Information Processing and Management 62 (2025) 104127

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Neurosymbolic graph enrichment for Grounded World Models

Stefano De Giorgis [a,*], Aldo Gangemi [a,b], Alessandro Russo [a]

[a] Institute for Cognitive Sciences and Technologies - National Research Council (CNR-ISTC), Italy
[b] Department of Philosophy - University of Bologna, Italy

ARTICLE INFO

Keywords:
Neurosymbolic AI
Knowledge representation
Knowledge extraction
Large language models
Graph KAG
Hybrid reasoning

ABSTRACT

The development of artificial intelligence systems capable of understanding and reasoning about complex real-world scenarios is a significant challenge. In this work we present a novel approach to enhance and exploit LLM reactive capability to address complex problems and interpret deeply contextual real-world meaning. We introduce a method and a tool for creating a multimodal, knowledge-augmented formal representation of meaning that combines the strengths of large language models with structured semantic representations. Our method begins with an image input, utilizing state-of-the-art large language models to generate a natural language description. This description is then transformed into an Meaning Representation (AMR) graph, which is formalized and enriched with logical design patterns, and layered semantics derived from linguistic and factual knowledge bases. The resulting graph is then fed back into the LLM to be extended with implicit knowledge activated by complex heuristic learning, including semantic implicatures, moral values, embodied cognition, and metaphorical representations. By bridging the gap between unstructured language models and formal semantic structures, our method opens new avenues for tackling intricate problems in natural language understanding and reasoning.

1. Introduction

The development of artificial intelligence systems capable of understanding and reasoning about complex real-world scenarios remains a significant challenge in computer science. This challenge is particularly evident when considering the distinct paradigms of generative AI and knowledge-based AI. Generative AIs, including Large Language Models (LLMs), function as signal processing machines: they learn activation patterns from input of any modality and provide symbolic output at inference time. In contrast, knowledge-based AIs operate as logical machines, extracting and designing symbolic representation patterns of the world with model-theoretical interpretations to ensure correct inference.

While both approaches have demonstrated remarkable capabilities, they often lack the comprehensive understanding necessary to build Grounded World Models (GWM) (Forrester, 1971; Ha & Schmidhuber, 2018), i.e., models of the world as experienced or constructed by humans (and other organisms). GWMs are multivaried, encompassing physical, neurocognitive, social, and cultural dimensions. This multidimensional approach is intended to better mirror the complexity of human understanding.

This paper presents a novel neurosymbolic approach that aims to bridge this gap by leveraging LLMs' power of automated generation from latent knowledge, and the heuristic power of ontology-based knowledge graphs. We called our system "Polanyi" from the mathematician Michael Polanyi, and its work on tacit and implicit knowledge.

At the core of our approach is the assumption that human-like understanding requires the integration of multiple layers of knowledge, ranging from sensorimotor experiences to abstract conceptual structures. As argued by Lakoff and Johnson (Lakoff,

* Corresponding author.
E-mail addresses: stefano.degiorgis@cnr.it, stefano.de.giorgis@gmail.com (S. De Giorgis).

https://doi.org/10.1016/j.ipm.2025.104127
Received 19 September 2024; Received in revised form 24 January 2025; Accepted 3 March 2025
Available online 22 March 2025

Johnson, et al., 1999), human cognition is grounded in embodied experiences, which give rise e.g. to image schemas and conceptual metaphors that shape our understanding of more abstract domains. An interesting consequence, which was not noticed until the appearance of LLMs (Nolfi, 2023), is that abstract knowledge (as represented in natural language, logical and statistical models, knowledge bases, sensor data) can work as a supramodal system that bears correspondences to the physical, social, cognitive worlds.

Consequentially, we propose Polanyi, a framework for developing GWMs trained with, and capable of generating, knowledge graphs. Training includes either in-context learning or fine-tuning of pre-trained (multimodal) LLMs reinforced to generate multilayered knowledge, including highly contextual implicit knowledge. Implicit knowledge includes e.g. presuppositions, conversational implicatures, factual impacts, image schemas, metonymic and symbolic coercions, event sequences, causal relations, emotion- and moral value-driven reasoning.

A key innovation in our approach is the use of LLMs not as expert systems but as reactive engines to extract implicit contextual knowledge. Besides using heuristic rules and curated knowledge bases, we leverage the reactive power embedded in LLMs to extract and formalize general knowledge across multiple semantic dimensions.

To operationalize this vision, we introduce a novel method to create knowledge-augmented formal representations of multimodal meaning. For example, starting from an image input, we reuse state-of-the-art LLMs to generate a natural language description. This description is then transformed into Abstract Meaning Representation (AMR) graphs (Bevilacqua & Navigli, 2020) which are then formalized as an ontology-based knowledge graph (Gangemi, Presutti, et al., 2017), and aligned to public knowledge bases, with the Text2AMR2FRED tool (Gangemi, Graciotti, Meloni, Nuzzolese, Presutti, Recupero, Russo, & Tripodi, 2023; Gangemi, et al., 2017; Meloni, Recupero, & Gangemi, 2017). The resulting graph is iteratively extended with implicit knowledge activated by heuristics for presuppositions, coercions, impact, etc.

Our approach implements a feedback loop that leverages multimodal, multilayered GWMs. This iterative process enables continuous refinement and expansion of the knowledge graph, enhancing the system's ability to adapt to new contexts and improve its understanding over time.

This neurosymbolic approach makes the process of knowledge base enrichment more agile and scalable compared to traditional methods. By combining the flexibility and generative power of LLMs with the structure and inferential capabilities of symbolic knowledge representations, we enable rapid expansion and refinement of knowledge graphs across diverse domains.

This hybrid architecture can be classified as a Type 2–3 neurosymbolic system in Kautz' taxonomy (Kautz, 2022).

The potential applications of this approach are far-reaching, spanning natural language understanding, visual reasoning, and complex problem-solving tasks. By providing a framework for grounding abstract language in embodied and commonsense knowledge, our method opens new avenues for developing AI systems capable of human-like reasoning and contextual understanding.

In the following sections, we detail the technical architecture of our implemented system and a short introduction to a public demo, and present experimental results demonstrating the efficacy of our approach across multiple knowledge domains.

The paper is organized as follows: Section 2 provides state-of-the-art positioning, Section 3 details the methodology, Section 4 describes the three-tier evaluation of the method, Section 5 illustrates ongoing and future works, and finally Section 6 concludes the paper.

2. Related work

Our approach is positioned in the quickly growing field of Large Language Models and Knowledge Graphs hybrid neurosymbolic methods. The synergy between neural and symbolic approaches has been extensively explored in multiple surveys (Belle, 2020; Besold et al., 2021; d'Avila Garcez & Lamb, 2023; Sarker, Zhou, Eberhart, & Hitzler, 2021; Yu, Yang, Liu, Wang, & Pan, 2023), some of them focusing explicitly on graph neural networks and symbolic components (Lamb, Garcez, Gori, Prates, Avelar, & Vardi, 2020), reasoning over graph structures (Zhang, Chen, Zhang, Ke, & Ding, 2021), dynamic knowledge graphs (Alam, Gesese, & Paris, 2024), and natural language inference (Yao, Xu, Shi, & Xu, 2018), reflecting the growing interest in hybrid methodologies.

While an exhaustive survey lies beyond the scope of this paper, we can position our method within the broader landscape of recent advancements: (i) within the landscape of interactions between LLMs and KGs in joint methods , and (ii) in relation to popular techniques such as Graph RAG (Retrieval-Augmented Generation) that have recently garnered significant attention.

Regarding joint LLMs and KGs methods, according to Pan et al. (2024), as shown in Fig. 1, our approach would be both "LLM augmented KG" and "Synergized LLMs + KGs", and in particular "LLMs - augmented KGs construction" focusing on "End-to-end construction" and "Distilling KGs from LLMs", as well as "LLMs-augmented KGs to text generation" focusing on "Leveraging knowledge from LLMs", and partially "LLM-augmented KG for question answering" adopting "LLMs as entity/relations extractors".

Considering Kautz' taxonomy of neurosymbolic systems (Kautz, 2022), as well as Hamilton's contextualization in the domain of natural language understanding (Hamilton, Nayak, Božić, & Longo, 2022), our pipeline qualifies as Type 2 ("Symbolic[Neuro] (Nested)") - Type 3 ("Neuro; Symbolic (Cooperative)") architecture. Type 2 is defined as nested architecture, where a symbolic reasoning system is the primary system with neural components driving certain internal decisions, while Type 3 covers cases in which a neural network focuses on one task, e.g. entity linking, word sense disambiguation, etc., and interacts via input/output with a symbolic reasoner specializing in a complementary task, in our case, building semantic tree dependencies in RDF format while aligning entities to Framester (Gangemi, Alam, Asprino, Presutti, & Recupero, 2016) hub of ontologies, maintaining correct OWL2 syntax.

Considering Graph RAG approaches, in recent years the integration of LLMs and knowledge graphs has gained significant traction (Zhu et al., 2024), particularly using models for knowledge graph completion (Yao, Mao, & Luo, 2019), and through the development of Retrieval-Augmented Generation (RAG) and Graph RAG techniques. Different models (GPT, Llama, Claude,

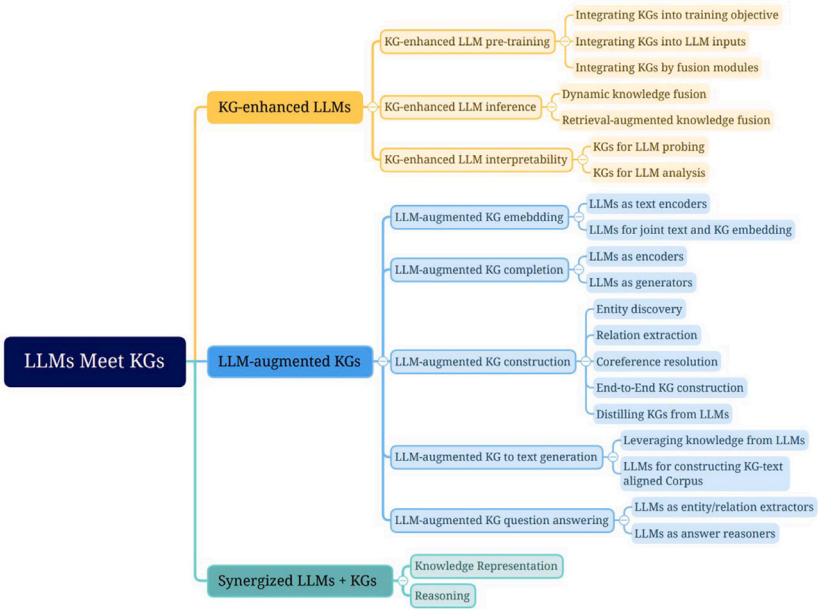

Fig. 1. Roadmap for LLMs and KGs interactions.
Source: Taken from Pan, Luo, Wang, Chen, Wang, and Wu (2024).

Gemini, etc.) have shown different specific capabilities but similar structural limitations (Meyer et al., 2023). While traditional RAG focuses on retrieving precise information from unstructured or vector-compressed resources, Graph RAG has demonstrated remarkable efficacy and compliance with graph-structured data (Edge et al., 2024). The Graph RAG approach extends beyond mere text vectorization, encompassing a two-fold process: entity recognition and relation extraction, culminating in the generation of semantic triples.

Classic applications of Graph RAG have been in question-answering (Sen, Mavadia, & Saffari, 2023), starting from textual information transposed to knowledge graph embeddings (Lu, Cong, & Huang, 2020), or summarization systems designed to retrieve specific information from established knowledge bases (Edge et al., 2024). However, certain layers of meaning remain challenging to capture and formalize, including moral reasoning, pragmatic implicatures, and tacit knowledge (Polanyi, 2009) derived from real-world experiences.

The topic of knowledge graph generation from text is of some relevance in several overlapping communities, such as the Semantic Web one (Tiwari et al., 2022), as well as the broader knowledge-graph-oriented one[1] (Edge et al., 2024).

For this reason, our research expands beyond Graph RAG. We leverage LLMs as latent reactors that can provide approximate, implicit, commonsense knowledge when activated with appropriate heuristics, rather than as tools for compressed information retrieval. Our approach, partially positioned as Knowledge Augmented Generation (KAG),[2] enables us to extract and formalize implicit information that humans intuitively infer, especially in visual comprehension tasks. The KAG method described in this paper assumes logically sound graphs, and falls under the Logic Augmented Generation paradigm (Gangemi & Nuzzolese, 2025).

By doing so, our work addresses a longstanding challenge in artificial intelligence – the Frame problem[3] – which concerns the ability of AI systems to contextualize information and make human-like inferences, massively reducing the combinatorial space of choices. Several past approaches, such as the work of Brachman (Brachman, 1977), Minsky (Minsky et al., 1974), and Fillmore (Fillmore et al., 2006) attempted to tackle this issue by using structured representation. Being able to faithfully capture the nature of a domain without enumerating all the implicit knowledge that it assumes when discussing or performing practices is indeed a cognitive and computationally hard task.

Concerning technical components, we include here some knowledge extraction tools/methods that have paved the way to what we present in this article.

FRED[4] (Gangemi, Presutti, et al., 2017) is a comprehensive tool for formal knowledge graph generation from natural language (using the OWL2 description logic Grau, Horrocks, Motik, Parsia, Patel-Schneider, & Sattler, 2008), including additional features

[1] https://www.linkedin.com/posts/tonyseale_the-microsoft-graphrag-library-has-recently-activity-7230121112158830593-m8At?utm_source=share&utm_medium=member_desktop

[2] For terminological reference, cf.: https://aidanhogan.com/talks/2024-09-04-wuwien-invited-talk.pdf.

[3] https://plato.stanford.edu/entries/frame-problem/

[4] Available at http://wit.istc.cnr.it/stlab-tools/fred/demo/.

such as entity recognition and entity linking, frame extraction, semantic role labeling, and word sense disambiguation. FRED uses Categorical Grammar (Lambek, 1988) and Boxer (Bos, 2008) to create a formal graph, which is then aligned to WordNet, FrameNet, DBpedia, and other public resources.

Abstract Meaning Representation (AMR) has been proposed long ago in computational linguistics as a solution for natural language representation that can be pragmatically extracted from text. Recently, an efficient parser (Bevilacqua & Navigli, 2020) based on transformer's end-to-end technology has opened new ways to make it scalable. AMR has a graph-based structure that informally encodes complex relationships and dependencies between concepts, facilitating a more nuanced representation of linguistic meaning. Compared to predecessors like Categorical Grammars, AMR abstracts away from surface-level syntactic variations, simplifying the application of logical and ontological patterns when creating knowledge graphs from text.

AMR2FRED (Gangemi, et al., 2017), a deterministic, rule-based system, which applies the formal design patterns of FRED's to AMR, and provides alignments reusing Framester (Gangemi, Alam, et al., 2016) as main hub of public resources. The Framester ontology hub (Gangemi, 2020; Gangemi, Alam, et al., 2016) provides a formal semantics to semantic frames (Fillmore et al., 2006) in a curated linked data version of multiple linguistic resources (e.g. FrameNet Baker, Fillmore, & Lowe, 1998, WordNet Miller, 1998, VerbNet Schuler, 2005, PropBank Kingsbury & Palmer, 2002), a cognitive layer including MetaNet (Gangemi, Presutti, & Alam, 2018) and ImageSchemaNet (De Giorgis, Gangemi, & Gromann, 2024), BabelNet (Navigli & Ponzetto, 2010), factual knowledge bases (e.g. DBpedia Auer, Bizer, Kobilarov, Lehmann, Cyganiak, & Ives, 2007, YAGO Suchanek, Kasneci, & Weikum, 2007, etc.), and ontology schemas (e.g. DOLCE-Zero Gangemi, Guarino, Masolo, & Oltramari, 2003), with formal links between them, resulting in a strongly connected RDF/OWL knowledge graph.

Considering LLMs, the modular method we present here is potentially agnostic to the LLMs used. Nevertheless, the current version of the tool implementing the method, available for testing online,[5] uses GPT-4o and Claude Sonnet 3.5 APIs. The results, shown in Section 4, are therefore based on these two top-scoring models.

3. Methodology

In this section we describe Polanyi and the developed approach, detailing (i) the technical structure, and (ii) the knowledge layers we have chosen to include in our heuristics. It is important to note that the selection of the heuristics is not exhaustive and can be incrementally expanded to incorporate additional knowledge areas as needed, as envisioned in Section 5. Additionally, we discuss the advantages and disadvantages of the implemented pipeline.

3.1. XKG generation pipeline

Our modular pipeline combines natural language processing, knowledge representation, and large language models (LLMs) to extract and enrich semantic information from textual or visual inputs.

The whole process is shown in Fig. 2, and begins on the top left, with either user-provided text or an LLM-generated description of an input image, using a carefully crafted prompt. Considering the case in which we provide a picture as starting input, following Fig. 2, this original content is passed to a Multimodal LLM (MLLM), prompted to provide a description of the picture in natural language. In our current implementation, after several tests, we retrieved that GPT-4o is the best model at providing a textual description when given as input an image. Therefore, starting from an image, we rely on GPT-4o to get as output a textual description. In Section 4 we provide an example of input and output; furthermore, all the prompts are provided as additional material on the GitHub repository.[6]

This textual representation is passed to the Text2AMR2FRED (TAF) tool (Gangemi et al., 2023), which transforms the input text into an Abstract Meaning Representation (AMR) graph using the SPRING parser (Bevilacqua, Blloshmi, & Navigli, 2021). As part of this step, entity linking is performed from the nodes of the AMR graph to Wikipedia entries using BLINK (Wu, Petroni, Josifoski, Riedel, & Zettlemoyer, 2019). Due to the opaque nature of the Text-to-AMR conversion, which does not explicitly map text segments to graph elements, entity linking is performed heuristically: BLINK identifies Wikipedia entity mentions in the text, and if AMR nodes are labeled with matching text segments, they are linked to the corresponding Wikidata entities.

The AMR graph is subsequently converted to an OWL-RDF knowledge graph relying on AMR2FRED (Gangemi, et al., 2017; Meloni et al., 2017) using FRED heuristics (Gangemi, Presutti, et al., 2017) and motifs (Gangemi, Recupero, Mongiovì, Nuzzolese, & Presutti, 2016). The enrichment of this RDF graph through word sense disambiguation (WSD) is done using the eWiSeR tool (Bevilacqua & Navigli, 2020). Similar to the entity linking process, WSD is applied to the original text, linking text segments to WordNet synsets (sets of contextual synonyms). In essence, when RDF graph entities have names (roughly corresponding to the final part of their URIs) that match disambiguated text segments, an `owl:equivalentClass` triple linking the entity to the WordNet synset is added to the graph. Leveraging Framester (Gangemi, Alam, et al., 2016), we augment the graph with additional `rdfs:subClassOf` triples for the disambiguated entities, connecting them to WordNet supersenses and DOLCE types, derived from querying Framester with the identified WordNet synsets. This RDF graph, the result of the text-to-AMR-to-FRED (TAF) pipeline, serves as our "Base Graph", shown in Fig. 2 as "Knowledge graph" in the middle of the pipeline.

From here, we re-inject this Base Graph to LLM for knowledge enrichment. For this step, after experimenting with a gold standard of formal knowledge graphs extracted from text (Gangemi, Recupero, et al., 2016), we found that Claude 3.5 Sonnet is the best

[5] https://arco.istc.cnr.it/itaf/

[6] https://github.com/StenDoipanni/XKG/tree/main/prompt

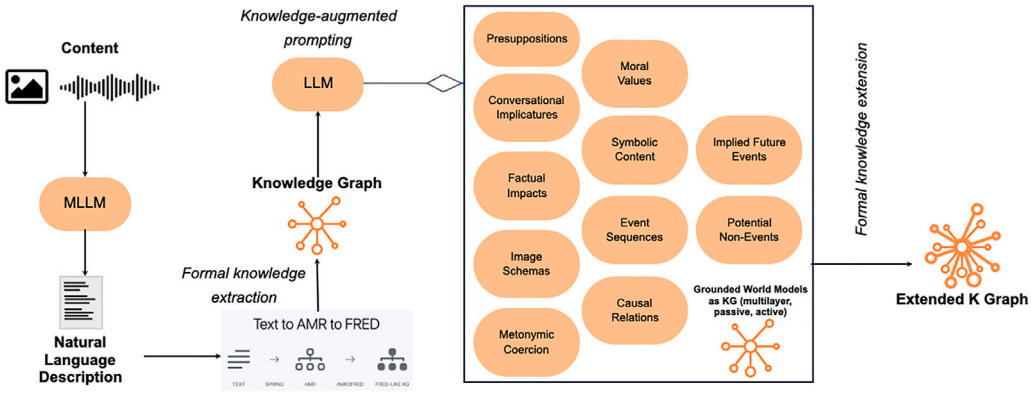

Fig. 2. Polanyi's hybrid knowledge enrichment pipeline.

model for the purpose. The "Knowledge Augmented prompting" step of the pipeline calls Claude 3.5 Sonnet API, and takes as input a specific base prompt, including the Base Graph expressed in Turtle syntax. This base prompt is then refined in a specific prompt for each heuristic, shown as salmon-colored boxes on the right of Fig. 2.

For each predefined enrichment heuristic, we prompt Claude to generate additional triples, expanding the implicit knowledge captured in the graph, anchoring the newly introduced triples to nodes of the Base Graph. Each heuristic produces a separate graph that combines the Base Graph with the heuristic's specific additional triples. Finally, we construct a comprehensive graph that merges the base graph with all the additional triples generated across all heuristics, resulting in a richer, multi-faceted semantic representation of the initial image or text input.

3.2. Knowledge heuristics

We describe here the 11 heuristics currently available in the implemented tool. They are chosen among the types of Polanyi's tacit knowledge (Polanyi, 2009) that have emerged as very hard for natural language understanding and automated knowledge extraction: pragmatics, common sense, event prediction and causality, shared assumptions and biases, sensorimotor competences, etc. The list is not proposed as a complete or near-complete typology of tacit knowledge types, but touches an important subset that is intensely studied in many disciplines. It should be considered as a bootstrap used to experiment with, and evaluate, automated tacit knowledge extraction, which constitutes an essential part of daily human understanding and sense-making activity in building and using Grounded World Models. The list includes: *Presuppositions*, based on previous background knowledge; *Conversational Implicatures* (Grice, 1975), which often contributes in making sense of incomplete information in linguistic exchanges; *Factual Impact*, which grounds linguistic entities to factual knowledge; *Image Schemas*, basic building blocks of cognition which grounds our way of conceiving the world in our sensori-motor bodily perception (grounding e.g. cognitive metaphors and several other entities); *Metonymic Coercions*, which allows understand propositions whose truth value would be zero, but differ from metaphorical speech grounding the relation between entities on the parthood-whole relation; *Moral Value Driven Coercion*, applied everyday in appraisal and moral evaluative processes, values nudge our daily behavior; *Symbolic Coercion*, in Peirce terminology (Peirce & Buchler, 1955), used to anchor meaning to various entities of the world; *Event Sequences*, determinant in our plan-making capability and ability to design plausible scenarios and outcomes; *Causal Relations*, establishing relations of cause–effect between processes and events, to avoid either having only (i) temporal sequences and (ii) statistical correlation; *Implied Future Events*, a specification of Event Sequences, for temporal projection in the future; and *Implied Non-Events*, an infinite set of events, but, referring to the Frame problem, focusing on those more closely related to a specific Event Sequence.

Presuppositions. Presuppositions[7] are implicit assumptions necessary for statements to be meaningful (Frege, 1892; Karttunen, 2016; Strawson, 1950). They play a crucial role in human cognition, enabling efficient communication through shared background knowledge. In natural language, presuppositions help infer unstated information and enhance contextual understanding. By formalizing and integrating these implicit meanings into knowledge structures, the approach captures nuanced understandings often taken for granted in human communication. For example, the statement "The athlete won the gold medal" presupposes that there was a competition, possibly an Olympic one, illustrating how presuppositions convey information beyond the explicit content of a sentence.

[7] https://plato.stanford.edu/entries/presupposition/

Conversational implicatures. Conversational implicatures[8] are implied meanings that arise from the context of a conversation rather than literal interpretation. Defined in Grice's pragmatics (Grice, 1975), they are essential to human communication, allowing speakers to convey more information than explicitly stated (Levinson, 2000). By formalizing and integrating these implicatures into knowledge structures, the approach captures nuanced, context-dependent interpretations that humans naturally derive from conversations. This enhances the system's ability to understand and reason about complex linguistic phenomena. For example, if someone asks "Is there a gas station nearby?" and receives the answer "There's one around the corner", the implicature here is that the gas station should be open and operational, even though this is not explicitly stated.

Factual impact. Factual impact refers to the physical, social, and cognitive consequences of events on participants, including expected emotions, sensations, and changes in mental states (Kahneman, 1991). This concept is crucial for understanding the full implications of human interactions and narratives. By formalizing these impacts, the approach creates a comprehensive representation of how events affect individuals on multiple levels. This enriched representation allows for inference and reasoning about e.g. the emotional and cognitive states of participants, adding human-like comprehension to the analysis of social interactions. For example, in the event of "winning a competition", the factual impact might include physical sensations (adrenaline rush), emotions (joy, pride), and cognitive changes (increased confidence, future goal-setting).

Image schemas. Image schemas are fundamental cognitive structures that help humans organize and interpret their experiences of the world (Johnson, 1987; Lakoff et al., 1999). These gestaltic schemas, such as container, path, balance, and force, underlie our perceptual understanding of sentences and situations (Besold, Hedblom, & Kutz, 2017). By integrating these schemas into knowledge representation, the approach captures a crucial aspect of human cognition and spatial reasoning (Hedblom, 2020). This allows for inference and reasoning about underlying spatial and conceptual structures that humans instinctively understand, also via cognitive metaphors (Lakoff & Johnson, 1980), enhancing the system's ability to represent complex narratives and interactions. For example, the container schema helps us understand phrases like "in the competition" or "out of the box", while the path schema underlies our comprehension of sentences describing movement or progress.

Metonymic coercion. Metonymic coercion is a linguistic phenomenon where a word's typical or literal sense is overridden by a specific, related sense within its extended semantics. This process is fundamental to human language comprehension, enabling more efficient and nuanced communication. By formalizing metonymic coercions in knowledge representation (Maudslay, Teufel, Bond, & Pustejovsky, 2024), the approach captures implicit semantic shifts that occur in everyday language. This allows the system to infer and reason about intended meanings behind metonymic expressions, bridging the gap between literal interpretations and context-dependent understandings. For example, in the sentence "The White House announced new policies", "White House" undergoes metonymic coercion to represent the U.S. government or administration, rather than the physical building.

Moral-value-driven coercion. Moral-value-driven coercion[9] is a linguistic phenomenon where the literal meaning of an action or statement is overridden by a morally or socially charged interpretation. This process reflects the underlying value systems (Graham et al., 2013; Schwartz, Melech, Lehmann, Burgess, Harris, & Owens, 2001) that guide human behavior and decision-making (Rozin, Lowery, Imada, & Haidt, 1999). By formalizing these coercions in knowledge representation, the approach captures implicit moral and social dynamics at play in human interactions. This allows the system to infer and reason about underlying ethical considerations and social norms that shape human behavior, adding depth to the understanding of complex narratives and interpersonal dynamics. For example, the statement "She always keeps her promises" might be coerced from a literal description of behavior to a moral judgment about the person's integrity and trustworthiness within a social context.

Symbolic coercions. Symbolic coercions involve the transformation of literal meanings into symbolic interpretations.[10] This process is fundamental to human cognition and communication (Peirce & Buchler, 1955), allowing for the anchoring of complex ideas to familiar concepts, objects, and imagery. By formalizing these coercions in knowledge representation, the approach captures implicit symbolic meanings that permeate human language and thought. This enables the system to recognize and reason about the underlying symbolic significance of actions and concepts, particularly in domains where abstract ideas are often expressed through concrete objective correlative (Eliot et al., 1920). For example, in sport events commentaries, the phrase "the cangaroos won the match" undergoes symbolic coercion from a literal description of marsupial mammals to representing Australia (including potential biases and stereotypes).

Event sequences. Event sequences (Motta, Daga, Gangemi, Gjelsvik, Osborne, & Salatino, 2024) capture the chronological relationship between events mentioned in a text, providing crucial information about the order of actions or occurrences. By formalizing these sequences, the approach creates a rich representation of implicit chronological information in narratives. This enables the system to infer and reason about the temporal flow of events, even when not explicitly stated. Such temporal reasoning is essential for understanding causality, narrative progression, and the logical flow of actions and reactions in complex scenarios. For example, in the sentence "After the flight, the athletes have two days before the competition", the system would recognize the sequence: $flight \rightarrow two\_days \rightarrow competition$, allowing for deeper comprehension of the narrative's timeline and potential causal relationships between events.

[8] https://plato.stanford.edu/entries/implicature/
[9] https://plato.stanford.edu/entries/value-theory/
[10] https://plato.stanford.edu/entries/peirce-semiotics/

Causal relations. Causal relations capture the cause-and-effect relationships between events or states mentioned in a text. By formalizing these relationships the approach creates a rich representation of implicit causal information. This enables the system to infer and reason about underlying causes and effects within complex scenarios, even when not explicitly stated. Understanding causal relations is essential for comprehending motivations, predicting outcomes, and constructing coherent mental models of situations. For example, in the sentence "The heavy rain forced the match to be postponed", the system would recognize the causal chain: $heavy\_rain \rightarrow match\_postponing$, allowing for deeper analysis of the event's progression and potential consequences.

Implied future events. Implied future events (Motta et al., 2024) involve inferring likely outcomes or consequences based on given information. This concept captures the human ability to anticipate future scenarios and make predictions based on current contexts and interactions. By formalizing these implied future events, we are able to reason about probable consequences of current actions and statements, even when not explicitly mentioned in the text. Such predictive reasoning is essential for understanding long-term significance of interactions. For example, in the sentence "The committee announced more strict regulations", the system might infer potential future events such as more severe checks, increase in bureaucratic practices, increase in controversy, etc.

Implied potential non-events. Implied potential non-events (Motta et al., 2024) are events that could have occurred but are prevented or made unlikely due to other circumstances or decisions mentioned in the text. This concept captures the understanding of alternative scenarios and the implications of characters' choices. Although this is an infinite set of entities, namely all the possible alternative scenarios that will never take place given a starting situation, we focus on those implicit alternatives that humans naturally consider when interpreting narratives. This enables the system to reason about not just what happens, but also what could have happened under different circumstances. Such counterfactual reasoning is essential for understanding motivations, the significance of decisions, and the broader implications of narrative events. For example, in the sentence "She decided not to compete", the system would recognize the implied potential non-event of participating to the competition, allowing for analysis of the decision's consequences and alternative outcomes.

The graphs produced from all these heuristics together form the set of Extended Knowledge Graphs (XKGs).

3.3. Prompting approach and techniques

Our approach and pipeline (Fig. 2) incorporate two key interaction points with LLMs: (i) extracting a textual description from an image, and (ii) enriching and extending the knowledge graph (KG) based on identified heuristics. Each of these interactions is driven by carefully crafted prompts tailored to each specific task, as described hereafter. The complete prompt list is available on GitHub.[11]

3.3.1. Prompt for generating image descriptions

The prompt for generating text from an image is designed and formulated to elicit concise, nuanced, and stylistically constrained outputs from the LLM. The prompt explicitly requires the LLM to describe the overt, implicit, and symbolic content of the image. The description must synthesize these layers of meaning into a brief, 100-word max text, emphasizing conciseness and integration of content. To better accommodate the capabilities of AMR parsers,[12] stylistic constraints are defined to ensure a continuous, flowing sentence structure and style. The model is prompted to avoid explicit textual references to the image itself, so as to focus on describing the content without directly acknowledging its medium (the image). The final description is expected to be a cohesive synthesis of explicit, implicit and symbolic elements, keeping the focus purely on content with a streamlined, connected narrative.

3.3.2. Prompting templates for heuristic-based KG extension

For each of the 11 heuristics, we have defined a dedicated prompting template to guide the interaction with the LLM. Although the content of each prompt is tailored to address the specifics of the corresponding heuristic, all prompts share a common structure, enabling consistency across the evaluation process. This shared structure allows for logical partitioning into distinct sections, each fulfilling a specific role in the prompting workflow and incorporating advanced prompting techniques. The sections of the prompts and related prompting techniques are as follows.

Role specification. The prompt starts by defining the role of the model explicitly (an "expert ontology engineer"). This sets the context and expertise required for the task. The *role assignment* technique (or *role-playing prompting technique*) establishes a persona for the model to emulate, focusing its behavior and expertise.

General instructions and guidance on input format. The second section provides detailed background information about the inputs (i.e., the text and the KG) and about the standards and technical framework used (e.g., OWL2, frame semantics, ontology design patterns, RDF in Turtle syntax, alignments to resources like PropBank, WordNet, etc.).

[11] https://github.com/StenDoipanni/XKG/tree/main/aft_rev/prompts

[12] AMR parses may be less precise when dealing with multi-sentence text and coreferences due to challenges in maintaining consistent representations across sentences and accurately resolving referential dependencies.

Task description and examples. The task and scope are explicitly described, reducing ambiguity for the LLM: to extend the KG based on the text by adding knowledge/triples about the heuristic at hand. The specific heuristic is defined, and inference examples are provided both in natural language and in their formal representation in OWL2, illustrating the expected output. These examples support *in-context learning* acting as *few-shot examples*, showing how the task should be performed and helping the model infer patterns and apply them to the provided input.

Output format instructions. The last section provides guidance on the expected structure and format of the output. *Clarifying instructions* are thus used to refine the model's output. Naming conventions for individuals, classes, and properties are prescribed to ensure consistency. Rules for ontology alignment and property declaration are explicitly mentioned. The model is explicitly instructed to produce output that only includes newly generated triples in a specified format, aiming at ensuring the results are directly usable in downstream processes without additional cleaning or parsing. The model is explicitly told what to add and what to avoid, and these explicit constraints aim at enforcing a layer of *validation through rules*.

3.4. Implementation and technical setup

Our approach to converting natural language text and images into OWL-RDF KGs integrates diverse services and tools within a unified processing framework. In the context of the overall pipeline shown in Fig. 2, the Text2AMR2FRED pipeline is a central component, designed and implemented to transform text into an RDF/OWL KG, which is then extended by an LLM based on the identified enrichment heuristics. As outlined before, the pipeline leverages state-of-the-art tools for AMR parsing, entity linking, AMR-to-OWL/OWL conversion, and word sense disambiguation. All components of the implemented pipeline are built using "off-the-shelf" tools and services, relying on publicly available pre-trained models. Below, we provide some insights into the configuration and settings of each component, as a blueprint for replicating and refining the pipeline.

The SPRING AMR parser serves as the backbone for generating AMR structures from textual inputs. We directly rely on the available implementation[13] and we perform text-to-AMR parsing using the provided model pre-trained on the AMR 3.0 dataset[14] and configured to produce AMR graphs serialized in PENMAN notation. In line with the approach adopted in SPRING, we integrate the entity linking step in the AMR parsing stage relying on the BLINK Entity Linker to handle *wikification*, i.e., the linking of nodes representing named entities in the AMR graph to Wikipedia pages. BLINK is used in its default configuration,[15] relying on the available pre-trained models and indexes[16] based on a specific Wikipedia dump. For each named entity mention, BLINK provides the top 10 matching entities, and the best match (if identified) is used for wikification in the AMR graph.

AMR2FRED is then responsible for converting AMR graphs into OWL-RDF KGs. This process leverages a predefined, deterministic set of conversion rules that are designed to: *(i)* align with the knowledge extraction and the logical and ontological representation patterns of the FRED machine reading system, and *(ii)* rely on the Framester knowledge graph as a public ontological resource to disambiguate the formal semantics of concepts and relations in the KG. The conversion and mapping rules define the process of translating the elements of AMR graphs (such as variables, concepts, predicates, and roles) into RDF triples, ensuring a structured and semantically sound representation. The conversion pipeline addresses two primary scenarios: (1) the mapping of predicate senses and predicate-specific core roles to their semantic counterpart as PropBank rolesets and local roles (available as part of the Framester resource); and (2) the translation of non-core roles, along with their associated nodes in the AMR graph. For this latter case, AMR2FRED implements a comprehensive translation process consisting of 15 distinct steps with more than 200 condition checks to account for all known edge cases and structural nuances in AMR graphs, ensuring that both nodes and relations are accurately translated into logically valid OWL semantics. For additional details on the conversion rules and implementation of this pipeline, please refer to the AMR2FRED GitHub repository.[17]

In a post-processing stage, the KG is refined and enriched by disambiguating underspecified concepts in the graph and linking them to WordNet synsets and to DOLCE foundational classes, as formalized in Framester. Entity disambiguation with respect to WordNet synsets relies on a word sense disambiguation (WSD) step performed over the initial input text using EWISER. As for SPRING, we directly rely on the available implementation[18] and we perform WSD using the provided pre-trained English model.[19] We specifically rely on EWISER as a Spacy plugin, with Spacy configured to use the `en_core_web_trf` English transformer pipeline.[20]

With the exception of AMR2FRED, implemented in Java, all components in the knowledge enrichment pipeline are implemented in Python. The overall pipeline that supports the results presented in this paper runs as a set of Docker containers hosted on a high end server with 112 cores, 2 NVIDIA A100 40 GB GPUs and 768 GB of RAM.

[13] https://github.com/SapienzaNLP/spring

[14] https://github.com/SapienzaNLP/spring?tab=readme-ov-file#text-to-amr-parsing

[15] https://github.com/facebookresearch/BLINK?tab=readme-ov-file#4-use-blink-in-your-codebase

[16] https://github.com/facebookresearch/BLINK?tab=readme-ov-file#2-download-the-blink-models

[17] https://github.com/infovillasimius/amr2Fred

[18] https://github.com/SapienzaNLP/ewiser

[19] The English checkpoint named "SemCor + tagged glosses + WordNet Examples" available at https://github.com/SapienzaNLP/ewiser?tab=readme-ov-file#externally-downloadable-resources.

[20] https://spacy.io/models/en#en_core_web_trf

Fig. 3. Gold medal winner Julien Alfred in the 100 m female competition at Paris 2024 Olympic games.(©Getty Images).

3.5. Discussion

A relatively long modular pipeline comes with inherent challenges and trade-offs. One potential drawback is the increased complexity in integration and coordination between modules, with the need to define and maintain consistent interfaces and data formats. Another concern is the potential accumulation of dependencies, inefficiencies and errors impacting runtime performance and quality of the final output. A pipeline relying on multiple external modules also pose challenges in maintaining usability for end-users, as this could lead to overwhelming or unintuitive configuration options, and require domain expertise to configure and adapt individual modules. The modular approach may lead to a system that is highly dependent on the capabilities of its individual components, inadvertently limiting the generality of the system.

Nevertheless, a modular pipeline offers significant advantages compared to a monolithic end-to-end black-box approach. The modular nature of the pipeline allows for individual components to be updated or replaced without disrupting the entire system. As advancements are made in fields like LLMs, AMR parsers, entity linking and other relevant technologies, these can be integrated into the pipeline. This ensures the system remains state-of-the-art without requiring a complete overhaul or without the need for extensive retraining or customization of a single model. A modular pipeline also provides granular control over each stage of the process, enabling to test and refine specific components independently. As a concrete example, this allows for evaluating the impact of different LLMs as discussed in this work. While the current implementation relies on general-purpose pre-trained models (as detailed before), the system can seamlessly integrate updates and improvements as newer versions of these pre-trained models become available. Similarly, each component or model can be fine-tuned to improve performance in specific applications or domains. The pipeline can be configured to suit various tasks or domains by swapping out, reconfiguring or specializing specific components or underlying models. The clear delineation of stages in the pipeline also enhances transparency and interpretability. Each module's outputs can be inspected and optimized independently, making it easier to identify potential issues and gain a clear understanding of how the KG is constructed.

4. Experimental evaluation

In this section, we evaluate the output generated by the full Polanyi's pipeline described in Section 3. To do so, we conduct three experiments.

In experiment 1 we start with an image input and we provides a detailed overview of the entire knowledge extension process, culminating in the complete enrichment of XKGs using all 11 heuristics.

We adopt a comprehensive three-tiered evaluation framework designed to rigorously assess our model's performance, and validate the integrity of the XKGs. The evaluation process encompasses: (i) logical validation of the triples, encompassing both syntactic correctness and proper anchoring to pre-existing nodes in the Base Graph; (ii) evaluation of foundational ontology alignment adequacy, and (iii) human assessment of the plausibility and adequacy of assertions within the triples generated for each heuristic.

To further ensure the robustness of our evaluation, we perform graph extension using an image captured after May 2024, shown in Fig. 3. This temporal selection is significant as it postdates the release of GPT-4o, thereby guaranteeing that the image is not part of the model's training dataset.

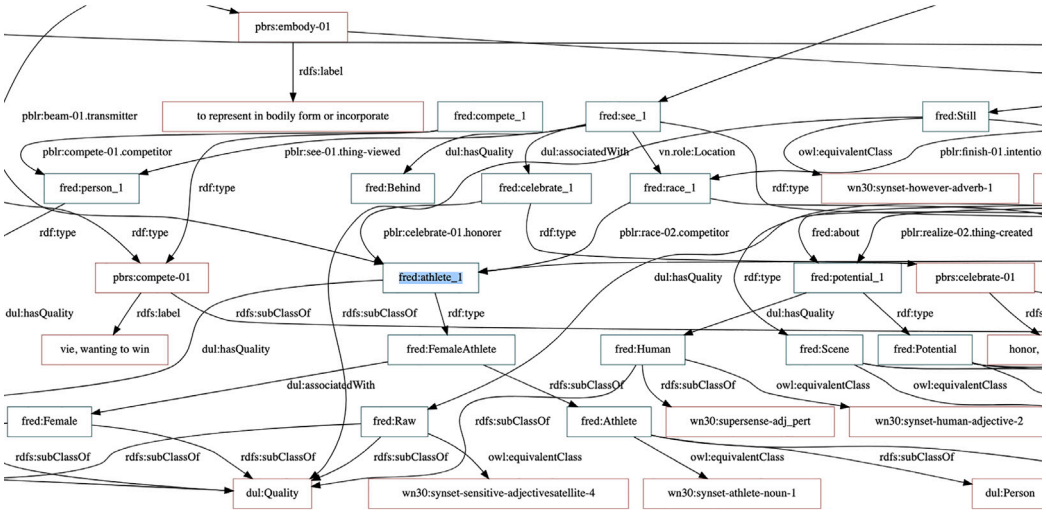

Fig. 4. An excerpt of the AMR2FRED RDF graph.

Due to space constraints, we present a detailed analysis and evaluation of a single knowledge extension instance, utilizing the image shown in Fig. 3, from the "sport" domain. Additional examples of knowledge extension, particularly focusing on "politics" and "everyday life" topics, are available in our dedicated GitHub repository.[21] We chose this image since it captures a moment of high emotional intensity. The athlete's expression clearly conveys what happened moments before and allows for informed speculation about subsequent events based on commonsense knowledge.

Experiment 2 starts from the same image, textual description, and base graph generated with Text2AMR2FRED, as in Experiment 1, but the graph enrichment is done not only via Claude 3.5 Sonnet, but also with GPT-4o and with the open-source Mistral Large 2 model (last update November 2024). Following He et al. (2023), Zhang, Zuo, Lin, and Wu (2024), these same models are then used as synthetic annotators, to check plausibility of the asserted enrichment triples. We define an evaluation framework where we measure and provide insights about triples production, mean ratings and standard deviation, as wells as "self-coherence", namely scores of each model on the triples originally generated by itself, inter-rater (pairwise) agreement, and score distribution.

In Experiment 3, finally, we test our approach on a downstream task: predicting plausible future events, given a scenario described in a text. We use as input 12 pieces of news from the New York Times, again, as the input image in Experiment 1 and 2, posterior to the latest LLMs' update, to ensure them not being part of the training set.

4.1. Experiment 1

Following the pipeline shown in Fig. 2, we give as input the image shown in Fig. 3 to GPT-4o, and we get the textual description shown in Box 1. To provide an example, highlighted in red is the anchoring text for Factual Impact knowledge, while in blue is the anchoring text for Moral Value-driven coercions.

Box 1 - GPT Textual Description

Jubilant female athlete celebrating victory on track, wearing Saint Lucia uniform with ''ALFRED'' on jersey, arms outstretched in triumphant pose, beaming with joy and exhilaration, fellow competitors visible behind her still finishing race, crowded stadium with cheering spectators in background, American flag visible, symbolic moment of personal and national pride, representation of hard work and dedication paying off, embodiment of Olympic spirit and international competition, capturing raw emotion and thrill of athletic achievement at highest level, inspiring scene of human potential realized.

Base graph. The textual description is then passed to Text2AMR2FRED to produce the RDF graph, available as additional material on GitHub.[22]

Fig. 4 shows an excerpt of the RDF graph obtained, in which it is possible to see the alignments to WordNet and PropBank, and to the DOLCE foundational ontology. Box 2 shows an excerpt of the Base Graph, in which it is possible to see how e.g. the individual node `fred:athlete_1` is aligned to DOLCE `dul:Person` and WordNet `wn:synset-athlete-noun-1`. Furthermore, at the

[21] https://github.com/StenDoipanni/XKG/tree/main/additional_use_cases

[22] Download and open in the browser this file: https://github.com/StenDoipanni/XKG/blob/main/resources/graphs/base-graph.html.

Table 1
Triples generated for the Base Graph detailed with origin of the arches and nodes.

Graph	Axioms	WordNet	PB roles	PB frames	VN roles	D0	DUL
Base Graph	293	33	23	20	1	4	10

bottom of Box 2 there is a natural language transposition of the triples above. In red the anchoring points for Factual Impact knowledge (among others), and in blue the anchor for knowledge related to Moral Value-driven coercion.

The Base Graph contains 293 OWL axioms (non-structural statements) as shown in Table 1, which correspond to 1436 RDF triples when serialized (including type declaration, label annotations, etc.). Out of these axioms, 21 are equivalence axioms, expressed via the `owl:equivalentClass` property.

Box 2 - Base graph

```
fred:Athlete rdfs:subClassOf dul:Person,
    wn30:supersense-noun_person ;
    owl:equivalentClass wn30:synset-athlete-noun-1 .

fred:celebrate_1 a pbrs:celebrate-01 ;
    vn.role:Location fred:track_1 ;
    pblr:celebrate-01.honored fred:win_1 ;
    pblr:celebrate-01.honorer fred:athlete_1 .

The athlete is a Person. The gesture of celebration takes place on the track. The victory is what is celebrated, the
athlete is the one celebrating.
```

Considering semantic web resources, the most retrieved are WordNet, PropBank and DOLCE: 33 entities from WordNet are retrieved, regarding both physical entities such as `wn:synset-flag-noun-1`, `wn:synset-uniform-noun-1` and `wn:synset-athlete-noun-1`, and more abstract ones such as `wn:synset-joy-noun-1`, and `wn:synset-international-adjective-1`.

From PropBank we have 23 local roles, namely instantiations of specific roles of a PropBank frame, including triples like:

`fred:wear_1 pblr:wear-01.clothing fred:uniform_1;`

`pblr:wear-01.wearer fred:athlete_1 .`

which means that the node "wear_1" (namely an occurrence of wearing retrieved by FRED) takes as "clothing" role the node "uniform_1" and as "wearer" role, the node "athlete_1". As for PropBank frames, we have 20 distinct entities, among others: `pbrs:win-01`, `pbrs:achieve-01`, and `pbrs:celebrate-01`. Finally, we retrieve 4 entities from DOLCE 0:`d0:CognitiveEntity`, `d0:Event`, `d0:Characteristic`, and `d0:Activity`; and 10 resources from DOLCE Ultralite, of which three are properties: `dul:associatedWith`, `dul:hasMember`, and `dul:hasQuality`, and 7 are entities, such as `dul:Person`, `dul:Situation`, and `dul:InformationEntity`. All the SPARQL queries to investigate the graphs are available as additional materials on the GitHub.[23]

Extended knowledge graphs. We summarize here some numbers about the 11 XKGs generated with the heuristics; all the graphs are available on GitHub.[24]

Table 2 omits the columns for D0 and DUL, since most of XKGs do not declare alignments to DOLCE classes. Two notable exceptions are: the `dul:precedes` property, used to state sequential order of entities, which is extensively present in Presuppositions as well as Event Sequences, and Implied Future Events XKGs; and the Image Schema XKG, which presents 14 alignments of image schemas to DUL classes; among them: "Balance" subClass of `dul:Quality`, "Collection" subClass of `dul:Collection`, "Cicle" subClass of `dul:Process`, "Path" subClass of `dul:SpaceRegion`, and "Container" as subClass of `dul:PhysicalObject`.

Box 3 shows and excerpt of triples generated in the Factual Impact XKG. We discuss the plausibility and correctness of alignments and triples in the next sections, via a three-tier evaluation.

[23] https://github.com/StenDoipanni/XKG/tree/main/resources/sparql-queries
[24] https://github.com/StenDoipanni/XKG/tree/main/resources/graphs

Table 2
Triples generated for the Base Graph and the 11 heuristics graphs, including axioms count, presence of WordNet entities, PropBank Roles, PropBank frames, and newly introduced Object Properties or Data Properties.

Graph	Axioms	WordNet	PB roles	PB entities	OP	DP
Presuppositions	32	–	–	3	–	11
Conversational Implicatures	36	6	–	–	14	–
Factual Impact	13	5	–	–	3	–
Image Schemas	63	11	–	–	1	–
Metonymic Coercion	26	–	5	5	7	–
Moral Value-driven Coercion	12	–	2	1	3	–
Symbolic Coercion	15	7	–	–	1	–
Event Sequences	15	–	–	–	1	–
Causal Relations	16	1	–	–	1	–
Implied Future Events	14	–	1	1	2	–
Potential Non-events	23	–	5	4	5	–

Box 3 - Factual Impact

```
fred:athlete_1 impact:hasExpectedEmotion impact:Joy,
    impact:Pride ;
    impact:hasExpectedPhysicalState impact:Exhilaration ;
impact:hasExpectedSocialImpact impact:NationalRecognition.

The athlete is expected to feel joy and pride, it is expected to be ecstatic, and to receive some form of national
recognition for winning.
```

A particularly salient aspect of XKGs is the introduction of novel properties, as delineated in the final two columns of Table 2. These columns, labeled OP and DP, represent newly introduced "Object Properties" and "Datatype Properties" respectively, offering insight into the ontological expansion and semantic extension of each knowledge graph.

Table 2 presents an overview of the composition of XKGs. The graphs exhibit significant variation in their structural complexity and semantic richness, reflecting the diverse nature of the conceptual domains being modeled. The "Axioms" column reveals considerable differences in the logical foundations of each graph, ranging from 12 axioms in the Moral Value-driven Coercion graph to a notable 63 in the Image Schemas graph. This substantial difference can be traced back to the nature of the reference image being a sport scene, which likely fosters a richer generation of triples, and representation of knowledge related to spatial relations, bodily movements, and physical interactions—concepts central to Image Schemas. The integration of external lexical-semantic resources is evident, with WordNet entities being prominently featured in several graphs, particularly in Conversational Implicatures (6) and Image Schemas (11). PropBank roles and entities are less frequently incorporated, with the Event Sequences graph showing the highest utilization (5 PB Entities), possibly indicating a focus on action-oriented semantics in depicting the sequential nature of plausible events. The graphs also exhibit varying degrees of expressiveness, as indicated by the introduction of new Object Properties (OP) and Datatype Properties (DP). The Conversational Implicatures graph stands out in this regard, introducing 14 new Object Properties, among which we see `:hasVictory`, `:hasEmotion`, `:hasNationality`, and `:hasSignificance`, showing several semantic layers collapsed in the same XKG, including a huge amount of implicit knowledge that could be further unpacked. Similarly, the Presuppositions graph adds 11 new Data Properties. It is noteworthy the fact that all of them takes as object a boolean 'True' vs 'False', and all of them takes as value 'True', pointing in the direction that, in lack of a certain information, e.g. the exact date in which the stadium was built, it is still possible to represent on the graph this information introducing an axiom like:

```
fred:stadium_1 fred:wasBuiltBefore true.
```

This heterogeneity in graph composition not only reflects the diverse requirements for semantic expressivity across different areas of knowledge representation but also underscores how different domain-related images e.g. sports-centric vs political, could influence the depth and breadth of implicit knowledge extraction.

Logical integrity and foundational ontologies compliance. To ensure the structural integrity and logical consistency of all XKGs, we use the Hermit 1.4.3.456 reasoner (Glimm, Horrocks, Motik, Stoilos, & Wang, 2014) as part of our evaluation framework, as well as OOPS! - OntOlogy Pitfall Scanner (Poveda-Villalón, Gómez-Pérez, & Suárez-Figueroa, 2014).

The reasoner's check is meant to verify soundness and consistency of the newly introduced LLM-generated triples, thereby ensuring that the extended knowledge remains sound and usable for downstream applications and inference tasks. In particular, we use Hermit via Protégé 5.5.0 interface and OOPS! web interface.[25] Passing each and every XKG graph to OOPS!, the pitfall scanner yields minor issues for all of them, mainly related to the absence of metadata such as `rdfs:comment` describing entities.

[25] https://oeg.fi.upm.es/index.php/en/technologies/292-oops/index.html

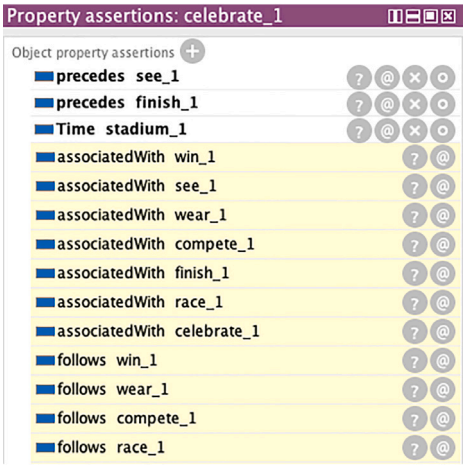

Fig. 5. Protégé visualization of inferences obtained running Hermit 1.2.3.456 on the Event Sequences extended graph.

Occasional instances of hallucination can been observed, particularly in the application of prefixes. An example occurs in the Metonymic coercion file, where the prefix "pbrs" (denoting PropBank Role Set, which PropBank utilized for frame-like structures) is erroneously employed instead of the correct prefix "pblr" (PropBank Local Role), which is the appropriate designation for occurrences of PropBank roles. Such inconsistencies, while minor, underscore the importance of rigorous post-processing and validation in ensuring the accuracy and reliability of the model's semantic annotations, as described in Section 5.

For soundness validation, we checked each XKG importing both the Base Graph and DOLCE Zero, in order to ensure complete coherence with the AMR2FRED original graph, as well as the DOLCE foundational ontology.

An isolated instance of inconsistency is identified in the Metonymic Coercions heuristic graph, suggesting the fact this XKG could present problems on several sides. This anomaly manifested as a conflicting chain of subsumptions ($athlete\_1 \rightarrow Athlete \rightarrow Person \rightarrow Agent \rightarrow Object$) wherein a single entity is erroneously classified as both an Object and an Event. Specifically, 'athlete_1' is incorrectly categorized as an instance of 'wn:cheer-01', rather than being appropriately designated as an agent of the cheering action. It is noteworthy that while the `d0:Event` class is deprecated, we have opted to retain it in our consistency verification process due to its relevance in identifying misalignments.

Other than checking for inconsistencies, found only in the Metonymic Coercion XKG, as mentioned above, reasoning over XKGs with the import of Base Graph and DOLCE 0, yields relevant inferences, as shown in Fig. 5. The Event Sequences XKG, in fact, making use of the `d0:precedes` property, allows to infer the order of events thanks to the transitivity of the property, positioning them in sequential order not explicitly stated. In this case, as shown in Fig. 5, the occurrence of celebration is inferred as sequentially posterior to the racing action, the competition event, the wearing action, and the winning event.

XKGs human evaluation. To ensure the quality and reliability of the generated triples, we conduct a comprehensive human evaluation process. The validation was performed by 5 annotators, all of whom possess proficiency in Resource Description Framework (RDF) and Turtle syntax, but are not domain experts across all 11 heuristics considered. Each triple produced in our extension undergoes scrutiny and is labeled using a 5-point Likert scale, where 1 represents "Not at all plausible/adequate" and 5 indicates "Completely plausible/adequate". This approach allows for a nuanced assessment of the triples' validity and relevance within their respective domains. By employing annotators with RDF expertise but varying levels of domain-specific knowledge, we aims to balance technical accuracy with a generalist perspective, mirroring real-world scenarios where RDF data may be consumed by users with diverse backgrounds.

Furthermore, to ensure the reliability and consistency of our ratings, we employs multiple statistical measures. Inter-rater agreement is calculated to assess the overall concordance among raters. Krippendorff's alpha is chosen for its ability to handle ordinal data and accommodate multiple raters, providing a robust measure of reliability. Cohen's Kappa, while typically used for binary ratings, is adapted to evaluate pairwise agreement between raters. Mean ratings are computed to provide a central tendency measure of the perceived quality of the generated triples. Finally, we calculate the standard deviation to quantify the dispersion of ratings, to get further insight into the consistency or variability of assessments across raters.

Mean ratings and standard deviation. It is important to highlight that, given the 5 point Likert scale, almost all the heuristics overall passed the threshold of 3, with the exception of Implied Future Events, which still shows a rating of $\mu = 2.94$. This means that, overall, the XKGs present at least "fairly plausible" knowledge extension for all the domains. This is *per se* a remarkable achievement, given the disparity among the domains, which coupled with symbolic reasoning inference regarding e.g. sequences of events, mentioned above, opens to very promising further neurosymbolic methodology exploration.

Furthermore, the analysis of mean ratings and standard deviations across various heuristics shown in Fig. 6 reveals interesting patterns in the performance and consistency of different evaluation criteria. Factual Impact, Conversational Implicatures, and Moral

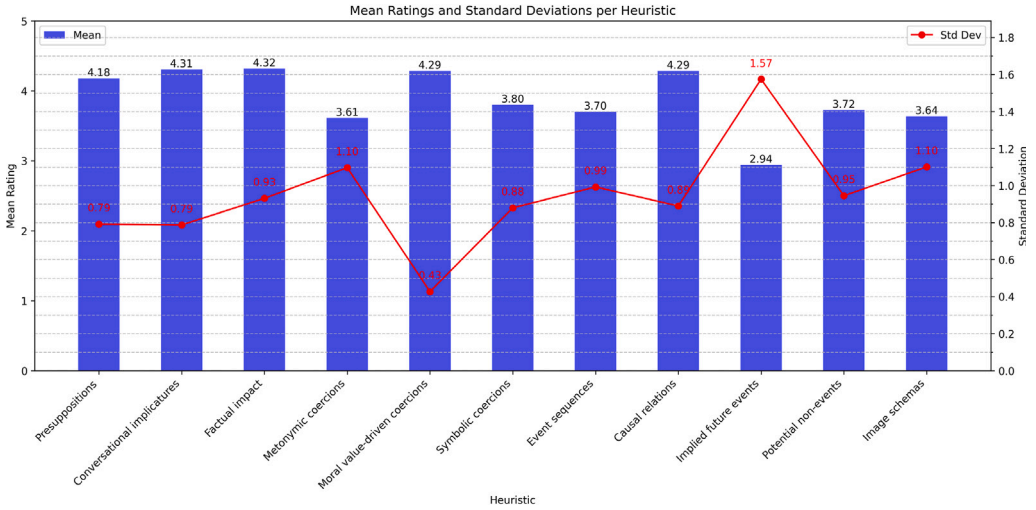

Fig. 6. Mean ratings and standard deviation per heuristics.

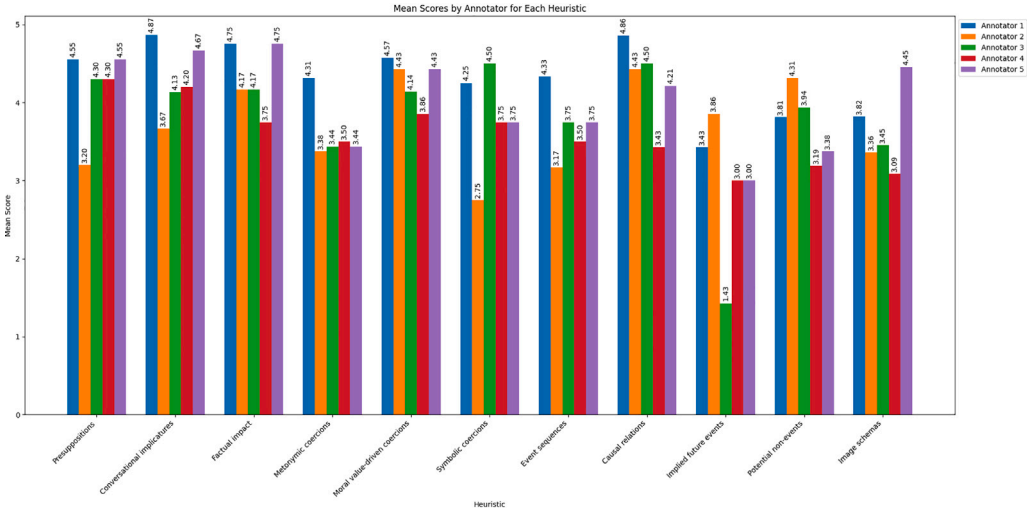

Fig. 7. Mean scores by annotator for each heuristic.

Value-driven Coercions emerge as the top-performing heuristics, with mean ratings exceeding $\mu > 4.29$ on a 5-point scale. This suggests a high degree of plausibility or adequacy in these areas. Conversely, Potential Non-Events and Image Schemas received the lowest mean ratings ($\mu = 2.94$ and $\mu = 3.64$, respectively), indicating potential areas for improvement in the language model's output. Notably, the standard deviations exhibit considerable variation, with Implied Future Events showing the highest variability ($\sigma = 1.57$) and Moral Value-driven Coercions demonstrating the most consistency ($\sigma = 0.47$). This disparity in standard deviations highlights the varying levels of agreement among raters across different heuristics. The data underscores the need for targeted refinements in certain aspects of the language model's performance, particularly in areas with lower mean ratings or higher standard deviations, to enhance the overall quality and consistency of generated content, which is discussed in Section 5. There might be a possible "remoteness effect" when users evaluate uncertain (as with future events) or very abstract (as with image schemas) implicit knowledge, which should be further investigated by involving field experts.

Mean scores per annotator. The analysis of mean scores by annotator for each heuristic, shown in Fig. 7, reveals significant insights into the evaluation process and possibly the nature of the heuristics themselves. Notably, there is considerable variation in scoring patterns across annotators, suggesting potential differences in interpretation or application of the evaluation criteria. Annotator 1 consistently provided higher scores across most heuristics, particularly for Conversational Implicatures (4.87) and Factual Impact (4.75), while Annotator 3 tended to score more conservatively, especially for Implied Future Events (1.43). The heuristics of Factual Impact and Causal Relations demonstrated relatively high consensus among annotators, with most scores clustering above 4.0, indicating their robustness and clarity. Conversely, Implied Future Events and Potential Non-events exhibited the widest disparity

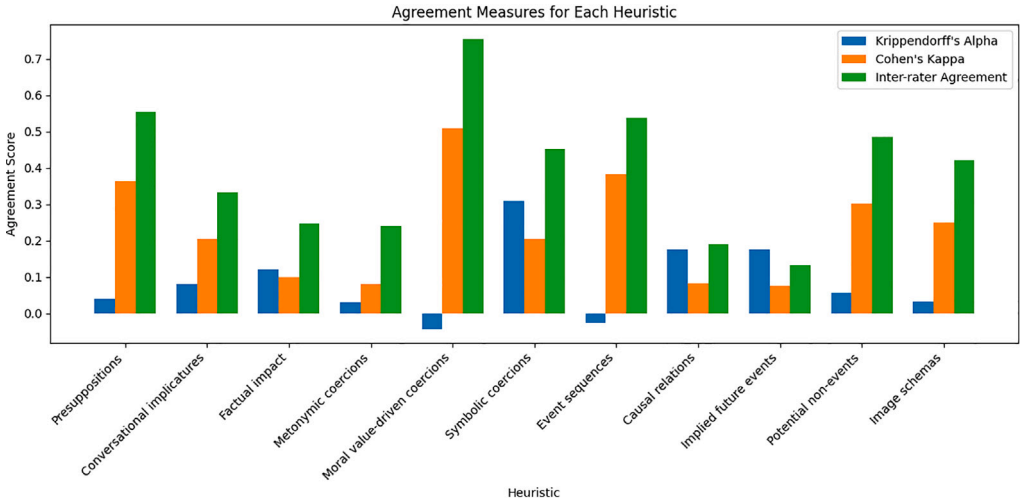

Fig. 8. Agreement measures.

in scores, ranging from 1.43 to 3.86 and 3.19 to 4.31 respectively, highlighting areas where the evaluation criteria may benefit from refinement. This variability underscores the subjective nature of certain heuristics and emphasizes the importance of clear guidelines and calibration sessions in future annotation tasks to enhance inter-rater reliability and the overall validity of the evaluation process.

Furthermore, the scoring patterns reveal distinct tendencies among the five annotators, showcasing varying levels of "generosity" in their evaluations: Annotator 1 emerges as the most generous evaluator, consistently providing the highest scores across most heuristics. This is particularly evident in Conversational Implicatures (4.87), Factual Impact (4.75), and Moral Value-driven Coercions (4.57). Their scores are frequently at least 0.5 points higher than the next highest annotator, suggesting a more lenient interpretation of the evaluation criteria. Annotator 5 appears to be the second most generous, often providing scores close to, but slightly below, Annotator 1. They show particular generosity in Presuppositions (4.55) and Factual Impact (4.75), matching Annotator 1 in the latter. Annotator 2 demonstrates a more moderate approach, typically scoring in the middle range compared to other annotators. However, they show higher scores for Causal Relations (4.43) and Potential Non-Events (4.31), indicating possible areas of expertise or confidence. Annotator 4 tends to be more conservative in their scoring, often providing lower scores than Annotators 1, 2, and 5. This is particularly noticeable in heuristics like Metonymic Coercions (3.44) and Implied Future Events (3.00). However, they align more closely with others on some heuristics like Presuppositions (4.30). Annotator 3 emerges as the most stringent evaluator overall. They consistently provide the lowest scores across multiple heuristics, most notably in Implied Future Events (1.43) and Image Schemas (3.09). This suggests a more critical approach to the evaluation process or possibly a stricter interpretation of the scoring criteria.

Agreement measures. Fig. 8 shows how Moral Value-driven Coercions demonstrate the highest overall agreement, with an inter-rater agreement of 0.75 and a Cohen's Kappa of 0.51, suggesting strong consistency among raters for this heuristic. Conversely, Implied Future Events show the lowest agreement across all measures, indicating a potential need for refinement in its definition or evaluation criteria. Notably, Krippendorff's Alpha consistently yields lower values compared to other measures, with several heuristics showing negative values, particularly for Metonymic Coercions and Moral Value-driven Coercions. This discrepancy between Krippendorff's Alpha and other measures warrants further investigation, as it may indicate sensitivity to specific patterns in the data or potential limitations in applying this metric to the current evaluation framework. The generally moderate to low agreement scores across most heuristics, especially for measures like Factual Impact and Conversational Implicatures, underscore the challenges in achieving consistent evaluations for potentially conceptually complex phenomena.

Following AAAI-20 talk by Henry Kautz,[26] these results confirm that the usage of LLMs in our hybrid pipeline moves significant steps not towards expert reasoning, but mostly towards commonsense reasoning.

4.2. Experiment 2

Starting from the same image (Fig. 3), textual description (Box 1) and Base Graph[27] from Experiment 1, we conducted a second experiment, generating enrichment triples for each heuristic, with three different LLMs: Claude 3.5 Sonnet, GPT-4o, and open source Mistral Large 2 (last update November 2024). After the triples generation, we use these same three models as synthetic annotators, as in He et al. (2023), Zhang et al. (2024), checking again, as in Experiment 1, the inter-rater agreement, mean ratings, and standard

[26] https://www.youtube.com/watch?v=_cQITY0SPiw
[27] https://github.com/StenDoipanni/XKG/blob/main/resources/graphs/base-graph.html

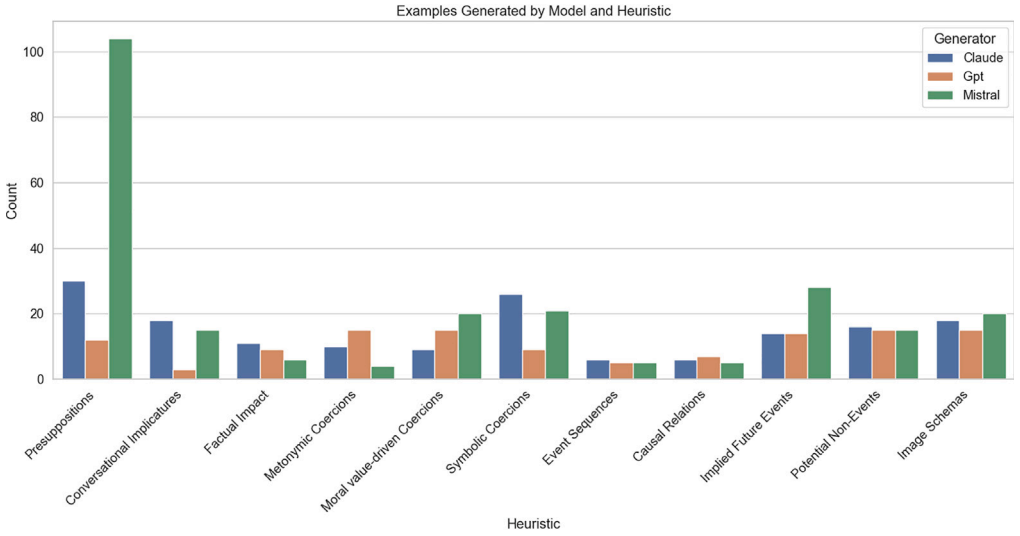

Fig. 9. Productivity in terms of number of triples generated by each model, per each heuristic.

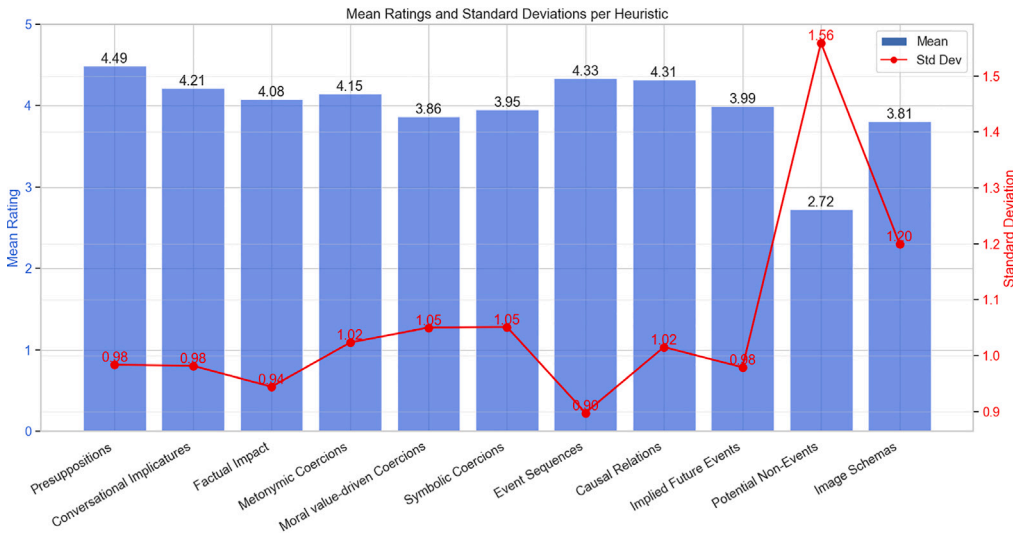

Fig. 10. Mean ratings and standard deviation for Claude, GPT-4o, and Mistral's annotations.

deviation. In addition we analyze score distribution and self-coherence, namely how a model scored the data previously produced by itself. All the diagrams for Experiment 2 are available on GitHub.[28] In the following we summarize and discuss the main findings.

Productivity. The analysis of example generation rates across different heuristics, shown in Fig. 9 reveals distinct productivity patterns among the models, which may help contextualize our findings detailed in the following paragraphs on plausibility assessment and inter-model agreement. While Claude demonstrates the most consistent production rate (10 – 25 examples per heuristic), Mistral shows remarkable variability, most notably producing an exceptional number of examples for Presuppositions (> 100 examples) while maintaining moderate output in other categories. GPT-4o generally exhibits lower and more variable production rates, with slight peaks in Metonymic and Moral value-driven Coercions. Interestingly, certain heuristics that showed high plausibility ratings and strong self-coherence, such as Event Sequences and Causal Relations, consistently elicit fewer examples across all models (5 – 10 examples).

Mean ratings and standard deviation. Our analysis of plausibility heuristics, shown in Fig. 10, reveals distinct patterns in both rating distributions and rater consensus across different heuristics. The data demonstrates generally robust mean ratings ($\mu > 3.80$)

[28] https://github.com/StenDoipanni/XKG/tree/main/aft_rev/outputs/metrics_plotting

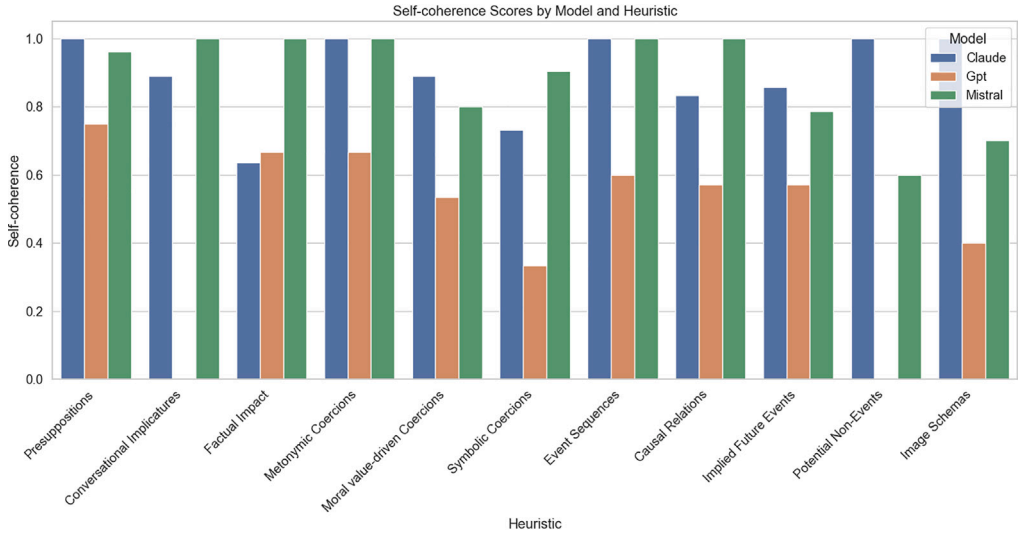

Fig. 11. Self-coherence scores by model and heuristic.

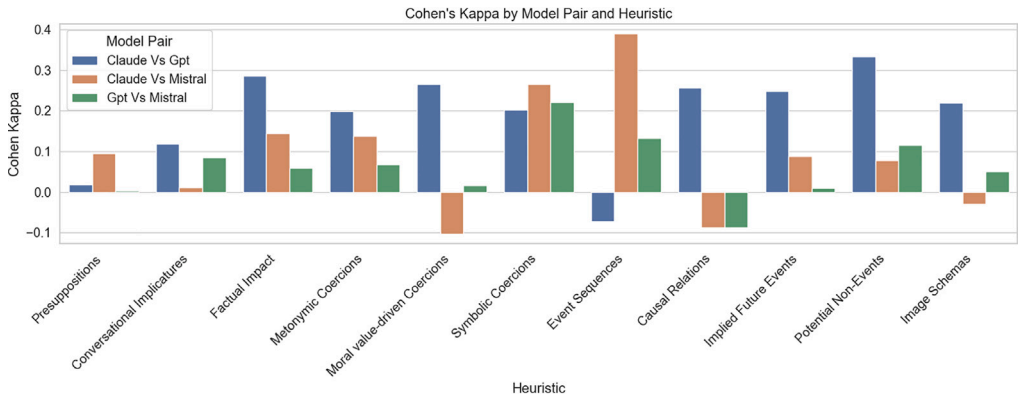

Fig. 12. Cohen's Kappa by model pairs, per each heuristic.

across most heuristics, with Presuppositions achieving the highest mean rating ($\mu = 4.49$) and Event Sequences and Conversational Implicatures following closely ($\mu = 4.33$ and $\mu = 4.21$ respectively). Notable exceptions emerge particularly for Potential Non-Events, which exhibits both the lowest mean rating ($\mu = 2.72$) and the highest standard deviation ($\sigma = 1.56$). This can be explained considering that the evaluation is focused on "plausibility" of the assertion in the triple, and therefore it is coherent that Potential Non-Events, pointing at possible events which are NOT plausible to happen, show low scores. The majority of heuristics demonstrate relatively stable standard deviations ($\sigma \approx 0.98 - 1.05$), with Event Sequences showing the most consistent ratings ($\sigma = 0.90$). The data particularly highlights the robust performance of presuppositional and event-sequence-based heuristics, suggesting these may serve as more reliable indicators in plausibility assessment frameworks.

Self coherence. Analysis of self-coherence metrics across the three language models, shown in Fig. 11, reveals Claude and Mistral as robust self-coherent, with scores > 0.8, with both models achieving perfect coherence (1.0) across several dimensions: Claude in Presuppositions, Metonymic Coercions, Event Sequences, and Potential Non-Events; Mistral in Conversational Implicatures, Factual Impact, Metonymic Coercions, and Implied Future Events. In contrast, GPT-4o exhibits consistently lower self-coherence scores (ranging from 0.35 to 0.75), never achieving perfect coherence in any category. This disparity is particularly pronounced in specific heuristics, notably Symbolic Coercions (GPT-4o ≈ 0.35) and Image Schemas, where all models show their lowest coherence scores (Claude ≈ 0.8, Mistral ≈ 0.7, GPT-4o ≈ 0.4) demonstrating once again, as in Experiment 1, the relevance of domain experts for more technical annotations.

Agreement measures. Inter-rater reliability analysis using Cohen's Kappa, shown in Fig. 12, reveals notably low agreement levels across model pairs in their plausibility assessments, with coefficients predominantly ranging from −0.1 to 0.4. The Claude-GPT pair demonstrates the most consistent positive agreement across heuristics, suggesting similar underlying evaluation criteria, while Mistral's assessments frequently diverge from both counterparts. Maximum agreement is observed in Event Sequences between

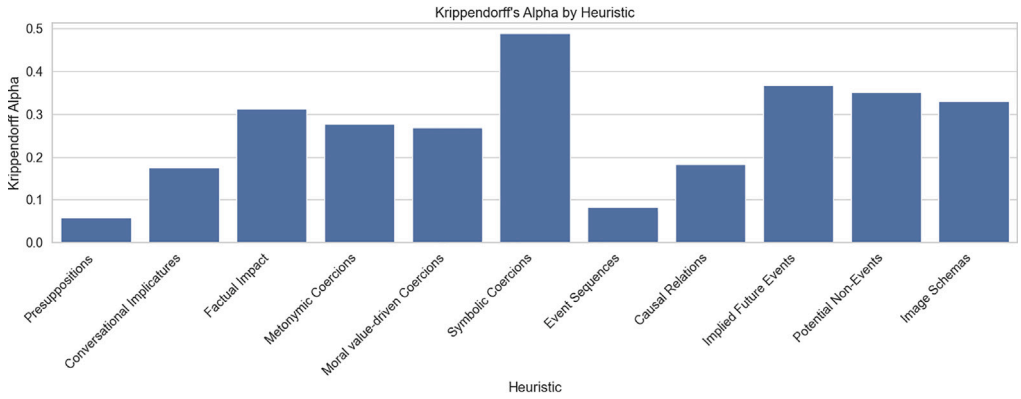

Fig. 13. Krippendorff's Alpha measure by heuristic.

Claude and Mistral ($\kappa \approx 0.4$) and in Potential Non-Events between Claude and GPT-4o ($\kappa \approx 0.35$), while certain heuristics, notably Causal Relations, show negative Kappa values for multiple model pairs, indicating worse-than-chance agreement. The GPT-Mistral pair consistently exhibits the lowest agreement levels, rarely exceeding $\kappa = 0.1$. This systematic variation in inter-rater reliability across both model pairs and heuristics suggests that there remains significant divergence in how different models evaluate specific aspects of plausibility, with particular disagreement in areas involving causal reasoning and image schemas.

Regarding the second measure we adopt, Krippendorff's Alpha analysis, shown in Fig. 13, reveals generally low inter-annotator agreement across all three models, with coefficients ranging from 0.05 to 0.48, indicating substantial divergence in plausibility assessments. Symbolic Coercions emerges as the dimension with highest cross-model consensus ($\alpha \approx 0.48$), followed by Implied Future Events and Potential Non-Events (both $\alpha \approx 0.37$), though none achieve conventionally accepted thresholds for strong agreement ($\alpha > 0.6$). Notably, Presuppositions and Event Sequences demonstrate remarkably low agreement ($\alpha \approx 0.05$ and 0.08 respectively), with the latter presenting an intriguing contrast to its relatively high pairwise agreement in Cohen's Kappa analysis. This discrepancy between pairwise and collective agreement metrics suggests that while certain model pairs may achieve moderate consensus on specific heuristics, establishing consistent evaluation criteria across all three architectures remains challenging.

This low inter-model agreement, coupled with our findings of substantial variation in production rates, particularly Mistral's exceptional output in Presuppositions, suggests that higher productivity does not necessarily correlate with more reliable or consistent plausibility assessment.

Discussion. Our comprehensive analysis reveals a complex interaction among language models and plausibility attribution across different heuristics. While all models demonstrate a tendency toward high mean plausibility ratings ($\mu > 3.8/5$ for most heuristics), confirming the results from Experiment n° 1, the low inter-model agreement (Krippendorff's $\alpha < 0.5$, Cohen's κ typically < 0.4) suggests that models often differ substantially in which specific scenarios they deem most plausible. This divergence is further complicated by varying levels of self-coherence, where newer architectures like Claude and Mistral demonstrate high internal consistency (self-coherence > 0.8), while GPT shows notably lower self-coherence ($0.35 - 0.75$) across all heuristics. This pattern—high overall plausibility ratings coupled with low inter-model agreement and variable self-coherence—suggests that while language models have generally evolved to view scenarios as plausible, they have developed distinctly different internal criteria for annotation's judgments. These findings highlight the need for more rigorous standardization in how we evaluate and compare plausibility judgments across different model architectures, particularly given their increasing deployment in real-world applications where consistent and reliable plausibility assessment is crucial.

Further analysis results including models' scores distribution and pairwise agreement per each heuristics are available on the GitHub.[29]

4.3. Experiment 3

A third experiment was done as a small downstream task: predicting plausible future events from a situation described in a text. We ran it with 12 pieces of news extracted from the New York Times on the same day as when the evaluation was performed to ensure the system did not know what happened later. One of the news has a purely historical context, and has been included as a double-check on the ability of the LLMs to single out, or alternatively describe, apparently 'future-improductive' texts.

Each piece of news has been passed to three LLMs (Claude Sonnet 3.5, ChatGPT-4o, and Gemini 1.5 Pro) with a straightforward prompt to make them educate guess plausible future events. Gemini did not provide any prediction, ChatGPT-4o generated single predictions for 7 news out of 12 (58.3%), Claude generated multiple predicted events for all news. For example, from *The Trump administration is giving Immigration and Customs Enforcement officials the power to quickly deport migrants who were allowed into the*

[29] https://github.com/StenDoipanni/XKG/tree/main/aft_rev/outputs/metrics_plotting

country temporarily under Biden-era programs, according to an internal government memo obtained by The New York Times., Claude guesses: 'Increased deportations'; 'Legal challenges'; 'Migrants avoiding detection'. We have validated all LLM predictions.

The downstream task evaluated the precision of our tool in extracting similar future events, and a 'Structure Multiplication Factor' (SMF), which measures how much our tacit knowledge extraction tool is better structured (in terms of nuanced event complexity) compared to the LLM's natural language prediction. The results are perfectly accurate (100% precision): the tool e.g. extracts all 6 LLM predictions for the Trump/ICE news, and adds 9 more institutional/power dynamics triples, so providing a SMF=2.5. The mean SMF for all news is 2.74 (std=0.37).

5. Ongoing and future work

In our ongoing efforts to refine and enhance Polanyi's pipeline, and the generation of XKGs, several specific improvements are being explored. Prompt refinement for in-context learning has emerged as a crucial area for development, particularly in light of insights gained from image schemas. Our current approach utilizes a standard template, but evidence suggests that more adapted heuristic-specific prompts could yield superior results.

Another area of focus is the refinement of property assignments. Currently, many triples are generated using the broad `dul:associatedWith` top property. Efforts are underway to specialize this property further, aligning newly introduced properties with DOLCE Zero. While this integration may increase the risk of ontological inconsistencies, it also promises enhanced inferential capabilities.

We are also reevaluating our validation methodology. The current system,[30] which informs annotators of the heuristic definition they are validating, may be influencing results. An alternative approach, such as presenting uncontextualised triples for evaluation, could potentially yield different outcomes in human assessment.

Lastly, we are exploring alignment with domain-specific ontologies, particularly those focused on Image Schemas (De Giorgis, Gangemi, & Gromann, 2024), moral and cultural values (De Giorgis, Gangemi, & Damiano, 2022), and other cognitive entities such as emotions. This could involve refining prompts to incorporate either complete ontology schemas (when feasible) or at least excerpts of top taxonomy classes, potentially leading to more nuanced and domain-specific knowledge.

Future work will focus on several key areas to enhance the capabilities and applicability of our graph enrichment process. A primary objective is the topicalization of enrichment when starting from images, which involves localizing triples related to specific heuristics to particular portions of an image. This can be achieved through the utilization of existing labeled repositories such as Visual Genome or ImageNet, or by implementing object-scene recognition algorithms that disambiguate to WordNet synsets before proceeding with triple generation.

Addressing the challenges of Wikidata entry alignment is another crucial area of development. While AMR2FRED proves reliable for the Base Graph, LLM-based alignment has shown significant inconsistencies, highlighting issues with precise information retrieval and hallucinations. A proposed solution involves using the "wd:" prefix with entity labels, followed by script-based verification using SPARQL engines like Qlever (Bast & Buchhold, 2017). This method, incorporating CamelCase string parsing, shows promise for specific named entities but may require further refinement for general conceptual entities.

Although our method is model-agnostic, practical results vary across different LLMs. Current implementation relies on state-of-the-art models requiring proprietary APIs. Future efforts will focus on refining prompts and segmenting processes to enable the use of smaller yet efficient models like Phi 3.5, potentially broadening accessibility.

Lastly, we envision expanding our service to accommodate user-defined prompts, opening up a vast array of possibilities for knowledge graph extension. This could include specialized applications such as e.g., color coercions for environmental elements (we assume that if there is a portion of image which represents the sea, or a forest, or snow, etc. there could be a specific color palette) or e.g. identifying potential safety hazards in images (this could have relevant impact in social robotics). The potential applications are diverse and hold significant relevance across various domains, underscoring the expansive capabilities of our approach in knowledge representation and reasoning.

6. Conclusions

Our research presents a significant step forward in the development of a hybrid neurosymbolic method for grounded world models that can effectively integrate multiple layers of knowledge, bridging the gap between the pattern-matching reactive capabilities of LLMs, and the structured reasoning applied on Knowledge Bases and traditional symbolic reasoning. By leveraging LLMs as repositories of implicit commonsense knowledge rather than expert systems, we have demonstrated a novel approach to knowledge base extension that is both agile and comprehensive.

These results also add to a growing corpus of results about the hidden correspondence of massive linguistic knowledge to the structure of human environments, be them physical, social, cognitive or purely abstract.

The multi-tiered evaluation framework we employed, encompassing logical validation, foundational ontology alignment, and human assessment, provides strong evidence for the efficacy of our approach and future promising refinements. The high plausibility ratings across most heuristics, particularly in areas such as Factual Impact, Conversational Implicatures, and Moral Value-driven Coercions, underscore the potential of our method to capture nuanced semantic information that goes beyond simple fact retrieval.

[30] Available here: https://github.com/StenDoipanni/XKG/blob/main/resources/XKG-human-validation-form.pdf.

However, the challenges revealed in our evaluation, particularly in areas like Implied Future Events and Image Schemas, point to important directions for future work. The variability in inter-rater agreement across different heuristics highlights the complexity of evaluating implicit knowledge and the need for continued refinement of our methods and evaluation criteria, even if uncertainty (for future events) and abstractness (for image schemas) of implicit knowledge are probably good reasons for lower agreement.

The neuro-symbolic nature of our approach, combining the strengths of neural networks and symbolic reasoning, opens up new possibilities for systems that can flexibly adapt to novel contexts while maintaining the ability to perform structured inference. This hybrid architecture addresses longstanding challenges in AI, such as the Frame problem, by providing a mechanism for dynamically integrating diverse knowledge sources and reasoning patterns.

Looking ahead, our work lays the foundation for several promising research directions. The potential for topicalization of enrichment on images, improved alignment with domain-specific ontologies, and the exploration of user-defined prompts for specialized applications all represent exciting avenues for extending the capabilities of our system.

In conclusion, our research contributes to the broader goal of developing systems capable of human-like reasoning and contextual understanding relevant in several areas ranging from natural language processing and computer vision to complex problem-solving and decision-making in real-world scenarios.

CRediT authorship contribution statement

Stefano De Giorgis: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Aldo Gangemi:** Writing – review & editing, Supervision, Software, Project administration, Methodology, Funding acquisition, Conceptualization. **Alessandro Russo:** Writing – review & editing, Software, Methodology, Conceptualization.

Acknowledgments

This work was supported by the Next Generation EU Program with the Future Artificial Intelligence Research (FAIR) project, code PE00000013, CUP 53C22003630006, and by the Italian Research Center on High Performance Computing, Big Data and Quantum Computing (ICSC).

Data availability

No data was used for the research described in the article.

References

Alam, Mehwish, Gesese, Genet Asefa, & Paris, Pierre-Henri (2024). Neurosymbolic methods for dynamic knowledge graphs. arXiv preprint arXiv:2409.04572.

Auer, Sören, Bizer, Christian, Kobilarov, Georgi, Lehmann, Jens, Cyganiak, Richard, & Ives, Zachary (2007). Dbpedia: A nucleus for a web of open data. In *The semantic web* (pp. 722–735). Springer.

Baker, Collin F., Fillmore, Charles J., & Lowe, John B. (1998). The berkeley framenet project. In *Proceedings of the 17th international conference on computational linguistics-volume 1* (pp. 86–90). Association for Computational Linguistics.

Bast, Hannah, & Buchhold, Björn (2017). Qlever: A query engine for efficient sparql+ text search. In *Proceedings of the 2017 ACM on conference on information and knowledge management* (pp. 647–656).

Belle, Vaishak (2020). Symbolic logic meets machine learning: A brief survey in infinite domains. In *International conference on scalable uncertainty management* (pp. 3–16). Springer.

Besold, Tarek R., d'Avila Garcez, Artur, Bader, Sebastian, Bowman, Howard, Domingos, Pedro, Hitzler, Pascal, et al. (2021). Neural-symbolic learning and reasoning: A survey and interpretation 1. In *Neuro-symbolic artificial intelligence: the state of the art* (pp. 1–51). IOS Press.

Besold, Tarek R., Hedblom, Maria M., & Kutz, Oliver (2017). A narrative in three acts: Using combinations of image schemas to model events. *Biologically Inspired Cognitive Architectures*, *19*, 10–20.

Bevilacqua, Michele, Blloshmi, Rexhina, & Navigli, Roberto (2021). One SPRING to rule them both: Symmetric AMR semantic parsing and generation without a complex pipeline. In *Proceedings of the AAAI conference on artificial intelligence*: *vol. 35*, (pp. 12564–12573).

Bevilacqua, Michele, & Navigli, Roberto (2020). Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the 58th annual meeting of the association for computational linguistics, Online* (pp. 2854–2864). Association for Computational Linguistics.

Bos, Johan (2008). Wide-coverage semantic analysis with boxer. In *Semantics in text processing. Step 2008 conference proceedings* (pp. 277–286).

Brachman, Ronald J. (1977). What's in a concept: structural foundations for semantic networks. *International Journal of Man-Machine Studies*, *9*(2), 127–152.

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

De Giorgis, Stefano, Gangemi, Aldo, & Damiano, Rossana (2022). Basic human values and moral foundations theory in valuenet ontology. In *International conference on knowledge engineering and knowledge management* (pp. 3–18). Springer.

De Giorgis, Stefano, Gangemi, Aldo, & Gromann, Dagmar (2024). Imageschemanet: a framester graph for embodied commonsense knowledge. *Semantic Web*, *15*(4), 1417–1441.

Edge, Darren, Trinh, Ha, Cheng, Newman, Bradley, Joshua, Chao, Alex, Mody, Apurva, et al. (2024). From local to global: A graph rag approach to query-focused summarization. arXiv preprint arXiv:2404.16130.

Eliot, Thomas Stearns, et al. (1920). Hamlet and his problems. In *The sacred wood: essays on poetry and criticism*: *vol. 4*, (pp. 95–104).

Fillmore, Charles J., et al. (2006). Frame semantics. In *Cognitive linguistics: basic readings*: *vol. 34*, (pp. 373–400).

Forrester, Jay W. (1971). Counterintuitive behavior of social systems. *Theory and Decision*, *2*(2), 109–140.

Frege, Gottlob (1892). On sense and reference.

Gangemi, Aldo (2020). Closing the loop between knowledge patterns in cognition and the semantic web. *Semantic Web*, *11*(1), 139–151.

Gangemi, Aldo, Alam, Mehwish, Asprino, Luigi, Presutti, Valentina, & Recupero, Diego Reforgiato (2016). Framester: A wide coverage linguistic linked data hub. In *Knowledge engineering and knowledge management: 20th international conference, EKAW 2016, Bologna, Italy, November 19-23 2016, proceedings 20* (pp. 239–254). Springer.

Gangemi, Aldo, Graciotti, Arianna, Meloni, Antonello, Nuzzolese, Andrea Giovanni, Presutti, Valentina, Recupero, Diego Reforgiato, et al. (2023). Text2AMR2FRED, a tool for transforming text into RDF/OWL knowledge graphs via abstract meaning representation. In *Proceedings of the ISWC 2023 posters, demos and industry tracks: from novel ideas to industrial practice co-located with 22nd international semantic web conference*.

Gangemi, Aldo, Guarino, Nicola, Masolo, Claudio, & Oltramari, Alessandro (2003). Sweetening wordnet with dolce. *AI Magazine*, *24*(3), 13.

Gangemi, Aldo, & Nuzzolese, Andrea Giovanni (2025). Logic augmented generation. *Journal of Web Semantics*, *85*, Article 100859.

Gangemi, Aldo, Presutti, Valentina, & Alam, Mehwish (2018). Amnestic forgery: An ontology of conceptual metaphors. In *Formal ontology in information systems - proceedings of the 10th international conference, FOIS 2018, Cape Town, South Africa, 19-21 2018*. IOS Press.

Gangemi, Aldo, Presutti, Valentina, Recupero, Diego Reforgiato, Nuzzolese, Andrea Giovanni, Draicchio, Francesco, & Mongiovì, Misael (2017). Semantic web machine reading with FRED. *Semantic Web*, *8*(6), 873–893.

Gangemi, Aldo, Recupero, Diego Reforgiato, Mongiovì, Misael, Nuzzolese, Andrea Giovanni, & Presutti, Valentina (2016). Identifying motifs for evaluating open knowledge extraction on the web. *Knowledge-Based Systems*, *108*, 33–41.

Gangemi, Aldo, et al. (2017). AMR2FRED, A tool for translating abstract meaning representation to motif-based linguistic knowledge graph. In *Proceedings of the extended semantic web conference* (pp. 43–47). DEU.

Glimm, Birte, Horrocks, Ian, Motik, Boris, Stoilos, Giorgos, & Wang, Zhe (2014). Hermit: an owl 2 reasoner. *Journal of automated reasoning*, *53*, 245–269.

Graham, Jesse, Haidt, Jonathan, Koleva, Sena, Motyl, Matt, Iyer, Ravi, Wojcik, Sean P., et al. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology: vol. 47*, (pp. 55–130). Elsevier.

Grau, Bernardo Cuenca, Horrocks, Ian, Motik, Boris, Parsia, Bijan, Patel-Schneider, Peter, & Sattler, Ulrike (2008). Owl 2: The next step for owl. *Journal of Web Semantics*, *6*(4), 309–322.

Grice, H. P. (1975). Logic and conversation. *Syntax and Semantics*, *3*.

Ha, David, & Schmidhuber, Jürgen (2018). World models. arXiv preprint arXiv:1803.10122.

Hamilton, Kyle, Nayak, Aparna, Božić, Bojan, & Longo, Luca (2022). Is neuro-symbolic AI meeting its promises in natural language processing? a structured review. *Semantic Web, (Preprint)*, 1–42.

He, Xingwei, Lin, Zhenghao, Gong, Yeyun, Jin, Alex, Zhang, Hang, Lin, Chen, et al. (2023). Annollm: Making large language models to be better crowdsourced annotators. arXiv preprint arXiv:2303.16854.

Hedblom, Maria M. (2020). *Image schemas and concept invention: cognitive, logical, and linguistic investigations*. Springer Nature.

Johnson, Mark (1987). *The body in the mind: the bodily basis of meaning, imagination, and reason*. Chicago and London: The University of Chicago Press.

Kahneman, Daniel (1991). Article commentary: Judgment and decision making: A personal view. *Psychological Science*, *2*(3), 142–145.

Karttunen, Lauri (2016). Presupposition: What went wrong? In *Semantics and linguistic theory* (pp. 705–731).

Kautz, Henry (2022). The third AI summer: AAAI Robert S. Engelmore Memorial Lecture. *AI Magazine*, *43*(1), 105–125.

Kingsbury, Paul R., & Palmer, Martha (2002). From treebank to propbank. In *LREC* (pp. 1989–1993).

Lakoff, George, & Johnson, Mark (1980). *Metaphors we live by*. University of Chicago Press.

Lakoff, George, Johnson, Mark, et al. (1999). *vol. 640*, *Philosophy in the flesh: the embodied mind and its challenge to western thought*. New York: Basic Books.

Lamb, Luís C., Garcez, Artur, Gori, Marco, Prates, Marcelo, Avelar, Pedro, & Vardi, Moshe (2020). Graph neural networks meet neural-symbolic computing: A survey and perspective. arXiv preprint arXiv:2003.00330.

Lambek, Joachim (1988). Categorial and categorical grammars. In *Categorial grammars and natural language structures* (pp. 297–317). Springer.

Levinson, Stephen C. (2000). *Presumptive meanings: the theory of generalized conversational implicature*. MIT Press.

Lu, Fengyuan, Cong, Peijin, & Huang, Xinli (2020). Utilizing textual information in knowledge graph embedding: A survey of methods and applications. *IEEE Access*, *8*, 92072–92088.

Maudslay, Rowan Hall, Teufel, Simone, Bond, Francis, & Pustejovsky, James (2024). Chainnet: Structured metaphor and metonymy in wordnet. arXiv preprint arXiv:2403.20308.

Meloni, Antonello, Recupero, Diego Reforgiato, & Gangemi, Aldo (2017). AMR2fred, a tool for translating abstract meaning representation to motif-based linguistic knowledge graphs. In *The semantic web: ESWC 2017 satellite events: ESWC 2017 satellite events, Portorož, Slovenia, May 28–June 1 2017, revised selected papers 14* (pp. 43–47). Springer.

Meyer, Lars-Peter, Stadler, Claus, Frey, Johannes, Radtke, Norman, Junghanns, Kurt, Meissner, Roy, et al. (2023). Llm-assisted knowledge graph engineering: Experiments with chatgpt. In *Working conference on artificial intelligence development for a resilient and sustainable tomorrow* (pp. 103–115). Springer Fachmedien Wiesbaden Wiesbaden.

Miller, George A. (1998). *WordNet: an electronic lexical database*. MIT Press.

Minsky, Marvin, et al. (1974). A framework for representing knowledge.

Motta, Enrico, Daga, Enrico, Gangemi, Aldo, Gjelsvik, M. L., Osborne, Francesco, & Salatino, Angelo (2024). The epistemology of fine-grained news classification. *Semantic Web*.

Navigli, Roberto, & Ponzetto, Simone Paolo (2010). Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 216–225).

Nolfi, Stefano (2023). On the unexpected abilities of large language models. *Adaptive Behavior*, Article 10597123241256754.

Pan, Shirui, Luo, Linhao, Wang, Yufei, Chen, Chen, Wang, Jiapu, & Wu, Xindong (2024). Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*.

Peirce, Charles Sanders, & Buchler, Justus (1955). Logic as semiotic: The theory of signs. In Justus Buchler (Ed.), *Philosophical writings of peirce* (p. 100). New York: Dover, 1902.

Polanyi, Michael (2009). The tacit dimension. In *Knowledge in organisations* (pp. 135–146). Routledge.

Poveda-Villalón, María, Gómez-Pérez, Asunción, & Suárez-Figueroa, Mari Carmen (2014). Oops!(ontology pitfall scanner!): An on-line tool for ontology evaluation. *International Journal on Semantic Web and Information Systems (IJSWIS)*, *10*(2), 7–34.

Rozin, Paul, Lowery, Laura, Imada, Sumio, & Haidt, Jonathan (1999). The CAD triad hypothesis: a mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). *Journal of Personality and Social Psychology*, *76*(4), 574, Publisher: American Psychological Association.

Sarker, Md Kamruzzaman, Zhou, Lu, Eberhart, Aaron, & Hitzler, Pascal (2021). Neuro-symbolic artificial intelligence. *AI Communications*, *34*(3), 197–209.

Schuler, Karin Kipper (2005). *VerbNet: a broad-coverage, comprehensive verb lexicon*. University of Pennsylvania.

Schwartz, Shalom H., Melech, Gila, Lehmann, Arielle, Burgess, Steven, Harris, Mari, & Owens, Vicki (2001). Extending the cross-cultural validity of the theory of basic human values with a different method of measurement. *Journal of Cross-Cultural Psychology*, *32*(5), 519–542, Publisher: Sage Publications Sage CA: Thousand Oaks, CA.

Sen, Priyanka, Mavadia, Sandeep, & Saffari, Amir (2023). Knowledge graph-augmented language models for complex question answering. In *Proceedings of the 1st workshop on natural language reasoning and structured explanations* (pp. 1–8).

Strawson, Peter F. (1950). On referring. *Mind*, *59*(235), 320–344.

Suchanek, Fabian M., Kasneci, Gjergji, & Weikum, Gerhard (2007). Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on world wide web* (pp. 697–706).

Tiwari, Sanju, Mihindukulasooriya, Nandana, Osborne, Francesco, Kontokostas, Dimitris, D'Souza, Jennifer, Kejriwal, Mayank, et al. (2022). Preface for the international workshop on knowledge graph generation from text. In *CEUR workshop proceedings*: *vol. 3184*, CEUR.

Wu, Ledell, Petroni, Fabio, Josifoski, Martin, Riedel, Sebastian, & Zettlemoyer, Luke (2019). Scalable zero-shot entity linking with dense entity retrieval. arXiv preprint arXiv:1911.03814.

Yao, Liang, Mao, Chengsheng, & Luo, Yuan (2019). Kg-bert: Bert for knowledge graph completion. arXiv preprint arXiv:1909.03193.

Yao, Yiqun, Xu, Jiaming, Shi, Jing, & Xu, Bo (2018). Learning to activate logic rules for textual reasoning. *Neural Networks*, *106*, 42–49.

Yu, Dongran, Yang, Bo, Liu, Dayou, Wang, Hui, & Pan, Shirui (2023). A survey on neural-symbolic learning systems. *Neural Networks*.

Zhang, Jing, Chen, Bo, Zhang, Lingxi, Ke, Xirui, & Ding, Haipeng (2021). Neural, symbolic and neural-symbolic reasoning on knowledge graphs. *AI Open*, *2*, 14–35.

Zhang, Zhihao, Zuo, Yuan, Lin, Chenghua, & Wu, Junjie (2024). Language model as an annotator: Unsupervised context-aware quality phrase generation. *Knowledge-Based Systems*, *283*, Article 111175.

Zhu, Yuqi, Wang, Xiaohan, Chen, Jing, Qiao, Shuofei, Ou, Yixin, Yao, Yunzhi, et al. (2024). Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities. *World Wide Web*, *27*(5), 58.