Mining for Species, Locations, Habitats, and Ecosystems from Scientific Papers in Invasion Biology: A Large-Scale Exploratory Study with Large Language Models

Jennifer D'Souza[1], **Zachary Laubach**[2], **Tarek Al Mustafa**[3], **Sina Zarrieß**[4], **Robert Frühstückl**[5], **Phyllis Illari**[6]
[1]TIB Leibniz Information Centre for Science and Technology,
[2]University of Colorado Boulder, [3]Friedrich Schiller University Jena, [4,5]Bielefeld University,
[6]University College London
jennifer.dsouza@tib.eu

Abstract

This study explores the use of large language models (LLMs), specifically GPT-4o, to extract key ecological entities—species, locations, habitats, and ecosystems—from invasion biology literature. This information is critical for understanding species spread, predicting future invasions, and informing conservation efforts. Without domain-specific fine-tuning, we assess the potential and limitations of GPT-4o, out-of-the-box, for this task, highlighting the role of LLMs in advancing automated knowledge extraction for ecological research and management.

1 Introduction

Human population growth and expansion drive the intentional and unintentional movement of species beyond their historic ranges, leading to significant ecological impacts (Roy et al., 2023). Invasion biology seeks to understand these impacts across ecological scales to conserve native species and maintain functional ecosystems that provide essential services (Cassey et al., 2018; Jeschke and Heger, 2018). However, alien species introductions occur at an accelerating pace globally, making it increasingly difficult for researchers to systematically track and categorize species, their locations, and relationships. This paper explores the potential of recent NLP technologies, specifically Information Extraction (IE) approaches based on Large Language Models (LLMs) (Amatriain et al., 2023; Jennifer D'Souza, 2025), as tools for predicting future invasions and their consequences.

The extraction and categorization of information from scientific publications is a well-known NLP task (Augenstein et al., 2017; Gábor et al., 2018; Luan et al., 2018; Brack et al., 2020; Dessì et al., 2020; D'Souza et al., 2021; Liu et al., 2021; Kabongo et al., 2021; D'Souza and Auer, 2022; D'Souza, 2024; Shamsabadi et al., 2024; D'Souza et al., 2024). While Named Entity Recognition (NER) and Relation Extraction (RE) have been extensively applied in the biomedical domain for network biology (Zhou et al., 2014), gene prioritization (Aerts et al., 2006), drug repositioning (Wang and Zhang, 2013), and curated database creation (Li et al., 2015), their application in invasion biology remains underexplored. To the best of our knowledge, the small-scale INAS dataset (Brinner et al., 2022) is the only invasion biology-specific resource with annotated hypotheses for scientific abstracts.

This paper investigates information extraction (IE) in invasion biology, encompassing both named entity recognition (NER) and relation extraction (RE). We simultaneously build on studies showing that jointly learning NER and RE can enhance overall performance (Giorgi et al., 2019) and on recent LLMs which may open new opportunities for IE. Thus, our central question is whether LLMs, with their advanced pattern recognition capabilities, can be effectively applied to a new domain to simultaneously identify entities and infer their relationships. We prompt LLMs to extract four key entities—species, location, habitat, and ecosystem—and qualitatively evaluate results by addressing: (i) the relevance of extracted entities and interactions, (ii) the types of inferred relationships, and (iii) the benefits of LLM workflows for large-scale data mining. This work makes two key contributions: (i) the release of a text data mining corpus of over 10,000 invasion biology papers, including full text for nearly 2,000, with structured information extracted by GPT-4o (`https://doi.org/10.5281/zenodo.13956882`); and (ii) a systematic workflow for schema discovery in IE tasks, broadly applicable for leveraging LLMs in open-ended IE objectives.

Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology2025), pages 16–23

2 Our Text Data Mining Corpus

As a first step, we compiled a literature corpus as the unstructured source of scientific information. Starting with the Invasion Biology Corpus (Mietchen et al., 2024), which catalogs metadata for 49,438 papers in Wikidata. Using their DOIs, we queried the ORKG ASK search engine's API to retrieve abstracts and full texts, leveraging ASK's broad coverage of over 80 million papers (Knoth et al., 2023). Of the 49,438 queried papers, 12,636 were available in ASK—9,802 with abstracts only and 2,834 with both abstracts and full texts—highlighting the challenge of limited open-access availability. Bibliometric analysis of these abstracts shows papers spanning 52 years (since 1950), with full texts available from 1990 onward. A snapshot of the past 20 years (Figure 1) shows 2016 as the peak year for abstracts (1,183) and 2017 for full texts (294). Figure 2 presents the distribution across the top ten publishers, with further details in our online repository.

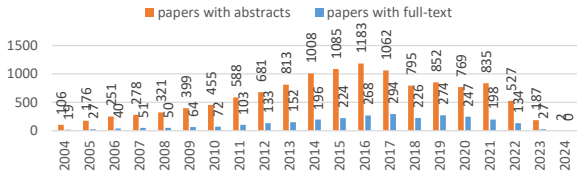

Figure 1: Distribution of papers in our corpus over the past 20 years.

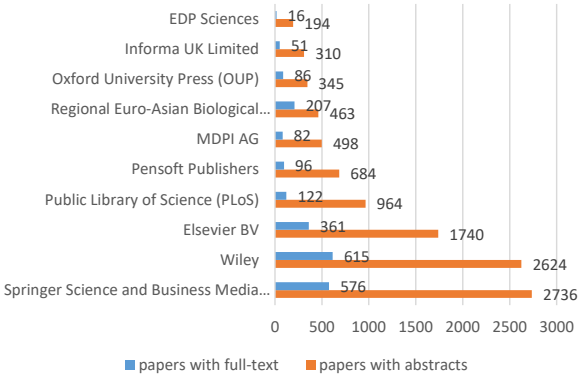

Figure 2: Distribution of papers in our corpus across the top ten publishers.

3 Information Extraction with LLMs

An IE task requires two prerequisites: 1) a collection of papers for processing, and 2) a schema defining the extraction targets.

3.1 Schema Discovery

Schema discovery is central to our approach, aiming to define a standardized semantic structure for IE from scientific papers. Without a predefined set of relations, our schema must flexibly capture extracted entities and their relationships. We achieve this in two stages: **specialize** and **generalize**. In the **specialize** stage, the LLM generates a schema for each paper in a given small sample, positing specialized extraction targets on four entities—*species*, *location*, *habitat*, and *ecosystem*. In the **generalize** stage, the LLM synthesizes a unified schema from multiple specialized schema instances, providing a flexible framework for relation extraction across all papers.

3.1.1 Stage Specialize: Schemas per Paper

The LLM operates in *completion* mode, guided by a SYSTEM PROMPT that defines its role as a "research assistant in invasion biology," tasked with extracting entity relationships. Initially, the prompt lacked precise entity definitions, but expert feedback led to refinements incorporating formal definitions, improving consistency (Table 1). The final system prompt aims to align the LLM for more accurate structured IE. The USER PROMPT then supplies each paper's title and abstract.

Results. Ten randomly selected papers were processed, with the resulting schemas available in our repository. Nine were true positives, while one was an outlier, indicating potential false positives in dataset filtering. Early schemas, such as Schema 1, employed basic entity categorization, whereas later schemas, like Schema 8, introduced more nuanced relationships by incorporating ecological and anthropogenic interactions. This evolution improved granularity and contextual relevance, capturing species dynamics within environmental conditions. Recurring patterns and study-specific distinctions emerged, with common themes—e.g., invasion biology, pollination networks, and anthropogenic impacts—highlighting research priorities. Standardized fields such as *species* and *location* ensured consistency, while tailored relationships, including "most effective pollinators" in Schema 2 and "competitive replacement" in Schema 5, provided contextual specificity. Integrating spatial and environmental parameters further reinforced the significance of habitats and ecosystems in ecological interactions.

Entity	Description
Species	Includes specific named species (e.g., Asterias amurensis) and broader categories (e.g., demersal fish, aquatic invertebrates), covering plants, animals, fungi, or microbes introduced to new environments where they establish, spread, and cause ecological or economic impacts. Higher-level taxonomic or functional groups are included when specific species are not identified, but generic terms like "invasive species" are excluded.
Location	Refers to study sites, from specific locations (e.g., "Port Phillip Bay, southern Australia") to broader regions (e.g., southern Australia, Amazon rainforest). Includes natural features (rivers, bays, mountains) and administrative areas (cities, states, countries).
Ecosystem	A system of interacting biological and abiotic components, often spanning multiple locations (e.g., the savannah ecosystem across Kenya and Tanzania).
Habitat	A specific part of an ecosystem where an organism lives, such as crocodiles in freshwater habitats (e.g., rivers) within the savannah ecosystem.

Table 1: Definitions of the four entities that encompass the information extraction (IE) aim of this paper.

3.1.2 Stage Generalize: Generic Schema

The goal of this stage was to develop a standardized schema in JSON format, capturing relationships among the four entities. The system prompt, similar to the specialize stage, defined the LLM's role as both a research assistant and an expert in semantic modeling. Inspired by prior schema discovery research (Baazizi et al., 2017, 2020), the LLM reviewed all individual schemas and proposed a unified structure. Since LLM outputs vary across runs, we prompted the model three times with: "Read the nine schema instances and generate a standardized schema in JSON format."

Results. The three generated JSON schema variants structured entity relationships with slight variations. Schema 1 emphasized geospatial precision, incorporating coordinates and linking habitats to ecosystems. Schema 2 detailed species roles (native, invasive) and introduced broader biological, physical, and anthropogenic interactions. Schema 3 focused on taxonomy, physiographic attributes, and habitat specificity. Despite minor differences, all schemas captured essential relations.

From these insights, we finalized a standardized schema, organizing data around species, locations, ecosystems, habitats, and relationships, each with structured properties tailored to ecological contexts. For instance, species include roles (e.g., invasive, native) and taxonomic classification, while locations integrate geopolitical and environmental details. Ecosystems and habitats are linked hierarchically, and relationships are classified by type (e.g., biological, ecological) and directionality. This schema enhances ecological network mapping, providing structured insights into species interactions across datasets. Table 2 presents a detailed breakdown.

3.2 Information Extraction

With a standardized semantic structure for extracting information from each paper, enabling easier downstream processing, the LLM-based IE task was conducted.

3.2.1 Stage Extract: Populate Schema

This stage now fulfills the main objective of this work, i.e. to extract information from a large-scale corpus (12,636 in our case) with an LLM to mine species, location, habitat, and ecosystem entities and their relations. The system prompt in this stage was close to the **specialize** stage system prompt where the role specified for the LLM was "research assistant in invasion biology or ecology tasked with reading and understanding scientific papers *to* extract relevant information *per the given predefined schema*."

3.3 Technical Details

The proprietary OpenAI GPT-4o model was used for all tasks in this paper. Schema generation in the **specialize** (Section 3.1.1) and **generalize** (Section 3.1.2) stages took only a few seconds per schema. The full extraction task in the **extract** stage (Section 3.2.1), applied to 12,636 papers, required approximately three days. The total cost was $1,000.

3.4 Results and Discussion

Of the 12,636 papers, the LLM classified 1,740 as outside invasion biology ("N/A"), leaving 10,896 for IE. This section summarizes the results.

Extraction Target	Extracted Item	Extracted Item Properties
Species	`name`: species_name	`role`: native/introduced/alien/invasive `taxonomy_level`: species/genus/family
Location	`name`: location_name	`category`: natural/administrative `geopolitical_info`: country/region/city `additional_details`: climatic/physiographic
Ecosystem	`name`: ecosystem_name	`type`: aquatic/terrestrial/marine `scope`: local/regional/global
Habitat	`name`: habitat_name	`type`: aquatic/terrestrial/marine `subcomponent_of`: ecosystem_name `specifics`: e.g., benthic, litoral
Relationships	`related_entities`: [entity1, entity2, ...]	`name`: relationship_name `type`: biological/physical/ecological/anthropogenic `directionality`: unidirectional/bidirectional `context`: relationship_contextual_description

Table 2: Standardized information extraction (IE) schema for four ecological entities, their relationships, and associated properties, pertinent to structure information from invasion biology scientific papers.

The extracted species roles reflect diverse ecological functions, origins, behaviors, and impacts in invasion biology. Broad categories include **native**, **alien**, **introduced**, **invasive**, and **naturalized**, alongside specific roles such as **agricultural weeds**, **biological control agents**, **pathogens**, **mutualists**, and **ecosystem engineers**. Some roles emphasize origins (**indigenous**, **non-native**, **cryptogenic**), behaviors (**colonizer**, **expanding**), or ecological functions (**symbiont**, **facilitator**, **pioneer**). Others capture ecosystem interactions (**co-introduced species**, **specialist herbivores**, **cryptic invaders**) or relate to conservation and management (**natural enemies**, **candidate biological control agents**, **quarantine pests**). This complexity underscores species' dynamic roles, informing biodiversity patterns, ecosystem impacts, and management strategies (full list here). A finer-grained analysis highlights **invasive species** as the most cited, including *Procambarus clarkii* (76 mentions), *Harmonia axyridis* (73), and *Rhinella marina* (68). **Native species** such as *Austropotamobius pallipes* and *Phragmites australis* (24 mentions each) appeared less frequently, while **introduced species** like *Oncorhynchus mykiss* and *Crassostrea gigas* showed varying ecological impacts. However, extraction also included generic terms (e.g., "native species," "native plants"), introducing noise due to the unsupervised nature of the task, highlighting the need for post-filtering (full list here).

The dataset highlights key **geopolitical locations**, with the most frequent countries being Australia (406), South Africa (248), New Zealand (236), Italy (187), and France (168). Regions include Europe (601), North America (348), the Mediterranean (117), Asia (112), and South America (98). Cities like Sydney (8), Hong Kong (7), and Rome (6) appear less frequently. The prominence of Europe and North America reflects their strong representation, while frequent mentions of Australia, South Africa, and New Zealand suggest a focus on biodiversity hotspots. The dataset spans continents, regions, countries, and cities, emphasizing a global perspective.

The extracted data provides a comprehensive view of **terrestrial, marine, and aquatic ecosystems**, highlighting their ecological diversity. Terrestrial ecosystems (93) dominate, with grasslands (42), forests (45), and agricultural landscapes (47) being the most cited. Mediterranean (37) and tropical ecosystems (26) reflect climate-specific regions, while urban ecosystems (46) underscore human-nature interactions. Marine ecosystems feature prominently, with the Mediterranean Sea (71) leading, followed by coral reefs (8) and the Baltic Sea (12). Aquatic ecosystems, especially freshwater systems (199), are well-represented, including lake (59), riverine (36), and wetland (40) ecosystems. Transitional zones such as estuarine (35) and coastal wetlands (10) further bridge freshwater and marine systems (full list here). Ad-

ditionally, the dataset captures **habitat-ecosystem relationships**, showcasing their ecological complexity. In aquatic systems, *pelagic zones* align with lake ecosystems, while *ballast water* links to marine environments. Marine habitats like *kelp beds* and *mussel beds* are associated with rocky subtidal and intertidal ecosystems, respectively. Human-modified environments, such as *artificial coastal defenses* linked to *biogenic reefs*, emphasize anthropogenic influences. Terrestrial systems highlight relationships like *forest habitats* in forest ecosystems, *soybean fields* in agricultural settings, and *urban areas* tied to urban ecosystems, underscoring the impact of land use. These insights illustrate the dataset's detailed representation of ecological interactions across environments.

The extracted information in our invasion biology corpus reveals diverse relation types, reflecting the field's interdisciplinary nature. **Ecological relations** dominate, with **invasion** (814), **competition** (429), **impact** (349), and **predation** (301) highlighting key species interactions and environmental changes. Other notable relations include **colonization** (179), **distribution** (179), and **habitat preference** (123), emphasizing species spread and habitat use. **Biological relations** such as **parasitism** (151), **hybridization** (74), and **pollination** (25) capture specific ecological interactions. **Physical relations** like **location**, **transport**, and **introduction location** focus on spatial and movement dynamics. **Anthropogenic relations**, including **introduction** (157) and **introduction pathway** (45), underscore the role of human activities in species dispersal. These relations collectively show the complexity of invasion biology.

The fully unsupervised IE task demonstrates the immense potential of LLMs as powerful tools for ecological research, assisting with tasks like systematic and scoping reviews. The insights presented here represent only a fraction of what can be derived from our corpus of over 10,000 papers, which we have made publicly available (`https://doi.org/10.5281/zenodo.13956882`). This work aligns with open information extraction (OIE) (Etzioni et al., 2008; Fader et al., 2011; Etzioni et al., 2011), traditionally reliant on syntactic patterns. However, LLMs surpass these methods by leveraging advanced semantic comprehension, enabling more effective analysis of complex relationships in large-scale corpora.

4 Recommendations for Future Work

Future work should explore integrating ontologies with LLMs to enhance information extraction (IE) and linked data creation, addressing key research questions: how LLMs can assist in ontology and knowledge graph construction (Kommineni et al., 2024), improve question answering through ontology support (Allemang and Sequeda, 2024), enable ontology learning from text (Babaei Giglou et al., 2023, 2024), and enhance representation learning (Ronzano and Nanavati, 2024). Ontologies, as *formal specifications of shared conceptualizations* (Studer et al., 1998), enable structured knowledge representation, yet their adoption is hindered by expertise barriers. Future research should investigate schema-driven IE, optimizing the information provided to LLMs, refining structured guidance (Caufield et al., 2024), and assessing how LLM-derived knowledge aligns with expert consensus. Ontologies can improve LLMs by supplying domain-specific definitions, guiding semantic modeling, enhancing entity and relation extraction, and integrating with retrieval-augmented generation (RAG) to reduce hallucinations (Soman et al., 2024). However, constraints must be considered in rapidly evolving fields, where rigid ontological structures may limit adaptability to emerging knowledge. Balancing structured knowledge integration with flexibility will be crucial for leveraging LLMs effectively across diverse domains.

5 Conclusion

This study highlights the potential of LLMs for advancing IE in invasion biology by extracting species, locations, habitats, and ecosystems from scientific literature. Through a standardized semantic schema, we demonstrated how LLMs can structure complex ecological data, enhancing research workflows. Our two-stage approach first extracts detailed, context-specific structures (specialize stage) and then integrates them into a flexible schema (generalize stage) balancing specificity and generality. This method enables structured representation of ecological complexity. The released dataset and schema support refining extraction methods, integrating ontologies, and broader ecological applications, underscoring LLMs' role in bridging unstructured data and structured knowledge in ecology.

Acknowledgments

The experiments reported in this paper using the OpenAI models were supported by the SCINEXT project grant (BMBF, German Federal Ministry of Education and Research, Grant ID: 01lS22070) and by the TIB Leibniz Information Centre for Science and Technology, Hannover, Germany. This work is also conducted in the context of the ZiF interdisciplinary research group "Mapping Evidence to Theory in Ecology," hosted at Bielefeld University.

Limitations

While this study highlights the potential of LLMs for IE in invasion biology, certain limitations remain. The extracted entities and relations were not evaluated against a gold-standard dataset, making it difficult to quantify precision and recall. A future inter-annotator agreement (IAA) study on a subset of the corpus (e.g., 20%) or a qualitative error analysis could enhance its reliability for researchers. Our approach also relies solely on OpenAI GPT-4o, without comparing alternative LLMs or prompting strategies, such as chain-of-thought prompting, which may improve extraction accuracy. Additionally, potential data contamination (Ranaldi et al., 2024) remains a concern, as LLMs may reproduce information seen during pre-training rather than extracting it anew. A systematic comparison against pre-training corpora would help assess this effect.

References

Stein Aerts, Diether Lambrechts, Sunit Maity, Peter Van Loo, Bert Coessens, Frederik De Smet, Leon-Charles Tranchevent, Bart De Moor, Peter Marynen, Bassem Hassan, et al. 2006. Gene prioritization through genomic data fusion. *Nature biotechnology*, 24(5):537–544.

Dean Allemang and Juan Sequeda. 2024. Increasing the llm accuracy for question answering: Ontologies to the rescue! *arXiv preprint arXiv:2405.11706*.

Xavier Amatriain, Ananth Sankar, Jie Bing, Praveen Kumar Bodigutla, Timothy J Hazen, and Michaeel Kazi. 2023. Transformer models: an introduction and catalog. *arXiv preprint arXiv:2302.07730*.

Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew Mccallum. 2017. Semeval 2017 task 10: Scienceie-extracting keyphrases and relations from scientific publications. In *Proceedings of the 11th SemEval (SemEval-2017)*, pages 546–555.

Mohamed-Amine Baazizi, Clément Berti, Dario Colazzo, Giorgio Ghelli, and Carlo Sartiani. 2020. Human-in-the-loop schema inference for massive json datasets. In *EDBT 2020-23nd International Conference on Extending Database Technology*, pages 635–638. OpenProceedings. org.

Mohamed-Amine Baazizi, Houssem Ben Lahmar, Dario Colazzo, Giorgio Ghelli, and Carlo Sartiani. 2017. Schema inference for massive json datasets. In *Extending Database Technology (EDBT)*.

Hamed Babaei Giglou, Jennifer D'Souza, and Sören Auer. 2023. Llms4ol: Large language models for ontology learning. In *International Semantic Web Conference*, pages 408–427. Springer.

Hamed Babaei Giglou, Jennifer D'Souza, and Sören Auer. 2024. Preface for llms4ol 2024: The 1st large language models for ontology learning challenge at the 23rd iswc. *Open Conference Proceedings*, 4:1–2.

Arthur Brack, Jennifer D'Souza, Anett Hoppe, Sören Auer, and Ralph Ewerth. 2020. Domain-independent extraction of scientific concepts from research articles. In *European Conference on Information Retrieval*, pages 251–266. Springer.

Marc Brinner, Tina Heger, and Sina Zarriess. 2022. Linking a hypothesis network from the domain of invasion biology to a corpus of scientific abstracts: The INAS dataset. In *Proceedings of the first Workshop on Information Extraction from Scientific Publications*, pages 32–42, Online. ACL.

Phillip Cassey, Pablo García-Díaz, Julie L Lockwood, and Tim M Blackburn. 2018. Invasion biology: searching for predictions and prevention, and avoiding lost causes. In *Invasion biology: hypotheses and evidence*, pages 3–13. CAB International Wallingford UK.

J Harry Caufield, Harshad Hegde, Vincent Emonet, Nomi L Harris, Marcin P Joachimiak, Nicolas Matentzoglu, HyeongSik Kim, Sierra Moxon, Justin T Reese, Melissa A Haendel, et al. 2024. Structured prompt interrogation and recursive extraction of semantics (spires): A method for populating knowledge bases using zero-shot learning. *Bioinformatics*, 40(3):btae104.

Danilo Dessì, Francesco Osborne, Diego Reforgiato Recupero, Davide Buscaldi, Enrico Motta, and Harald Sack. 2020. Ai-kg: an automatically generated knowledge graph of artificial intelligence. In *The Semantic Web–ISWC 2020: 19th International Semantic Web Conference, Athens, Greece, November 2–6, 2020, Proceedings, Part II 19*, pages 127–143. Springer.

Jennifer D'Souza, Sören Auer, and Ted Pedersen. 2021. SemEval-2021 task 11: NLPContributionGraph - structuring scholarly NLP contributions for a research knowledge graph. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 364–376, Online. Association for Computational Linguistics.

Jennifer D'Souza. 2024. Agriculture named entity recognition—towards fair, reusable scholarly contributions in agriculture. *Knowledge*, 4(1):1–26.

Jennifer D'Souza and Sören Auer. 2022. Computer science named entity recognition in the open research knowledge graph. In *International Conference on Asian Digital Libraries*, pages 35–45. Springer.

Jennifer D'Souza, Salomon Kabongo, Hamed Babaei Giglou, and Yue Zhang. 2024. Overview of the clef 2024 simpletext task 4: Sota? tracking the state-of-the-art in scholarly publications. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024)*, pages 3163–3173.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S Weld. 2008. Open information extraction from the web. *Communications of the ACM*, 51(12):68–74.

Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, et al. 2011. Open information extraction: The second generation. In *Twenty-Second International Joint Conference on Artificial Intelligence*. Citeseer.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 1535–1545.

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haïfa Zargayouna, and Thierry Charnois. 2018. SemEval-2018 task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 679–688, New Orleans, Louisiana. Association for Computational Linguistics.

John Giorgi, Xindi Wang, Nicola Sahar, Won Young Shin, Gary D Bader, and Bo Wang. 2019. End-to-end named entity recognition and relation extraction using pre-trained language models. *arXiv preprint arXiv:1912.13415*.

Jennifer D'Souza. 2025. A catalog of transformer models.

Jonathan M Jeschke and Tina Heger. 2018. *Invasion biology: hypotheses and evidence*. CAB International.

Salomon Kabongo, Jennifer D'Souza, and Sören Auer. 2021. Automated mining of leaderboards for empirical ai research. In *International Conference on Asian Digital Libraries*, pages 453–470.

Petr Knoth, Drahomira Herrmannova, Matteo Cancellieri, Lucas Anastasiou, Nancy Pontika, Samuel Pearce, Bikash Gyawali, and David Pride. 2023. Core: a global aggregation service for open access papers. *Scientific Data*, 10(1):366.

Vamsi Krishna Kommineni, Birgitta König-Ries, and Sheeba Samuel. 2024. From human experts to machines: An llm supported approach to ontology and knowledge graph construction. *arXiv preprint arXiv:2403.08345*.

Gang Li, Karen E Ross, Cecilia N Arighi, Yifan Peng, Cathy H Wu, and K Vijay-Shanker. 2015. mirtex: a text mining system for mirna-gene relation extraction. *PLoS computational biology*, 11(9):e1004391.

Haoyang Liu, M Janina Sarol, and Halil Kilicoglu. 2021. Uiuc_bionlp at semeval-2021 task 11: A cascade of neural models for structuring scholarly nlp contributions. *arXiv preprint arXiv:2105.05435*.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 EMNLP*, pages 3219–3232.

Daniel Mietchen, Jonathan M. Jeschke, Maud Bernard-Verdier, Tina Heger, Camille Musseau, and Steph Tyszka. 2024. Invasion biology corpus 2024-07.

Federico Ranaldi, Elena Sofia Ruzzetti, Dario Onorati, Leonardo Ranaldi, Cristina Giannone, Andrea Favalli, Raniero Romagnoli, and Fabio Massimo Zanzotto. 2024. Investigating the impact of data contamination of large language models in text-to-SQL translation. In *Findings of ACL*, pages 13909–13920, Bangkok, Thailand. ACL.

Francesco Ronzano and Jay Nanavati. 2024. Towards ontology-enhanced representation learning for large language models. *arXiv preprint arXiv:2405.20527*.

Helen E Roy, Aníbal Pauchard, Peter Stoett, Tanara Renard Truong, Sven Bacher, Bella S Galil, Philip E Hulme, Tohru Ikeda, Kavileveettil Sankaran, Melodie A McGeoch, et al. 2023. Ipbes invasive alien species assessment: summary for policymakers. *IPBES*.

Mahsa Shamsabadi, Jennifer D'Souza, and Sören Auer. 2024. Large language models for scientific information extraction: An empirical study for virology. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 374–392, St. Julian's, Malta. Association for Computational Linguistics.

Karthik Soman, Peter W Rose, John H Morris, Rabia E Akbas, Brett Smith, Braian Peetoom, Catalina Villouta-Reyes, Gabriel Cerono, Yongmei Shi, Angela Rizk-Jackson, et al. 2024. Biomedical knowledge graph-optimized prompt generation for large language models. *Bioinformatics*, 40(9):btae560.

Rudi Studer, V Richard Benjamins, and Dieter Fensel. 1998. Knowledge engineering: Principles and methods. *Data & knowledge engineering*, 25(1-2):161–197.

Zhong-Yi Wang and Hong-Yu Zhang. 2013. Rational drug repositioning by medical genetics. *Nature biotechnology*, 31(12):1080–1082.

XueZhong Zhou, Jörg Menche, Albert-László Barabási, and Amitabh Sharma. 2014. Human symptoms–disease network. *Nature communications*, 5(1):4212.