. (2024) develop a decision framework where LLMs under uncertainty prefer giving any answer rather than abstaining, echoing human loss aversion.
- Suri et al. (2023) demonstrate hallucination rates rise when models are not penalized for incorrect answers.

References I


Hamilton, K., Nayak, A., Božić, B., & Longo, L. (2022, 2024). Is Neuro-Symbolic AI Meeting its Promise in NLP ?


Belle, V. (2024). On the relevance of logic for AI.


Oltramari, A. (2023, 2024). A Path Towards High-Level Reasoning Through Cognitive Neuro-Symbolic Systems.


Colelough, B.C. & Regli, W. (2025). Neuro-Symbolic AI in 2024 : A Systematic Review.


Roy, K., Wu, S., & Oltramari, A. (2024). Neurosymbolic Cognitive Methods.


Hersche, M., Camposampiero, G., & Wattenhofer, R. (2024). Towards Learning to Reason.


Sui, P., Duede, E., Wu, S., & So, R. J. (2024). Confabulation in LLMs.


Rawte, V., Chakraborty, S., Pathak, A., et al. (2023). Hallucination Taxonomy.


Guo, X., & Vosoughi, S. (2024). Serial Position Effects in LLMs.


Suri, G., Slater, L. R., Ziaee, A., & Nguyen, M. (2023). Decision Heuristics in LLMs.


Ross, J., Kim, Y., & Lo, A. W. (2024). Behavioral Biases in LLMs.


Khalifa, M., et al. (2024). Source-Aware Training for Knowledge Attribution.

References II


Pavlick, E. (2023). Symbol Grounding in LLMs.


Lior, G., Nacchace, L., & Stanovsky, G. (2025). WildFrame Dataset.


Cleti, M., Jano, P. (2024). Hallucinations in LLMs : Types, Causes, and Approaches for Enhanced Reliability


Echterhoff, J., Liu, Y., Alessa, A., McAuley, J., He, Z. (2024). Cognitive Bias in Decision-Making with LLMs


Malberg, S., Poletukhin, R., Schuster, C.M., Groh, G. (2025). A Comprehensive Evaluation of Cognitive Biases in LLMs


Lyu, Y., Ren, S., Feng, Y., Wang, Z., Chen, Z., Ren, Z., Rijkel, M. (2025). Cognitive Debiasing Large Language Models for Decision-Making