A Survey of Neurosymbolic Approaches to Propaganda and Persuasion Technique Detection

Alysa Ziying Tan
Nanyang Technological University
s190109@ntu.edu.sg

Daniel Kainovan Handoyo
University of Edinburgh
D.K.Handoyo@sms.ed.ac.uk

Abstract

Propaganda and persuasion technique detection, the task of identifying how misleading content persuades rather than merely what it claims, occupies a distinctive and underexplored position within computational misinformation research. This task requires interpretable and rhetorically grounded reasoning over persuasion techniques, making it particularly well aligned with neurosymbolic artificial intelligence (AI). This survey presents the first systematic review of neurosymbolic approaches for propaganda and persuasion technique detection, together with relevant structured-reasoning approaches from closely related areas that offer methodological foundations for these tasks. It synthesizes work spanning argument mining, knowledge-augmented learning, logic-based fallacy detection, structured reasoning with large language models (LLMs), and hybrid neural-symbolic architectures. We further organize existing methods through a unified taxonomy of neurosymbolic integration paradigms and review foundational rhetoric theory, benchmark datasets, and evaluation methodologies. We identify key open challenges, including subtle and implicit persuasion, cross-lingual generalization, explanation faithfulness, and the growing challenge of LLM-generated propaganda. Finally, we outline future research directions and argue that neurosymbolic AI provides a promising foundation for bridging predictive performance with human-interpretable rhetorical reasoning in next-generation misinformation detection systems.

1 Introduction

The proliferation of online misinformation has driven sustained research interest in automated detection systems. Much of this effort has centered on *fact verification*: determining whether a claim is supported, refuted, or unverifiable given evidence (Thorne et al., 2018; Zeng et al., 2021; Guo et al., 2022). However, factuality alone is often insufficient for understanding how misinformation operates in practice. Much manipulative communication is implicit, shaping beliefs and behavior through rhetorical framing, emotional appeals, selective emphasis, and other discourse strategies rather than outright false claims (Rashkin et al., 2017; Da San Martino et al., 2019; Pennycook et al., 2021). This has motivated growing attention toward the detection of *propaganda* and *persuasion techniques*, with a focus extending beyond what the content claims to how it attempts to influence its audience.

Propaganda detection and persuasion technique detection are closely related but conceptually distinct tasks. Propaganda detection is commonly formulated as determining whether content is propagandistic, while persuasion technique detection identifies the rhetorical mechanisms underlying influence and where they occur within text (Barrón-Cedeno et al., 2019; Da San Martino et al., 2020). For example, given the statement "Vaccines are deliberately used to control the population," a propaganda detection system may classify the statement as propagandistic, whereas a persuasion analysis system seeks to identify the underlying rhetorical mechanisms, such as *Appeal to Fear* or *Loaded Language*. Existing taxonomies include techniques such as *Name Calling / Labeling*, *Flag-Waving*, *Loaded Language*, and argumentative strategies including *Whataboutism*, *Red Herring*, and *Straw Man*, many of which originate from rhetoric, political communication, and argumentation theory (Da San Martino et al., 2019, 2020; Dimitrov et al., 2021; Piskorski et al., 2023). Throughout this survey, we use *propaganda and persuasion technique detection* as an umbrella term for methods that identify such rhetorical manipulation strategies.

This distinction between detecting *what is false* and understanding *how persuasion operates* has important implications for computational modeling. Unlike fact verification, which is commonly formulated as evidence retrieval and natural language inference (NLI) (Thorne et al., 2018; Zeng et al., 2021; Fadeeva et al., 2024), persuasion analysis fundamentally depends on linguistic and rhetorical reasoning. Techniques such as *False Dilemma*, *Whataboutism*, and *Appeal to Fear* rely on argumentative structure, discourse organization, pragmatic implication, or audience interpretation rather than surface lexical patterns alone. More broadly, persuasion analysis requires reasoning over discourse structure, argumentation, pragmatics, commonsense knowledge, and symbolic rhetorical categories (Lawrence and Reed, 2019; Grice, 1975; William and Thompson, 1988). Figure 1 summarizes these reasoning requirements and motivates why propaganda and persuasion technique detection is particularly well aligned with neurosym-

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 2688–2699
November 6-10, 2026 ©2026 Association for Computational Linguistics

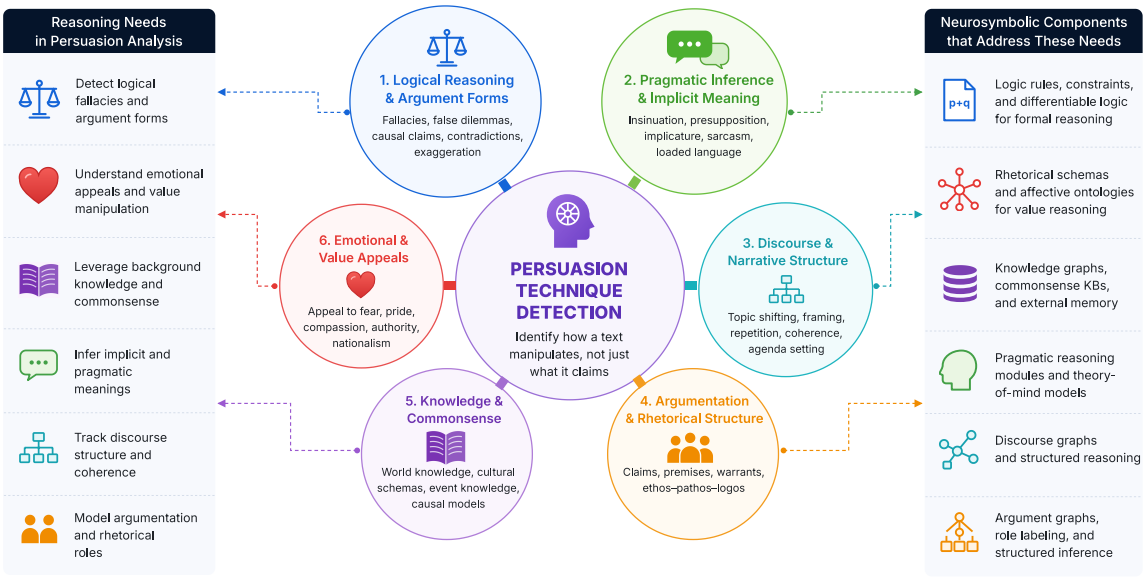


Figure 1: Reasoning requirements underlying propaganda and persuasion technique detection and corresponding neurosymbolic components that address them. Persuasion analysis extends beyond factual verification and surface semantic matching to require reasoning over logical structure, pragmatics, discourse organization, argumentation, commonsense knowledge, and emotional framing, naturally motivating neurosymbolic approaches.

bolic AI.

Why Neurosymbolic AI? Neurosymbolic AI (NeSy) aims to combine neural representation learning with symbolic reasoning mechanisms such as logic rules, discourse structures, argument graphs, ontologies, and knowledge-guided inference (Susskind et al., 2021; Garcez and Lamb, 2023; Bhuyan et al., 2024). Propaganda and persuasion analysis naturally aligns with this paradigm because persuasion techniques are typically organized into explicit taxonomies, many correspond to identifiable rhetorical or argumentative structures, and effective analysis often benefits from interpretable reasoning processes rather than black-box prediction alone.

Recent advances in transformer architectures and large language models (LLMs) have substantially improved propaganda and persuasion analysis (Yu et al., 2022). Nevertheless, purely neural approaches continue to face important limitations. They often rely on shallow lexical correlations, struggle with implicit or compositional reasoning, exhibit limited robustness under domain shifts, and provide limited transparency for high-stakes settings (Atanasova et al., 2020). These limitations are particularly consequential for persuasion analysis, where interpretability and reasoning transparency are central requirements. NeSy approaches offer a promising direction by integrating structured knowledge, symbolic constraints, discourse-aware inference, and interpretable intermediate reasoning processes within neural systems.

Scope and Contributions of This Survey. Research at the intersection of propaganda analysis and neurosymbolic AI remains fragmented across multiple communities, including natural language processing, argument mining, computational rhetoric, explainable AI, and knowledge-guided reasoning. Several surveys examine adjacent areas relevant to this work. Automated fact verification and fact-checking have been reviewed extensively by Zhou and Zafarani (2020), Zeng et al. (2021), and Guo et al. (2022), while misinformation and fake news detection are surveyed by Bondielli and Marcelloni (2019). Related reasoning-oriented areas have also received dedicated attention, including argument mining (Lawrence and Reed, 2019) and explainable fact verification (Kotonya and Toni, 2020). More broadly, neurosymbolic AI has been surveyed as a general paradigm integrating neural learning with symbolic reasoning (Susskind et al., 2021; Garcez and Lamb, 2023; DeLong et al., 2024; Bhuyan et al., 2024). Recent surveys of computational persuasion and online persuasion review broader ML/NLP methods, persuasive language, argumentative perspectives, and explainability (Bassi et al., 2024; Bozdag et al., 2026). In contrast, this survey focuses more narrowly on propaganda and persuasion technique detection through structured reasoning mechanisms and neurosymbolic integration.

To the best of our knowledge, this survey is the first to systematically examine NeSy approaches for propaganda and persuasion technique detection. We further provide the first unified treatment connecting propaganda analysis with argumentation theory, structured reasoning, and modern NeSy frameworks. We review methods incorporating symbolic knowledge, explicit reasoning mechanisms, or hybrid neural-symbolic architectures, spanning rule-guided learning,

knowledge-augmented systems, graph-based reasoning, and reasoning-oriented LLM frameworks. We additionally compare benchmark datasets and evaluation protocols, introduce a unified taxonomy of NeSy integration paradigms, and identify key open challenges and future directions for building more interpretable, robust, and reasoning-aware persuasion analysis systems.

2 Background

2.1 Rhetoric, Propaganda, and Persuasion Taxonomies

The computational study of propaganda and persuasion is grounded in longstanding traditions in rhetoric and argumentation theory. Cope et al. (1877)'s *Rhetoric* distinguished three persuasive appeals: *ethos* (credibility), *pathos* (emotion), and *logos* (reasoning) (Cope et al., 1877). These foundations continue to shape contemporary understandings of persuasive communication, as many propaganda techniques operate through emotional appeals, perceived credibility, and argumentative structure.

Early systematic taxonomies emerged through the Institute for Propaganda Analysis (IPA) (Lee and Lee, 1939), which codified seven canonical techniques in the 1930s: *Name Calling*, *Glittering Generalities*, *Transfer*, *Testimonial*, *Plain Folks*, *Card Stacking*, and *Bandwagon*. While historically influential, these categories are relatively coarse for modern computational modeling, where systems must identify fine-grained spans and distinguish closely related techniques. Contemporary work additionally draws on argumentation theory and informal logic (Hamblin, 2022; Macagno, 2025), alongside broader persuasion science (Cialdini, 2018).

Within NLP, the most influential fine-grained benchmark taxonomy originates from Da San Martino et al. (2019), who introduced the Propaganda Techniques Corpus (PTC), annotated at the fragment level with 18 propaganda techniques. These include affective and evaluative devices such as *Loaded Language*, *Name Calling / Labeling*, and *Appeal to Fear / Prejudice*, alongside argumentation-oriented techniques such as *Causal Oversimplification*, *Black-and-White Fallacy*, and *Whataboutism, Red Herring, and Straw Man*. This taxonomy informed SemEval-2020 Task 11 (Da San Martino et al., 2020), while subsequent benchmarks extended persuasion analysis toward multilingual, cross-domain, and multimodal settings (Piskorski et al., 2023; Dimitrov et al., 2024; Piskorski et al., 2025).

These taxonomies exhibit a property particularly relevant to neurosymbolic modeling because they define persuasion techniques using rhetorical, argumentative, and logical criteria that extend beyond statistical regularities. Techniques such as *Whataboutism*, *Black-and-White Fallacy*, and *Appeal to Fear / Prejudice* correspond to recognizable discourse or reasoning patterns, making propaganda and persuasion analysis naturally compatible with symbolic representations and structured reasoning frameworks.

2.2 Tasks and Benchmark Datasets

Propaganda Technique Corpus (PTC). Introduced by Da San Martino et al. (2019), the PTC remains the foundational benchmark for fine-grained propaganda analysis. It consists of 451 news articles annotated at both the sentence level (propaganda versus non-propaganda) and span level (18 persuasion techniques). The dataset highlights two persistent challenges for computational modeling: (i) substantial class imbalance, with propaganda spans constituting a minority of text, and (ii) ambiguity in annotating subtle rhetorical phenomena disagreement on subtle techniques.

SemEval-2020 Task 11. Da San Martino et al. (2020) formalized propaganda analysis through two subtasks: *Span Identification (SI)*, which detects propagandistic text fragments, and *Technique Classification (TC)*, which identifies the persuasion technique expressed by each span. The benchmark established standardized evaluation protocols and remains the primary evaluation setting for fine-grained persuasion technique detection.

Multimodal and Multilingual Extensions. Subsequent benchmarks expanded beyond English news text. SemEval-2021 Task 6 (Dimitrov et al., 2021) introduced multimodal persuasion analysis in internet memes, requiring joint reasoning over textual and visual content. SemEval-2024 Task 4 (Dimitrov et al., 2024) further advanced this direction through multilingual meme-based persuasion detection with textual and multimodal subtasks. SemEval-2023 Task 3 (Piskorski et al., 2023) extended persuasion analysis toward multilingual and cross-domain settings spanning multiple languages and genres, while the SlavicNLP 2025 shared task (Piskorski et al., 2025) broadened evaluation to parliamentary debates and social media across five Slavic languages. These datasets create new challenges for robustness, cross-lingual generalization and domain transfer.

Document-Level and Upstream Resources. Beyond span-level datasets, Barrón-Cedeno et al. (2019) introduced the Proppy corpus for document-level propaganda detection. Related shared tasks such as CLEF CheckThat! (Nakov et al., 2021) target upstream components including claim detection and check-worthiness estimation, which frequently serve as earlier stages within broader misinformation processing pipelines.

Collectively, these benchmarks reflect a progression toward increasingly fine-grained, multilingual, and multimodal persuasion analysis, while exposing core challenges for NeSy methods, including implicit reasoning, long-range discourse understanding, interpretability, and robustness across domains and languages.

2.3 Neurosymbolic AI: Scope and Working Definition

Neurosymbolic AI broadly refers to computational approaches that integrate neural learning with symbolic representations or reasoning mechanisms (Garcez and Lamb, 2023). In this survey, we adopt a functional rather than doctrinal definition of NeSy, focusing on systems that combine neural language understanding with explicit structured reasoning or interpretable intermediate representations that substantively influence inference. Symbolic components may include logic rules, rhetorical schemas, discourse structures, argument graphs, ontologies, knowledge graphs, or structured reasoning traces.

This definition encompasses not only classical hybrid neural-symbolic architectures (Garcez et al., 2015), but also modern reasoning-oriented large language model (LLM) frameworks that generate explicit intermediate structures, such as chain-of-thought rationales, argument decompositions, or graph-based reasoning representations, rather than relying solely on end-to-end black-box prediction (Wei et al., 2022; Yao et al., 2022; Zhang et al., 2024; Xu et al., 2025b; Sun et al., 2025). Conversely, purely neural classifiers operating entirely through latent representations fall outside the primary scope of this survey, except where included as baselines or points of comparison.

Under this perspective, NeSy approaches are particularly well-aligned with propaganda and persuasion technique detection, where effective analysis often depends on reasoning over logical structure, discourse organization, pragmatic implication, rhetorical strategies, and world knowledge rather than surface lexical patterns alone.

3 Linguistic and Rhetorical Reasoning Foundations

Propaganda and persuasion technique detection is fundamentally a reasoning task that extends beyond surface semantic classification. Unlike standard text classification, persuasion analysis often requires understanding how rhetorical structure, discourse organization, and contextual framing shape audience interpretation (Contalbo et al., 2024; Turk et al., 2025; Bozdag et al., 2026). Many persuasion techniques are not defined by explicit falsehoods or lexical cues, but by the logical, pragmatic, emotional, and narrative functions they perform. For example, *Whataboutism* relies on discourse diversion, *False Dilemma* imposes an artificially constrained argumentative structure, and *Loaded Language* operates through connotation rather than propositional content. Effective persuasion analysis therefore frequently depends on reasoning over argumentation, discourse relations, pragmatic inference, and world knowledge.

Rather than viewing persuasion analysis solely as classification, we frame it as a multi-layered reasoning problem involving logical and argumentative structure, pragmatic and implicit communication, discourse organization, and commonsense knowledge. These foundations motivate NeSy approaches and provide the conceptual basis for the methods surveyed in subsequent sections.

3.1 Logical and Argumentative Structures

Many persuasion techniques correspond to identifiable argumentative patterns rather than isolated lexical features. Techniques such as *Straw Man*, *False Cause*, *Red Herring*, and *Whataboutism* derive meaning from how claims are framed, contrasted, justified, or causally linked (Lawrence and Reed, 2019; Hamblin, 2022; Macagno, 2025). For example, *False Dilemma* manipulates reasoning by presenting artificially limited alternatives, while *Whataboutism* redirects attention through argumentative diversion.

These characteristics connect persuasion analysis to argument mining and computational fallacy detection (Lippi and Torroni, 2016). Argument mining identifies claims, premises, and support or attack relations, while fallacy detection targets flawed reasoning patterns. This motivates symbolic representations such as logical rules, argument graphs, and rhetorical schemas for modeling persuasion more explicitly and interpretably.

3.2 Pragmatic and Implicit Persuasion

Many persuasion techniques operate through pragmatic inference rather than explicit semantic content. Emotional framing, value signaling, presupposition, insinuation, sarcasm, and implied causality often depend on shared social knowledge and audience interpretation (Grice, 1975; Cialdini, 2018). For example, *Appeal to Fear* may imply negative consequences without explicitly stating them, while techniques such as *Flag-Waving* and *Appeal to Authority* derive persuasive force from identity, values, or perceived legitimacy rather than propositional content alone. Similarly, *Loaded Language* operates through affective framing and connotation (Da San Martino et al., 2019).

Such phenomena challenge purely distributional language models because persuasive intent is often implicit, context-dependent, culturally grounded, and intertwined with emotional or value-based interpretation (Grice, 1975). Effective analysis may require reasoning about unstated assumptions, speaker intent, audience expectations, commonsense knowledge, and broader social or cultural context (Speer et al., 2017).

3.3 Discourse and Narrative Structure

Propaganda frequently operates at the discourse level rather than within isolated sentences. Techniques such as *Repetition*, *Flag-Waving*, and *Causal Oversimplification* often emerge through patterns distributed across longer contexts, requiring models to track narrative framing, topic transitions, and discourse coherence (William and Thompson, 1988; Taboada and Mann, 2006).

These challenges connect persuasion analysis to discourse parsing, framing analysis, and narrative understanding (William and Thompson, 1988; Boydstun, 2013; Piper et al., 2021). Narrative framing can selectively emphasize actors, causal relationships, or perspectives in ways that shape audience interpretation. Such observations motivate discourse representations, graph-based reasoning, and hierarchical symbolic structures for persuasion analysis.

3.4 Commonsense and World Knowledge

Persuasion often depends on commonsense and world knowledge, including cultural narratives and socially shared assumptions. Statements intended to provoke fear, patriotism, distrust, or outrage frequently rely on beliefs shared between speaker and audience, making reasoning beyond literal text essential (Grice, 1975).

This challenge is particularly important for subtle and context-dependent persuasion. Conspiracy narratives, for example, frequently rely on implied causal chains and latent assumptions that remain unstated (Van Prooijen, 2018). Knowledge graphs, ontologies, commonsense reasoning systems, and structured world representations therefore provide promising mechanisms for enhancing persuasion analysis (Speer et al., 2017; Hitzler and Sarker, 2022).

Taken together, these logical, pragmatic, discourse-level, and knowledge-driven characteristics highlight why persuasion analysis is fundamentally a structured reasoning problem. Models must reason not only over linguistic patterns, but also argumentative structure, rhetorical intent, contextual implication, and world knowledge.

4 Neurosymbolic Approaches for Propaganda and Persuasion Analysis

Recent work increasingly explores NeSy approaches that integrate neural language understanding with structured reasoning mechanisms. These approaches aim to improve interpretability, robustness, and reasoning beyond latent neural representations. Different paradigms support different reasoning capabilities and system properties (Table 1). Rule-guided systems emphasize logical consistency and interpretability, graph-based methods support discourse and narrative reasoning, knowledge-augmented approaches improve contextual grounding, and reasoning-oriented LLM frameworks provide flexible pragmatic and commonsense reasoning. These paradigms underscore the need for multiple interacting forms of reasoning in persuasion analysis

4.1 Logic- and Rule-Guided Neural Learning

Logic-guided approaches incorporate symbolic rules, logical constraints, or rhetorical heuristics into neural pipelines to regularize model behavior and enforce argumentative or rhetorical consistency (Garcez et al., 2015; Hu et al., 2016; Pauk and Pacheco, 2024). Constraints may encode relationships between persuasion techniques, discourse patterns, or argumentative components.

Early systems relied heavily on linguistic patterns and lexicons (Rashkin et al., 2017), while recent methods combine neural architectures with differentiable logic, posterior regularization, or soft constraint optimization (Garcez and Lamb, 2023; Pan et al., 2023). These mechanisms can encourage consistency between argumentative relations and persuasion labels or enforce dependencies between rhetorical categories.

These approaches are conceptually well suited to techniques with explicit argumentative structure, such as *False Dilemma*, *Straw Man*, and *Whataboutism*, although direct comparative evidence remains limited. Their primary strengths lie in interpretability and logical transparency, while explicit constraints can provide moderate robustness against superficial lexical variation. However, hand-crafted rules often generalize poorly across domains and subtle pragmatic contexts, limiting scalability to open-domain persuasion analysis.

4.2 Knowledge-Augmented and Ontology-Guided Approaches

Knowledge-augmented approaches incorporate external resources such as knowledge graphs, embeddings, ontologies, and commonsense reasoning systems to provide contextual grounding beyond surface text (Cao et al., 2021; Ibrahim et al., 2024; Faye et al., 2026). Resources such as ConceptNet (Speer et al., 2017), Wikidata (Vrandečić and Krötzsch, 2014), and structured knowledge representations enable models to reason over entities, events, ideological narratives, and latent causal structure. Ontological representations additionally support hierarchical persuasion taxonomies, consistency-aware reasoning, and interpretable inference (Garcez and Lamb, 2023).

Compared to rule-guided systems, knowledge-augmented approaches generally provide stronger commonsense grounding and broader contextual flexibility while maintaining moderate interpretability through explicit knowledge structures. Their robustness depends heavily on knowledge coverage and quality, while scalability is constrained by knowledge acquisition and integration costs.

4.3 Graph-based and Structured Reasoning Architectures

Graph-based approaches model interactions between entities, claims, arguments, discourse units, or narrative components through structured relational representations (William and Thompson, 1988; Lawrence and Reed, 2019). Unlike flat sequence models, they explicitly capture long-range dependencies and discourse organization. Common structures include argument graphs, discourse dependency graphs, heterogeneous information networks, and entity-relation graphs (Arisaka et al., 2022). Graph neural networks (GNNs) (Wu

Approach	Logical / Arg.	Pragmatic / Implicit	Discourse / Narrative	Commonsense / World Know.	Interpretability	Robustness	Scalability
Rule-/Logic-Guided	Very High	Low	Low	Low	Very High	Moderate	Low
Knowledge-Augmented	Moderate	Moderate	Low–Mod.	Very High	Moderate	Moderate	Moderate
Graph-Based Reasoning	High	Moderate	Very High	Moderate	Mod.–High	Moderate	Moderate
Hybrid Neural-Symbolic	Very High	Mod.–High	High	High	High	High	Moderate
Reasoning-Oriented LLM	Mod.–High	High	High	Very High	Moderate	Low–Mod.	Mod.–High
Pure Neural Models	Low–Mod.	Moderate	Moderate	Low–Mod.	Low	Low	Very High
	Very High	High	Mod.–High	Moderate	Low–Mod.	Low	

Table 1: Comparison of reasoning paradigms for propaganda and persuasion analysis across reasoning capabilities and system properties. The complementary strengths of different paradigms motivate hybrid NeSy approaches.

et al., 2020) enable information propagation across these structures, supporting multi-hop reasoning over interconnected discourse elements (Zhou et al., 2020).

These architectures are particularly effective for discourse and narrative reasoning and provide moderate interpretability through explicit relational structure. Their structured representations can further improve robustness by modeling dependencies beyond surface lexical patterns. However, graph construction often depends on upstream systems such as discourse parsers or argument mining pipelines (Lawrence and Reed, 2019), introducing structural ambiguity and error propagation. Scalability may also become challenging in large open-domain settings due to graph construction and reasoning overhead.

4.4 Hybrid Neural-Symbolic Frameworks

Hybrid neural-symbolic systems jointly integrate neural representation learning and symbolic reasoning within unified architectures rather than treating symbolic processing as external post-processing (Garcez et al., 2015; Garcez and Lamb, 2023; Yu et al., 2023; Cai et al., 2025). Representative approaches include differentiable reasoning modules, symbolic memory systems, neuro-symbolic program induction, and neural-symbolic verification pipelines (Yu et al., 2023; Wooding et al., 2026). In persuasion analysis, such systems may combine transformer encoders with reasoning over rhetorical schemas, logical constraints, or discourse structures.

Hybrid frameworks frequently provide strong balance across interpretability, logical consistency, discourse reasoning, robustness, and representational flexibility. Their integration of explicit symbolic reasoning can improve transparency and resilience beyond surface lexical correlations, though these benefits often come at the cost of greater architectural complexity, computational overhead, and moderate scalability challenges.

4.5 Reasoning-Oriented LLM Frameworks

LLMs introduce reasoning paradigms that increasingly blur the boundary between neural modeling and structured reasoning. Recent methods employ chain-of-thought prompting, self-reflection, tool use, debate mechanisms, and structured prompting strategies (Wei et al., 2022; Yao et al., 2022; Zhang et al., 2024; Xu et al., 2025b; Sun et al., 2025). Although not symbolic in the classical sense, these systems often generate intermediate structures such as rationale chains, argument decompositions, or discourse analyses. Such capabilities support pragmatic reasoning, commonsense inference, and open-domain contextual understanding while enabling adaptation to evolving persuasion strategies with limited task-specific engineering.

However, these reasoning processes remain largely statistically grounded rather than formally constrained. Generated rationales may appear plausible without faithfully reflecting underlying model behavior, raising concerns regarding hallucination, robustness, reasoning consistency, and explanation faithfulness (Kim et al., 2025; Salido et al., 2026). Multi-stage inference and long-context reasoning can further introduce computational overhead despite relatively strong cross-domain adaptability (Lin et al., 2025; Shojaee et al., 2026).

Taken together, recent advances increasingly frame persuasion analysis as a structured reasoning problem rather than shallow classification. Existing paradigms exhibit complementary strengths across logical, pragmatic, discourse, and commonsense reasoning dimensions, motivating continued development of hybrid and multi-level NeSy systems.

5 Explainability and Interpretability in Neurosymbolic Persuasion Analysis

Explainability is particularly important in propaganda and persuasion technique detection due to the subjective, rhetorical, and socially sensitive nature of the task. Unlike conventional text classification, persuasion analysis often involves subtle argumentative ma-

nipulation, emotional framing, and implicit rhetorical intent (Da San Martino et al., 2019). In domains such as political communication, journalism, and content moderation, predictions without meaningful justification may be insufficient for human oversight and trust (Kotonya and Toni, 2020). Interpretability is therefore a central requirement for responsible deployment.

NeSy approaches are especially attractive because symbolic structures such as logical rules, argument graphs, discourse representations, and explicit reasoning traces naturally support more interpretable inference than purely neural black-box systems (Garcez and Lamb, 2023). Beyond producing persuasion labels, these systems can expose intermediate reasoning processes underlying predictions.

5.1 Rationales and Symbolic Explanations

Persuasion analysis systems employ diverse explanation paradigms. A common approach generates rationales highlighting influential words, phrases, or spans contributing to predictions. Attention visualizations, saliency maps, and rationale extraction methods improve local transparency by identifying evidence supporting model outputs (Atanasova et al., 2020; Jacovi and Goldberg, 2020). Recent explanation-enhanced resources such as PropXplain (Hasanain et al., 2025) further connect propaganda labels with natural-language rationales, illustrating how LLMs can be trained to generate both technique predictions and explanatory justifications.

NeSy systems additionally support structured explanations. Rule-guided approaches may expose symbolic constraints or rhetorical heuristics, graph-based methods can visualize argumentative or discourse relations, and hybrid architectures may generate intermediate representations such as logical constraints, argument graphs, or rhetorical schemas (Garcez and Lamb, 2023). Recent reasoning-oriented LLM frameworks further extend this paradigm through rationale chains, argument decompositions, and multi-step reasoning traces (Wei et al., 2022).

These explanation forms differ in granularity and interpretive depth. Span-level rationales primarily localize evidence, while symbolic explanations expose higher-level argumentative and rhetorical structure. This distinction is particularly important as persuasive effects often emerge through discourse organization, emotional framing, and interactions between argumentative components rather than isolated lexical cues.

5.2 Human-Centered Evaluation

As persuasion analysis is context-dependent and inherently subjective, explanation quality often requires human evaluation. Existing work commonly employs expert annotation, user studies, and human evaluation protocols assessing usefulness, coherence, interpretability, and alignment with human reasoning (Jacovi and Goldberg, 2020). Human-centered evaluation is particularly important because persuasion techniques may admit multiple valid interpretations across cultural, political, or social contexts. Related techniques such as *Appeal to Fear*, *Loaded Language*, and *Exaggeration* frequently exhibit ambiguous boundaries, contributing to annotator disagreement and evaluation challenges (Piskorski et al., 2023; Bassi et al., 2025).

Evaluation requirements further vary across stakeholders. Expert analysts may prioritize reasoning transparency and argumentative detail, whereas general users may require simpler explanations. Effective persuasion analysis systems may therefore require explanation mechanisms adapted to different users and decision-making settings.

5.3 Faithfulness and Reasoning Transparency

A central challenge concerns whether explanations faithfully reflect the model's actual reasoning process. This issue is particularly important for reasoning-oriented LLM frameworks, where coherent explanations may not correspond to the underlying inference mechanism (Jacovi and Goldberg, 2020; Shen et al., 2026). In many cases, generated rationales function as post-hoc justifications rather than causally grounded reasoning traces. This limitation is especially problematic for persuasion analysis, where users may rely on explanations to understand why rhetoric is considered manipulative. Explanation fluency should therefore not be conflated with reasoning transparency.

NeSy approaches attempt to address this challenge by incorporating explicit symbolic structures and intermediate representations that expose inference pathways more directly (Garcez and Lamb, 2023; Gurrapu et al., 2023). As summarized in Table 1, approaches with stronger symbolic integration generally provide improved interpretability and reasoning transparency, though often with trade-offs in scalability and engineering complexity. Compared to post-hoc explanations, these mechanisms potentially improve alignment between model behavior and generated explanations.

Nevertheless, faithfulness remains difficult to guarantee even in NeSy systems. Symbolic pipelines often depend on imperfect upstream components such as discourse parsers, argument mining systems, or knowledge extraction modules, which may introduce structural ambiguity. Moreover, persuasive communication remains inherently context-dependent and socially grounded, making rhetorical reasoning difficult to formalize completely. Developing explanations that are simultaneously faithful, interpretable, robust, and socially reliable therefore remains an important open challenge.

6 Evaluation

6.1 Standard Metrics and Their Limitations

Evaluation of propaganda and persuasion technique detection largely follows SemEval benchmark protocols, with micro-averaged F1 over persuasion techniques re-

maining the dominant metric, particularly for *Technique Classification (TC)*. While useful for standardized comparison, strong class imbalance can bias performance toward frequent techniques (e.g., *Loaded Language*) while masking weaknesses on rarer categories such as *Repetition*. Aggregate metrics further provide limited insight into reasoning quality, robustness, or interpretability. Fine-grained persuasion analysis also remains substantially more challenging than binary propaganda detection (Da San Martino et al., 2020), motivating approaches that incorporate structured reasoning and symbolic guidance.

6.2 Robustness Evaluation

Robustness is a critical but relatively underexplored evaluation dimension. Neural propaganda classifiers are often sensitive to paraphrasing, distribution shifts, and domain transfer (Le et al., 2020). Persuasion techniques may remain rhetorically unchanged despite substantial lexical variation, making robustness particularly important. Cross-domain transfer presents an additional challenge. Models trained on news propaganda frequently generalize poorly to social media, political speech, or health misinformation settings (Alam et al., 2022). NeSy methods potentially improve robustness by relying more heavily on argumentative structure, discourse organization, or symbolic representations rather than surface lexical patterns alone.

6.3 Explanation Evaluation Metrics

As persuasion analysis increasingly emphasizes interpretability, explanation quality represents an important complementary evaluation dimension. Prior work evaluates explanations along dimensions such as **faithfulness** (whether explanations reflect model reasoning), **plausibility** (whether explanations appear convincing to humans), **completeness** (whether relevant evidence is sufficiently covered), and **conciseness** (whether explanations remain appropriately minimal) (Atanasova et al., 2020; Xu et al., 2025a). Despite growing interest in explainable persuasion analysis, evaluation of explanation quality remains comparatively underdeveloped. More systematic evaluation protocols for reasoning transparency, explanation faithfulness, and human interpretability remain important directions for future work.

7 Open Challenges and Future Directions

Subtle and Implicit Persuasion. Widely studied techniques such as *Loaded Language*, *Repetition*, and *Name Calling* often exhibit stronger lexical signals than subtler forms of persuasion. Techniques such as *Doubt*, *Minimization*, and *Whataboutism* frequently depend on context, intent, pragmatic inference, and speaker modeling that current systems handle poorly (Da San Martino et al., 2019). Future NeSy approaches may benefit from richer representations of speaker goals, common ground, and context-dependent interpretation.

LLM-Generated Propaganda. LLMs introduce a new challenge by enabling fluent, stylistically diverse propaganda generation (Zellers et al., 2019; Barfar and Sommerfeldt, 2026). Unlike human-authored propaganda, synthetic persuasion can be optimized to evade existing detectors while preserving rhetorical effectiveness. Reasoning-centric detection approaches that model argumentative and rhetorical structure may improve resilience, but evaluation on LLM-generated propaganda remains limited. Developing benchmarks and adversarially robust NeSy methods is therefore an important direction.

Multilingual and Cross-Cultural Propaganda. Persuasion norms vary substantially across languages and cultures. A rhetorical strategy perceived as manipulative in one context may be interpreted as acceptable communication in another. Extending symbolic taxonomies across cultural settings therefore requires grounding in sociolinguistic and rhetorical theory. Although multilingual benchmarks such as SemEval-2023 (Piskorski et al., 2023) and SlavicNLP 2025 (Piskorski et al., 2025) provide initial infrastructure across languages, genres, and domains, culturally adaptive NeSy reasoning remains largely unexplored.

Integration with Fact Verification Pipelines. Propaganda detection and fact verification are complementary rather than competing tasks. Manipulative communication often relies on selective truth-telling, framing, or emphasis rather than factual falsehood alone. Pipeline architectures that combine persuasion analysis with claim verification represent a promising direction, where symbolic representations of rhetorical techniques can guide evidence retrieval and downstream reasoning.

Annotation Standards and Subjectivity. Low inter-annotator agreement in persuasion analysis reflects genuine ambiguity in rhetorical interpretation (Da San Martino et al., 2019). Recent analyses of persuasion-technique annotation campaigns further show that disagreement is not merely noise, but can reflect uncertainty about technique boundaries, annotator expertise, and the correctness of final labels (Bassi et al., 2025). Future work may benefit from annotation protocols informed by rhetoric and argumentation theory, expert annotator involvement, and benchmarks that explicitly model disagreement rather than collapsing to single labels. Probabilistic NeSy approaches that represent uncertainty over persuasion techniques offer one promising direction.

Evaluation of Explanation Quality. Despite growing interest in explainable persuasion analysis, standardized evaluation protocols remain limited. Future work should establish benchmarks targeting explanation faithfulness, interpretability, and downstream utility, while incorporating domain experts such as rhetoricians, journalists, and fact-checkers into evaluation pipelines.

8 Conclusion

We presented the first comprehensive survey of NeSy approaches for propaganda and persuasion technique detection. We argued that persuasion analysis occupies a distinctive position within computational misinformation research: persuasion techniques are rhetorically grounded, often symbolically defined, and require discourse-, argument-, and knowledge-level reasoning beyond surface semantic classification. These properties make propaganda and persuasion analysis particularly well aligned with NeSy approaches that integrate neural learning with structured reasoning. We organized existing work through a unified taxonomy spanning rule-guided learning, knowledge-augmented systems, graph-based reasoning, hybrid NeSy frameworks, and reasoning-oriented LLM approaches. Across these paradigms, NeSy methods offer promising mechanisms for improving interpretability, robustness, and reasoning transparency, while challenges remain in subtle persuasion, cross-cultural generalization, explanation evaluation, and resilience to increasingly sophisticated LLM-generated propaganda. As persuasive communication evolves, NeSy provides a promising foundation for more interpretable, robust, and reasoning-aware persuasion analysis systems.

Acknowledgment

This work was conducted as part of the IMDA SG Digital Leadership Accelerator Mentorship Program for Young Talent. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the organization.

Limitations

This work has several limitations arising from the evolving and interdisciplinary nature of the field. Propaganda and persuasion taxonomies remain socially situated constructs, with definitions of persuasion techniques, annotation protocols, and rhetorical interpretations varying across datasets, languages, and cultural contexts. These differences influence how methods are developed and evaluated. Also, as our survey adopts a functional definition of neurosymbolic AI that includes reasoning-oriented large language model (LLM) frameworks with explicit intermediate reasoning structures, alternative definitions may draw different boundaries around neurosymbolic reasoning. We view these limitations as characteristic of an emerging field whose conceptual foundations are still evolving.

References

Firoj Alam, Hamdy Mubarak, Wajdi Zaghouani, Giovanni Da San Martino, and Preslav Nakov. 2022. Overview of the wanlp 2022 shared task on propaganda detection in arabic. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 108–118.

Ryuta Arisaka, Ryoma Nakai, Yusuke Kawamoto, and Takayuki Ito. 2022. Theme aspect argumentation model for handling fallacies. *arXiv preprint arXiv:2205.15141*.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7352–7364.

Arash Barfar and Lee Sommerfeldt. 2026. Propaganda by prompt: Tracing hidden linguistic strategies in large language models. *Information Processing & Management*, 63(2):104403.

Alberto Barrón-Cedeno, Israa Jaradat, Giovanni Da San Martino, and Preslav Nakov. 2019. Proppy: Organizing the news based on their propagandistic content. *Information Processing & Management*, 56(5):1849–1864.

Davide Bassi, Dimitar Iliyanov Dimitrov, Bernardo D'Auria, Firoj Alam, Maram Hasanain, Christian Moro, Luisa Orrù, Gian Piero Turchi, Preslav Nakov, and Giovanni Da San Martino. 2025. Annotating the annotators: analysis, insights and modelling from an annotation campaign on persuasion techniques detection. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17918–17929.

Davide Bassi, Søren Fomsgaard, and Martín Pereira-Fariña. 2024. Decoding persuasion: a survey on ml and nlp methods for the study of online persuasion. *Frontiers in Communication*, 9:1457433.

Bikram Pratim Bhuyan, Amar Ramdane-Cherif, Ravi Tomar, and TP Singh. 2024. Neuro-symbolic artificial intelligence: a survey. *Neural Computing and Applications*, 36(21):12809–12844.

Alessandro Bondielli and Francesco Marcelloni. 2019. A survey on fake news and rumour detection techniques. *Information sciences*, 497:38–55.

Amber E Boydstun. 2013. *Making the news: Politics, the media & agenda setting*. University of Chicago Press.

Nimet Beyza Bozdag, Shuhaib Mehri, Xiaocheng Yang, Hyeonjeong Ha, Zirui Cheng, Esin Durmus, Jiaxuan You, Heng Ji, Gokhan Tur, and Dilek Hakkani-Tür. 2026. Must read: A comprehensive survey of computational persuasion. *ACM Computing Surveys*, 58(12):1–39.

Zhixi Cai, Fucai Ke, Simindokht Jahangard, Maria Garcia de la Banda, Reza Haffari, Peter J Stuckey, and Hamid Rezatofighi. 2025. Naver: A neuro-symbolic compositional automaton for visual grounding with

explicit logic reasoning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 24078–24089. IEEE.

Xianshuai Cao, Yuliang Shi, Han Yu, Jihu Wang, Xinjun Wang, Zhongmin Yan, and Zhiyong Chen. 2021. Dekr: description enhanced knowledge graph for machine learning method recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 203–212.

Robert B Cialdini. 2018. *Influence: The Psychology of Persuasion, Revised Edition*. Skillsoft.

Michele Luca Contalbo, Francesco Guerra, and Matteo Paganelli. 2024. Argument relation classification through discourse markers and adversarial training. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 18949–18954.

Edward Meredith Cope, John Edwin Sandys, and 1 others. 1877. *The rhetoric of Aristotle*, volume 2. WC Brown.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. Semeval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the fourteenth workshop on semantic evaluation*, pages 1377–1414.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeno, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 5636–5646.

Lauren Nicole DeLong, Ramon Fernández Mir, and Jacques D Fleuriot. 2024. Neurosymbolic ai for reasoning over knowledge graphs: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 36(5):7822–7842.

Dimitar Dimitrov, Firoj Alam, Maram Hasanain, Abul Hasnat, Fabrizio Silvestri, Preslav Nakov, and Giovanni Da San Martino. 2024. Semeval-2024 task 4: Multilingual detection of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2009–2026.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Semeval-2021 task 6: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021)*, pages 70–98.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, and 1 others. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9367–9385.

Géraud Faye, Benjamin Icard, Morgane Casanova, Guillaume Gadek, Guillaume Gravier, Wassila Ouerdane, Céline Hudelot, Sylvain Gatepaille, and Paul Égré. 2026. Reliable news or propagandist news? a neurosymbolic model using genre, topic, and persuasion techniques to improve robustness in classification. In *Proceedings of the 1st Workshop on Information Disorder (InDor)@ LREC*, volume 2026, pages 55–65.

Artur d'Avila Garcez and Luis C Lamb. 2023. Neurosymbolic ai: The 3 rd wave. *Artificial Intelligence Review*, 56(11):12387–12406.

Artur S d'Avila Garcez, Tarek R Besold, Luc De Raedt, Peter Földiak, Pascal Hitzler, Thomas Icard, Kai-Uwe Kühnberger, Luis C Lamb, Risto Miikkulainen, and Daniel L Silver. 2015. Neural-symbolic learning and reasoning: Contributions and challenges. In *AAAI Spring Symposia*, pages 18–21.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the association for computational linguistics*, 10:178–206.

Sai Gurrapu, Ajay Kulkarni, Lifu Huang, Ismini Lourentzou, and Feras A Batarseh. 2023. Rationalization for explainable nlp: a survey. *Frontiers in artificial intelligence*, 6:1225093.

Charles Leonard Hamblin. 2022. Fallacies. Advanced Reasoning Forum.

Maram Hasanain, Md Arid Hasan, Mohamed Bayan Kmainasi, Elisa Sartori, Ali Ezzat Shahroor, Giovanni Da San Martino, and Firoj Alam. 2025. Propxplain: Can llms enable explainable propaganda detection? *arXiv preprint arXiv:2502.16550*.

Pascal Hitzler and Md Kamruzzaman Sarker. 2022. Neuro-symbolic artificial intelligence: The state of the art.

Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard Hovy, and Eric Xing. 2016. Harnessing deep neural networks with logic rules. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2410–2420.

Nourhan Ibrahim, Samar Aboulela, Ahmed Ibrahim, and Rasha Kashef. 2024. A survey on augmenting knowledge graphs (kgs) with large language models (llms): models, evaluation metrics, benchmarks, and challenges. *Discover Artificial Intelligence*, 4(1):76.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4198–4205.

Jonathan Kim, Anna Podlasek, Kie Shidara, Feng Liu, Ahmed Alaa, and Danilo Bernardo. 2025. Limitations of large language models in clinical problem-solving arising from inflexible reasoning. *Scientific reports*, 15(1):39426.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational linguistics*, 45(4):765–818.

Thai Le, Suhang Wang, and Dongwon Lee. 2020. Malcom: Generating malicious comments to attack neural fake news detection models. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 282–291. IEEE.

Alfred Lee and Elizabeth Briant Lee. 1939. The fine art of propaganda.

Bill Yuchen Lin, Ronan Le Bras, Kyle Richardson, Ashish Sabharwal, Radha Poovendran, Peter Clark, and Yejin Choi. 2025. Zebralogic: On the scaling limits of llms for logical reasoning. *arXiv preprint arXiv:2502.01100*.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):1–25.

Fabrizio Macagno. 2025. Argumentation schemes. *The Routledge Handbook of Argumentation Theory*, pages 169–182.

Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeño, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Watheq Mansour, and 1 others. 2021. Overview of the clef–2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 264–291. Springer.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. 2023. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3806–3824.

Matt Pauk and Maria Leonor Pacheco. 2024. Pauk at semeval-2024 task 4: A neuro-symbolic method for consistent classification of propaganda techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1424–1434.

Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand. 2021. Shifting attention to accuracy can reduce misinformation online. *Nature*, 592(7855):590–595.

Andrew Piper, Richard Jean So, and David Bamman. 2021. Narrative theory for computational narrative understanding. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 298–311.

Jakub Piskorski, Dimitar Iliyanov Dimitrov, Filip Dobranić, Marina Ernst, Jacek Haneczok, Ivan Koychev, Nikola Ljubešić, Michal Marcinczuk, Arkadiusz Modzelewski, Ivo Moravski, and 1 others. 2025. Slavicnlp 2025 shared task: detection and classification of persuasion techniques in parliamentary debates and social media. In *Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025)*, pages 254–275.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the 17th international workshop on semantic evaluation (SemEval-2023)*, pages 2343–2361.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937.

Eva Sánchez Salido, Julio Gonzalo, and Guillermo Marco. 2026. On the limits of llm reasoning: Evidence from contamination, translation, and answer modification in multiple-choice benchmarks. *IEEE Access*, 14:9384–9393.

Xu Shen, Song Wang, Zhen Tan, Laura Yao, Xinyu Zhao, Kaidi Xu, Xin Wang, and Tianlong Chen. 2026. Faithcot-bench: Benchmarking instance-level faithfulness of chain-of-thought reasoning. In *International Conference on Learning Representations*, volume 2026, pages 65918–65948.

Parshin Shojaee, Iman Mirzadeh, Maxwell Horton, Samy Bengio, Mehrdad Farajtabar, and 1 others. 2026. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *Advances in Neural Information Processing Systems*, 38:108018–108059.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, and 1 others. 2025. A survey of reasoning with foundation models: Concepts, methodologies, and outlook. *ACM Computing Surveys*, 57(11):1–43.

Zachary Susskind, Bryce Arden, Lizy K John, Patrick Stockton, and Eugene B John. 2021. Neuro-symbolic ai: An emerging class of ai workloads and their characterization. *arXiv preprint arXiv:2109.06133*.

Maite Taboada and William C Mann. 2006. Rhetorical structure theory: Looking back and moving ahead. *Discourse studies*, 8(3):423–459.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.

Nawar Turk, Sevag Kaspar, and Leila Kosseim. 2025. On the influence of discourse relations in persuasive texts. *arXiv preprint arXiv:2510.26124*.

Jan-Willem Van Prooijen. 2018. *The psychology of conspiracy theories*. Routledge.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Mann William and Sandra Thompson. 1988. Rhetorical structure theory: Towards a functional theory of text organization. *Text*, 8(3):243–281.

Ben Wooding, Kiersten Brennan, Anne M Tumlin, Hongchao Zhang, and Taylor T Johnson. 2026. A self-correcting neuro-symbolic ai reasoning framework. In *International Symposium on AI Verification*, pages 191–211. Springer.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24.

Austin Xu, Srijan Bansal, Yifei Ming, Semih Yavuz, and Shafiq Joty. 2025a. Does context matter? contextualjudgebench for evaluating llm-based judges in contextual settings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9541–9564.

Yige Xu, Xu Guo, Zhiwei Zeng, and Chunyan Miao. 2025b. Softcot: Soft chain-of-thought for efficient reasoning with llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23336–23351.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Dongran Yu, Bo Yang, Dayou Liu, Hui Wang, and Shirui Pan. 2023. A survey on neural-symbolic learning systems. *Neural Networks*, 166:105–126.

Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2022. A survey of knowledge-enhanced text generation. *ACM Computing Surveys*, 54(11s):1–38.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *Advances in neural information processing systems*, 32.

Xia Zeng, Amani S Abumansour, and Arkaitz Zubiaga. 2021. Automated fact-checking: A survey. *Language and Linguistics Compass*, 15(10):e12438.

Xuan Zhang, Chao Du, Tianyu Pang, Qian Liu, Wei Gao, and Min Lin. 2024. Chain of preference optimization: Improving chain-of-thought reasoning in llms. *Advances in Neural Information Processing Systems*, 37:333–356.

Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2020. Graph neural networks: A review of methods and applications. *AI open*, 1:57–81.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.