Analysis and Application of the Intelligence Task Ontology (ITO) in AI Benchmarking – a critical essay

Introduction

Blagec et al's manuscript describes the rationale, methodology, and outputs of the Intelligence Tasks Ontology (ITO), a tool designed to systematize knowledge around the development and benchmarking of Artificial Intelligence (AI) models (Blagec *et al.*, 2022). This essay provides a summary and critical evaluation of the paper, discussing both the strengths and weaknesses of their approach, as well as the broader impact and applications of ITO.

Manuscript summary

ITO was created in response to the rapid acceleration in AI research, exemplified by a four-fold increase in published manuscripts between 2000-2019 (Stanford Human-Centered Artificial Intelligence, 2025). This growth brought about a proliferation of tasks, datasets, methods, and benchmarking results, making it difficult for researchers to monitor progress across AI subfields. Without systematic mapping, important patterns can remain hidden, hampering efforts to prioritize research and resulting in resource waste due to poor allocation or duplication.

The main purpose of ITO is to provide a comprehensive framework for organizing information related to AI tasks, benchmarks, and performance metrics. Its structured approach allows for a global analysis of AI capabilities and supports meta-research into advances across AI research domains. ITO makes it possible to track where new AI methods succeed or fail, or compare progress across different types of tasks and domains, and offers a taxonomic foundation for annotating datasets and benchmarks. This organized structure helps developers identify suitable models for specific tasks and facilitates integration of AI with other fields, such as bioinformatics.

To construct ITO, the authors used a mix of automated data extraction and manual curation by experts. The primary data source was Papers With Code (PWC), the largest public repository of AI benchmarks and results at the time of publication (*Papers With Code*, no date), which combines automated extraction from arXiv records and manual input by the AI community. PWC data were first converted to established knowledge representation standards (RDF/OWL). AI specialists then curated these data, standardizing over 800 names for performance metrics and organizing AI processes (or tasks) into a process-centric hierarchy spanning 16 parent classes, such as Natural Language Processing. Importantly, this effort leveraged established ontologies for greater consistency and interoperability.

The resulting knowledge graph contained over 685,000 relationships, linking more than 50,000 entities across 9,000 classes, capturing more than 26,000 benchmark outcomes using 3,633 datasets, and covering work from 2000 through 2021. The authors demonstrated ITO's effectiveness in specific use cases, such as analyzing which performance metrics are most prevalent in different research domains, and tracking progress in 16 AI research areas (Blagec *et al.*, 2020; Ott *et al.*, 2022). Through its ontological approach, ITO directly addresses the challenge of summarizing and analyzing trends and capabilities within AI, paving the way for more deliberate research planning and a deeper understanding of the field's development.

Critical Evaluation of ITO's Approach and Ontological Framework

ITO's methodology combines automated data collection with detailed manual curation. This hybrid ontology development approach facilitates scalability while maintaining data integrity and clarity (Yun *et al.*, 2021). One of the key innovations is the rigorous standardization of

performance metrics. By establishing a single canonical hierarchy for over 800 metrics, ITO makes it possible to conduct meaningful cross-study comparisons that were previously hindered by inconsistent terminology in AI research. However, the ontology currently lacks an effective way to distinguish metrics that share names but differ in definitions or implementations across disciplines. A potential enhancement here could include an additional attribute specifying a commonly accepted definition for each metric. The extent of manual curation required may also present a barrier, especially for long-term sustainability, as shown by the lack of updates to ITO since its release ('OpenBioLink/ITO', 2025).

Building on PWC data gave ITO a strong, highly-relevant grounding, helping ensure that the resulting ontology was both complete and concise (Raad and Cruz, 2025). Using an established, well-annotated data source expedited automated extraction and made updates easier, while also providing a framework for further expert annotation. However, heavy reliance on a single source introduces potential bias that may skew the portrayal of the AI landscape primarily towards what is present and well-annotated in PWC. This includes potential issues stemming from common publication biases and limited coverage of grey literature or alternative streams of research.

The process-centric hierarchical structure of ITO, where AI tasks (termed "processes") are classified into 16 major categories and further subdivided, mirrors the way researchers conceptualize AI capabilities and makes the system intuitive. However, this structure complicates the modelling of cross-modal tasks that naturally span several categories. Managing these via multiple inheritance leads to a tangled hierarchy, a violation of ontological best practices, as highlighted by its high tangledness metric (Gruber, 1995). The introduction of a universal "Benchmarking" superclass aids in data querying but further increases the complexity of the hierarchy. Although such tangledness can be justified for meta-research use, the authors themselves acknowledge that splitting the class hierarchy into collateral branches could mitigate this problem, although such changes have yet to be implemented.

While ITO's representation of benchmarks effectively maps relations between datasets, tasks, and results, it omits critical contextual information, such as evaluation protocols, hardware specifications, or training conditions. These omissions limit users' ability to make fair comparisons across benchmark results, ultimately reducing the knowledge graph's value for analyzing the evolution of AI research.

Considering the ontology-based knowledge graph approach more broadly, ITO illustrates some clear advantages. The system is flexible and can accommodate the addition of new tasks, metrics, or relationships with only minor manual input. Its graph structure is ideal for capturing the complex, multidimensional relationships inherent in AI research, which would be hard to express in a traditional database. The use of W3C standards (RDF/OWL) provides compatibility with other semantic data sources and allows the use of automated reasoning tools. These standards also enable consistency checking and can identify implicit relationships between AI tasks, a critical function as the field continues to expand in complexity.

The ontology approach is not without challenges, however. The complexity of knowledge graphs can render them less accessible to researchers without semantic web experience. As ITO grows, query performance and reasoning may slow, though this can be addressed through indexing and optimization (Hogan *et al.*, 2021). More problematic is the difficulty in representing uncertainty or conflicting information, since benchmark results often involve caveats, are context-dependent, or may be contested. Yet ontologies typically lack mechanisms to encode such nuances (Gottlob *et al.*, 2012). Finally, ongoing curation is labor-intensive, raising sustainability

concerns without additional resources or institutional support, as shown by ITO's stagnation since publication.

These issues echo the broader divisions within AI: symbolic or knowledge-based AI, as exemplified by ITO, provides structure and explainability and works well for well-defined, discrete knowledge domains. Yet, it falls short in representing nuance and uncertainty and can be hard to scale. By contrast, connectionist approaches, such as deep learning, are highly adaptable, scalable, and well-suited to complex or ambiguous domains, though they are less interpretable, highly-dependent on training data quality, and prone to accuracy issues.

Real-world application and future implications

Despite these challenges, ITO's ontology-based approach is capable of effectively representing the interconnected, rapidly evolving world of AI, provided that its limitations are understood and augmented with other methods where needed. As such, ITO can support several practical applications in AI research practice. Research funders could use ITO to identify underexplored capabilities, making informed choices about which AI gaps to address through funding or where to promote promising methodologies. Researchers could utilize ITO as a discovery resource, finding relevant datasets, benchmarks, and top-performing models when moving into new fields, a particularly valuable function for interdisciplinary research. The system's standardized performance metrics could help the research community move toward more consistent and reproducible evaluation, addressing current issues in replicability and comparability (Pineau *et al.*, 2020). In the industrial context, ITO could enable more informed decision-making by offering a clear, comprehensive view of AI capabilities across tasks and domains. This is particularly useful in areas such as healthcare or autonomous vehicles, where the need to clearly understand AI limitations and mitigate associated risks are particularly high.

Perhaps most transformative however is ITO's potential to shape benchmarking standards. By offering a canonical mapping of tasks and metrics, ITO gives journals and conferences a reference for standardizing reporting rules, gradually bringing greater consistency to AI evaluation. Its underlying structure supports rapid, automated aggregation of benchmarking results, a vital feature as the rate of AI publications and models continues to rise. ITO also opens the door for collaborative, community-driven development. If maintained as a "living" ontology, research teams across the world could contribute benchmarking outcomes directly, sharing the curation load and ensuring the resource remains current and comprehensive.

Conclusions

In essence, ITO marks a potentially significant advance for the AI research community, providing a rigorous, ontology-driven system for organizing the field's complex landscape of tasks, benchmarks, and performance measures. The framework provides value for structuring, standardizing, and enabling meta-research, though its long-term impact will rely on sustained maintenance, broader integration of data sources, and active community participation. By addressing both the promise and the current practical limitations, this approach sets a valuable precedent for managing complexity in fast-moving research domains.

References:

Blagec, K. *et al.* (2020) *A critical analysis of metrics used for measuring progress in artificial intelligence*, *arXiv.org*. Available at: https://arxiv.org/abs/2008.02577v2 (Accessed: 13 June 2025).

Blagec, K. *et al.* (2022) 'A curated, ontology-based, large-scale knowledge graph of artificial intelligence tasks and benchmarks', *Scientific Data*, 9(1), p. 322. Available at: https://doi.org/10.1038/s41597-022-01435-x.

Gottlob, G. *et al.* (2012) 'Datalog and Its Extensions for Semantic Web Databases', in *Reasoning Web. Semantic Technologies for Advanced Query Answering*. *Reasoning Web International Summer School*, Springer, Berlin, Heidelberg, pp. 54–77. Available at: https://doi.org/10.1007/978-3-642-33158-9_2.

Gruber, T.R. (1995) 'Toward principles for the design of ontologies used for knowledge sharing?', *International Journal of Human-Computer Studies*, 43(5), pp. 907–928. Available at: https://doi.org/10.1006/ijhc.1995.1081.

Hogan, A. *et al.* (2021) 'Knowledge Graphs', *ACM Comput. Surv.*, 54(4), p. 71:1-71:37. Available at: https://doi.org/10.1145/3447772.

'OpenBioLink/ITO' (2025). OpenBioLink. Available at: https://github.com/OpenBioLink/ITO (Accessed: 13 June 2025).

Ott, S. *et al.* (2022) 'Mapping global dynamics of benchmark creation and saturation in artificial intelligence', *Nature Communications*, 13(1), p. 6793. Available at: https://doi.org/10.1038/s41467-022-34591-0.

Papers With Code (no date). Available at: https://paperswithcode.com/ (Accessed: 13 June 2025).

Pineau, J. *et al.* (2020) 'Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program)'. arXiv. Available at: https://doi.org/10.48550/arXiv.2003.12206.

Raad, J. and Cruz, C. (2025) 'A Survey on Ontology Evaluation Methods', in. *7th International Conference on Knowledge Engineering and Ontology Development*, pp. 179–186. Available at: https://www.scitepress.org/Link.aspx?doi=10.5220/0005591001790186 (Accessed: 13 June 2025).

Stanford Human-Centered Artificial Intelligence (2025) 'Artificial Intelligence Index Report 2025', *Artificial Intelligence* [Preprint].

Yun, W. *et al.* (2021) 'Knowledge modeling: A survey of processes and techniques', *International Journal of Intelligent Systems*, 36(4), pp. 1686–1720. Available at: https://doi.org/10.1002/int.22357.