REFERENCES

12. GPT-4o mini: advancing cost-efficient intelligence, 2024. URL `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/`.

14. Claude 3.5 Sonnet \ Anthropic, 2024. URL `https://www.anthropic.com/claude/sonnet`.

Sunitha Abburu and Suresh Babu Golla. Ontology and nlp support for building disaster knowledge base. In *2017 2nd International Conference on Communication and Electronics Systems (ICCES)*, pp. 98–103. IEEE, 2017.

Ankit Agrawal and Alok Choudhary. Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. *Apl Materials*, 4(5), 2016.

Suzi A Aleksander, James P. Balhoff, Seth Carbon, J. Michael Cherry, Harold J. Drabkin, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L. Harris, David P. Hill, Raymond Lee, Huaiyu Mi, Sierra Taylor Moxon, Chris J. Mungall, Anushya Muruganugan, Tremayne Mushayahama, Paul W. Sternberg, Paul D. Thomas, Kimberly M. Van Auken, Jolene Ramsey, Deborah A. Siegele, Rex L. Chisholm, Petra Fey, Maria Cristina Aspromonte, María Victoria Nugnes, Federica Quaglia, Silvio C. E. Tosatto, Michelle G. Giglio, Suvarna Nadendla, Giulia Antonazzo, Helen Attrill, Gilberto dos Santos, Steven J. Marygold, Victor B. Strelets, Christopher J. Tabone, Jim Thurmond, Pinglei Zhou, Saadullah H. Ahmed, Praoparn Asanitthong, Diana Luna Buitrago, Meltem N Erdol, Matthew C. Gage, Mohamed Ali Kadhum, Kan Yan Chloe Li, Miao Long, Aleksandra Michalak, Angeline Pesala, Armalya Pritazahra, Shirin C. C. Saverimuttu, Renzhi Su, Kate E. Thurlow, Ruth C. Lovering, Colin Logie, Snezhana Oliferenko, Judith A. Blake, Karen R. Christie, Lori E. Corbani, Mary E Dolan, Harold J. Drabkin, David P. Hill, Li Ni, Dmitry Sitnikov, Cynthia L. Smith, Alayne Cuzick, James Seager, Laurel D. Cooper, Justin Elser, Pankaj Jaiswal, Parul Gupta, Pankaj Jaiswal, Sushma Naithani, Manuel Lera-Ramírez, Kim M Rutherford, Valerie Wood, Jeffrey L De Pons, Melinda R. Dwinell, G. Thomas Hayman, Mary L. Kaldunski, Anne E. Kwitek, Stanley J. F. Laulederkind, Marek A Tutaj, Mahima Vedi, Shur-Jen Wang, Peter D'Eustachio, Lucila Aimo, Kristian B. Axelsen, Alan J Bridge, Nevila Hyka-Nouspikel, Anne Morgat, Suzi A Aleksander, J. Michael Cherry, Stacia R. Engel, Kalpana Karra, Stuart R. Miyasato, Robert S. Nash, Marek S. Skrzypek, Shuai Weng, Edith D. Wong, Erika Bakker, Tanya Z. Berardini, Leonore Reiser, Andrea H. Auchincloss, Kristian B. Axelsen, Ghislaine Argoud-Puy, Marie-Claude Blatter, Emmanuel Boutet, Lionel Breuza, Alan J Bridge, Cristina Casals-Casas, Elisabeth Coudert, Anne Estreicher, Maria Livia Famiglietti, Marc Feuermann, Arnaud Gos, Nadine Gruaz-Gumowski, Chantal Hulo, Nevila Hyka-Nouspikel, Florence Jungo, Philippe le Mercier, Damien Lieberherr, Patrick Masson, Anne Morgat, Ivo Pedruzzi, Lucille Pourcel, Sylvain Poux, Catherine Rivoire, Shyamala Sundaram, Alex Bateman, Emily H. Bowler-Barnett, Hema Bye-A-Jee, Paul Denny, Alex Ignatchenko, Rizwan Ishtiaq, Antonia Lock, Yvonne Lussi, Michele Magrane, Maria Jesus Martin, Sandra E. Orchard, Pedro Raposo, Elena Speretta, Nidhi Tyagi, Kate Warner, Rossana Zaru, Alexander D. Diehl, Raymond Lee, Juancarlos Chan, Stavros Diamantakis, Daniela Raciti, Magdalena Zarowiecki, Malcolm E. Fisher, Christina James-Zorn, Virgilio G. Ponferrada, Aaron M. Zorn, Sridhar Ramachandran, Leyla Ruzicka, and Monte Westerfield. The gene ontology knowledgebase in 2023. *Genetics*, 224, 2023. URL `https://api.semanticscholar.org/CorpusID:257311316`.

2023 Anthropic, 2023. Anthropic. (2023). Claude (Oct 8 version) [Large language model]. https://www.anthropic.com/.

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

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. Seven failure points when engineering a retrieval augmented generation system. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI*, pp. 194–199, 2024.

Alexander S. Behr, Marc Völkenrath, and Norbert Kockmann. Ontology extension with nlp-based concept extraction for domain experts in catalytic sciences. *Knowledge and Information Systems*, 65:5503–5522, 2023. URL https://api.semanticscholar.org/CorpusID:259929231.

Subhra Bikash Bhattacharyya. *Introduction to SNOMED CT*. Springer, 2015.

Manish Bhattarai, Javier E Santos, Shawn Jones, Ayan Biswas, Boian Alexandrov, and Daniel O'Malley. Enhancing code translation in language models with few-shot learning via retrieval-augmented generation. *arXiv preprint arXiv:2407.19619*, 2024.

Abeba Birhane, Atoosa Kasirzadeh, David Leslie, and Sandra Wachter. Science in the age of large language models. *Nature Reviews Physics*, 5(5):277–280, 2023.

Chris Bizon, Steven Cox, James Balhoff, Yaphet Kebede, Patrick Wang, Kenneth Morton, Karamarie Fecho, and Alexander Tropsha. Robokop kg and kgb: integrated knowledge graphs from federated sources. *Journal of chemical information and modeling*, 59(12):4968–4973, 2019.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023a.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*, 2023b.

Giovanni Ciatto, Andrea Agiollo, Matteo Magnini, and Andrea Omicini. Large language models as oracles for instantiating ontologies with domain-specific knowledge. *arXiv preprint arXiv:2404.04108*, 2024.

Roberto Confalonieri and Giancarlo Guizzardi. On the multiple roles of ontologies in explanations for neuro-symbolic ai. *Neurosymbolic Artificial Intelligence*, (Preprint):1–15, 2024.

Gene Ontology Consortium. The gene ontology resource: 20 years and still going strong. *Nucleic acids research*, 47(D1):D330–D338, 2019.

Maricela Claudia Bravo Contreras, Luis Fernando Hoyos Reyes, and José Alejandro Reyes Ortiz. Methodology for ontology design and construction. *Contaduría y Administración*, 2019. URL https://api.semanticscholar.org/CorpusID:159049319.

Lorraine J. Daston. The history of science and the history of knowledge. *KNOW: A Journal on the Formation of Knowledge*, 1:131 – 154, 2017. URL https://api.semanticscholar.org/CorpusID:164680540.

Kirill Degtyarenko, Paula De Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. Chebi: a database and ontology for chemical entities of biological interest. *Nucleic acids research*, 36(suppl_1):D344–D350, 2007.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

Jiaojiao Fang, Qingqing Chen, Zhi Li, Junjie Mao, and Yadong Li. The synthesis of single-atom catalysts for heterogeneous catalysis. *Chemical communications*, 2023. URL https://api.semanticscholar.org/CorpusID:256146078.

Maurice Funk, Simon Hosemann, Jean Christoph Jung, and Carsten Lutz. Towards ontology construction with language models. *arXiv preprint arXiv:2309.09898*, 2023.

Bartosz A. Grzybowski, Tomasz Badowski, Karol Molga, and Sara Szymkuć. Network search algorithms and scoring functions for advanced-level computerized synthesis planning. *WIREs Comput. Mol. Sci.*, 13(1):e1630, 2023. doi: https://doi.org/10.1002/wcms.1630.

Nicola Guarino, Daniel Oberle, and Steffen Staab. *What Is an Ontology?*, pp. 1–17. 05 2009. doi: 10.1007/978-3-540-92673-3_0.

Chi Han, Qifan Wang, Hao Peng, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. Lm-infinite: Zero-shot extreme length generalization for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3991–4008, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Tom Hope, Jason Portenoy, Kishore Vasan, Jonathan Borchardt, Eric Horvitz, Daniel S Weld, Marti A Hearst, and Jevin West. Scisight: Combining faceted navigation and research group detection for covid-19 exploratory scientific search. *arXiv preprint arXiv:2005.12668*, 2020.

Bairu Hou, Jinghan Jia, Yihua Zhang, Guanhua Zhang, Yang Zhang, Sijia Liu, and Shiyu Chang. Textgrad: Advancing robustness evaluation in nlp by gradient-driven optimization. *ArXiv*, abs/2212.09254, 2022. URL https://api.semanticscholar.org/CorpusID:254854553.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*, 2023.

Rizwan Iqbal, Masrah Azrifah Azmi Murad, Aida Mustapha, Nurfadhlina Mohd Sharef, et al. An analysis of ontology engineering methodologies: A literature review. *Research journal of applied sciences, engineering and technology*, 6(16):2993–3000, 2013.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 874–880, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.74. URL https://aclanthology.org/2021.eacl-main.74.

Kevin Maik Jablonka, Philippe Schwaller, Andres Ortega-Guerrero, and Berend Smit. Leveraging large language models for predictive chemistry. *Nat. Mach. Intell.*, 6(2):161–169, February 2024. ISSN 2522-5839. doi: 10.1038/s42256-023-00788-1.

Aashish Jain and Daisuke Kihara. Nntox: gene ontology-based protein toxicity prediction using neural network. *Scientific reports*, 9(1):17923, 2019.

C. Maria Keet. An introduction to ontology engineering. In *An Introduction to Ontology Engineering*, 2018.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, et al. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*, 2023.

Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G Rodriques, and Andrew D White. Paperqa: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*, 2023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474, 2020.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL https://arxiv.org/abs/2005.11401.

Xinzhe Li. A review of prominent paradigms for llm-based agents: Tool use (including rag), planning, and feedback learning. 2024. URL https://api.semanticscholar.org/CorpusID:270370985.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Hao Liu, Matei Zaharia, and Pieter Abbeel. Ring attention with blockwise transformers for near-infinite context. *arXiv preprint arXiv:2310.01889*, 2023.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024a.

Shiyu Liu, Yihong Chen, Chuchu Chen, Yaohua Wu, Juanshan Du, Xiaochi Feng, Qinglian Wu, Peishi Qi, Huazhe Wang, Nanqi Ren, and Wan-Qian Guo. From single-atom catalysis to dual-atom catalysis: A comprehensive review of their application in advanced oxidation processes. *Separation and Purification Technology*, 2024b. URL https://api.semanticscholar.org/CorpusID:269992019.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics*, 9: 329–345, 2021.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

OpenAI. Gpt-4 technical report. *Preprint at https://arxiv.org/abs/2303.08774*, 2023.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pp. 248–260. PMLR, 2022.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*, 2023.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. Kilt: a benchmark for knowledge intensive language tasks. *arXiv preprint arXiv:2009.02252*, 2020.

Robert G Raskin and Michael J Pan. Knowledge representation in the semantic web for earth and environmental terminology (sweet). *Computers & geosciences*, 31(9):1119–1125, 2005.

Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

Philippe Schwaller, Daniel Probst, Alain C Vaucher, Vishnu H Nair, David Kreutter, Teodoro Laino, and Jean-Louis Reymond. Mapping the space of chemical reactions using attention-based neural networks. *Nature machine intelligence*, 3(2):144–152, 2021.

ES Shahul, Jithin James, Luis Espinosa Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation. In *Conference of the European Chapter of the Association for Computational Linguistics*, 2023. URL `https://api.semanticscholar.org/CorpusID:263152733`.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*, 2021.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

David Soong, Sriram Sridhar, Han Si, Jan-Samuel Wagner, Ana Caroline Costa Sá, Christina Y Yu, Kubra Karagoz, Meijian Guan, Sanyam Kumar, Hisham Hamadeh, et al. Improving accuracy of gpt-3/4 results on biomedical data using a retrieval-augmented language model. *PLOS Digital Health*, 3(8):e0000568, 2024.

Manu Suvarna, Alain C. Vaucher, Sharon Mitchell, Teodoro Laino, and Javier Pérez-Ramírez. Language models and protocol standardization guidelines for accelerating synthesis planning in heterogeneous catalysis. *Nature Communications*, 14, 2023. URL `https://api.semanticscholar.org/CorpusID:265550759`.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 2024.

Sabrina Toro, Anna V Anagnostopoulos, Sue Bello, Kai Blumberg, Rhiannon Cameron, Leigh Carmody, Alexander D Diehl, Damion Dooley, William Duncan, Petra Fey, et al. Dynamic retrieval augmented generation of ontologies using artificial intelligence (dragon-ai). *arXiv preprint arXiv:2312.10904*, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Amalie Trewartha, Nicholas Walker, Haoyan Huo, Sanghoon Lee, Kevin Cruse, John Dagdelen, Alexander Dunn, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. *Patterns*, 3 (4), 2022.

David Vallet, Miriam Fernández, and Pablo Castells. An ontology-based information retrieval model. In *The Semantic Web: Research and Applications: Second European Semantic Web Conference, ESWC 2005, Heraklion, Crete, Greece, May 29–June 1, 2005. Proceedings 2*, pp. 455–470. Springer, 2005.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Adv. Neural Inf. Process. Syst.*, 30, 2017.

Karin Verspoor. 'fighting fire with fire'—using llms to combat llm hallucinations, 2024.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. Scifact-open: Towards open-domain scientific claim verification. *arXiv preprint arXiv:2210.13777*, 2022.

Aiqin Wang, Jun Li, and Zhang Tao. Heterogeneous single-atom catalysis. *Nature Reviews Chemistry*, 2:65–81, 2018. URL https://api.semanticscholar.org/CorpusID:139163163.

Chengrui Wang, Qingqing Long, Meng Xiao, Xunxin Cai, Chengjun Wu, Zhen Meng, Xuezhi Wang, and Yuanchun Zhou. Biorag: A rag-llm framework for biological question reasoning. *ArXiv*, abs/2408.01107, 2024. URL https://api.semanticscholar.org/CorpusID:271693700.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

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

Wiley-API. Text and Data Mining - Librarians, 2024. URL https://onlinelibrary.wiley.com/library-info/resources/text-and-datamining.

Kevin Wu, Eric Wu, and James Zou. How faithful are rag models? quantifying the tug-of-war between rag and llms' internal prior. *arXiv preprint arXiv:2404.10198*, 2024.

Wenhan Xiong, Xiang Lorraine Li, Srini Iyer, Jingfei Du, Patrick Lewis, William Yang Wang, Yashar Mehdad, Wen-tau Yih, Sebastian Riedel, Douwe Kiela, et al. Answering complex open-domain questions with multi-hop dense retrieval. *arXiv preprint arXiv:2009.12756*, 2020.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.*, 59(8):3370–3388, 2019.

Li Yin. AdalFlow: The Library for Large Language Model (LLM) Applications, 7 2024. URL https://github.com/SylphAI-Inc/AdalFlow.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Xiao Zhang, Ji Wu, Zhiyang He, Xien Liu, and Ying Su. Medical exam question answering with large-scale reading comprehension. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, and Bin Cui. Retrieval-augmented generation for ai-generated content: A survey. *ArXiv*, abs/2402.19473, 2024. URL https://api.semanticscholar.org/CorpusID:268091298.

A Appendix

A.1 Control experiments on biomedical benchmarks

A.1.1 Performance and analysis

Method	TM	MedMCQA	MedQA	MMLU-Med
ZeroShot	✗	62.06	67.16	80.06
CoT	✗	60.91	**69.99**	76.70
OntoRAG-simple	✗	**64.12**	68.34	79.26
	✓	61.80	68.11	80.01
OntoRAG-HyA	✗	**64.04**	67.64	79.96
	✓	62.13	**69.36**	**80.65**

Table 1: Performance comparison of methods on 3 biomedical benchmarks. TM denotes "*translation module*", refering to a variation of the fusion operator $\mathcal{F}$ in which an LLM translates ontological context into natural language.

A.1.2 Effects of ontological relevance.

We hypothesize that weak performance in some areas when using OntoRAG might be due to vocabulary discrepancies as an effect of decreased ontological relevance. The assess this, we conduct an analysis where for each question in a given benchmark, the number of retrieved concepts from an ontology is computed, and the mean across the benchmark is correlated to performance (accuracy), for a given method. That is, each ontorag variation contributes one point to the correlation analysis. The goal is to determine whether high ontological relevance correlates with higher accuracy.

The results in Table 2 indicate an overall positive and usually strong correlation between ontological relevance and downstream performance.

Benchmark	Correlation
MedQA	0.7852
MMLU-Med	0.7506
MedMCQA	0.1018

Table 2: Correlation values for different benchmarks

A.2 Medical Ontologies

We first evaluate our methodology by first gauging its performance on a well known LLM question and answer (QA) benchmark, Multi-Subject Multi-Choice Dataset for Medical domain (MedMCQA) (Pal et al., 2022). This is a popular benchmark for evaluating LLM performance on multiple choice questions from various areas in the medical domain. Questions from this dataset were first divided based on their medical domain (dentistry, pediatrics, etc.) which then guided the selection of ontologies to place into the OntoRAG pipeline. The selected ontologies were limited to a biochemical ontology (https://bioportal.bioontology.org/ontologies/REX) a general medical term/ diagnostic ontology (https://bioportal.bioontology.org/ontologies/SNOMEDCT), and the widely-used gene ontology (GO) Aleksander et al. (2023) in an attempt to cover most of the concepts present in the QA dataset. These ontologies were also chosen due to their public availability and their professional quality. The benchmark was was curated to only include concepts that appear within the utilized ontologies. The final dataset contained around 4000 questions with the number of questions ranging from 27 to 400 for each medical domain. As with the results presented in the main document, the OntoRAG system offers similar or improved performance over the baseline zero-shot and CoT methods, with a significant improvements in the areas of genetics, anatomy, and microbiology. These improvements correlate with the fact that we used ontologies most relevant to these fields.

No. Entries	Question Class	ZeroShot	CoT	OntoRAG
405	Unknown	**0.83**	0.78	0.82
311	Biochemistry	0.81	0.78	**0.83**
283	Physiology	**0.82**	0.79	**0.82**
130	Medicine	**0.88**	0.83	0.86
92	Preventive Medicine	**0.75**	0.65	0.71
88	Microbiology	0.58	0.57	**0.61**
80	Gynaecology & Obstetrics	**0.82**	0.78	**0.82**
77	Anatomy	0.77	0.77	**0.91**
72	Pharmacology	0.78	**0.79**	0.76
68	Pediatrics	0.85	**0.87**	0.85
49	Psychiatry	0.73	**0.76**	0.73
33	Surgery	**0.73**	0.67	0.61
23	Dental	**0.74**	0.65	**0.74**
18	Genetics	0.83	0.78	**0.89**
18	Orthopaedics	**0.83**	0.67	**0.83**
16	Neurology	**0.88**	0.81	0.81

Table 3: Accuracy of OntoRAG against baselines on MMLU-Med, by question class. The table shows the accuracy of each method by type of question. OntoRAG-HyA-TM was used here.

A.3 ONTORAG DETAILS

OntoRAG is implemented using the DSPy library Khattab et al. (2023). The library abstracts the interface with an LLM into Signatures and Modules. The Signatures abstract the prompting of the LLM into classes with Input and Output properties, while the Modules define the flow of information that the pipeline implements.

The below Module is defined as the OntoRAG base module, and defines some standard routines used in every other sub-module used in this work.

Figure 5: OntoRAG implementations used in this work. Only *Simple* and *HyQ* are shown here. These represent variations in the retrieval type (i.e. direct or hypothetical answer). Variations in the fusion operator F are defined as part of the BaseOntoRAG class, see Appendix A.3.

```python
class ORAG_Simple(BaseOntoRAG):
    """Simple Ontorag"""
    def forward(self, q: str):
        ctxt = self.retr(q)
        answer = self.predictor(
            question=q,
            context=context
        )
        return answer
```

OntoRAG Simple

```python
class ORAG_HyA(BaseOntoRAG):
    """Ontorag with Hypot. answer
    ↪ """
    def forward(self, q: str):
        # Hypothetical answer
        ctxt0 = self.retr(q)
        hans = self.hya(
            question=q,
            context=ctxt0
        )
        # Query concepts in HyA
        ctxt1 = self.retr(
            hans.answer
        )
        answer = self.predictor(
            question=q,
            context=ctxt1
        )
        return answer
```

OntoRAG-HyA

Algorithm 1 OntoRAG base class.

```
class BaseOntoRAG(dspy.Module):
    retriever: dspy.Retrieve
    ontoretriever: OntoRetriever

    def forward(self, query: str) -> dspy.Prediction:
        """Forward pass of the OntoRAG pipeline."""
        pass

    def retrieve(self, query: str, ctxt_doc: str|None) -> str:
        """Retrieve and format."""
        ctxt_doc, ctxt_onto = "", ""

        if ctxt_doc is None:
            ctxt_dict = self.retrieve_doc(query)
            ctxt_doc = self.format_context(ctxt_dict)

        if self.ontoretriever.ontology.ontologies:
            ctxt_ontoj = self.ontoretriever(query)
            ctxt_onto = self.format_onto_context(ctxt_ontoj)

        ctxt = self.fuse_contexts(ctxt_doc, ctxt_onto)
        return ctxt

    def format_context(self, context: List[Dict]) -> str:
        """Format context."""
        contexts = [p["text"] for c in context for p in c["passages"]]
        return "\n".join(deduplicate(contexts))

    def format_onto_context(self, context: List[Dict]) -> str:
        """Format ontology context."""
        return json.dumps(context, indent=2)

    def fuse_contexts(self, ctxt_doc: str, ctxt_onto: str) -> str:
        """Fuse document and ontology contexts."""
        return ctxt_doc + ctxt_onto
```

A specific implementation of OntoRAG looks as follows: First, a Signature is defined, where inputs and outputs are defined.

The Modules are written to handle the inputs in the Signature, and to produce the outputs.

A.3.1 Ontology Retrieval Operator

The operator $\mathcal{O}$ defined in eq. 2, works by first extracting concepts from a statement s and returning the most similar ontological concepts $\{o\}$ in the ontology. The concepts are retrieved by 1. extracting concepts from the input query, and 2. retrieving ontological context from each of those concepts. The complete ontology retrieval pipeline is illustrated in pseudo-code 4.

In our implementation, retrieval works by extracting concepts using the spacy "en_core_web_sm" parser. The pipeline then searches in the loaded ontology, and if found retrieves the parents, children, as well as the definition, if any.

Algorithm 2 MedQnA: Medical Question Answering Signature

```
class MedQnA(dspy.Signature):
    """Answer a question with a detailed response based on the
    given context. If the context is not relevant or there is no
    context, answer based on
    your knowledge."""

    context: str = dspy.InputField(
        desc="Context: This information shows the relationships between
        relevant concepts:"
    )
    question: str = dspy.InputField(
        desc="Here is the question you need to answer:"
    )
    reasoning: str = dspy.OutputField(
        desc="Reasoning: Let’s think step by step in order to ${reasoning}"
    )
    choice_answer: str = dspy.OutputField(desc="Answer: ${answer}")
```

Algorithm 3 SimpleORAG: Simple Ontology-enhanced Retrieval-Augmented Generation

```
class SimpleORAG(BaseOntoRAG):

    def __init__(
        self,
        ontology: Union[str, OntoRetriever],
        context: None|str,
    ):
        super().__init__()
        self.predictor = dspy.Predict(MedQnA)
        if isinstance(ontology, str):
            self.ontoretriever = OntoRetriever(ontology_path=ontology)
        else:
            self.ontoretriever = ontology

    def forward(self, qprompt: str) -> dspy.Prediction:
        context = self.retrieve(qprompt)
        answer = self.predictor(question=qprompt, context=context)
        return answer
```

A.3.2 Working example of OntoRAG.

Here we need to show an example of a variation of ontorag.

A.4 OntoGen Details

A.4.1 Self consistency

The improvement of LLMs' capabilities to generate high-quality, hallucination-free answers is currently a highly active area of research. Many generic methods have been proposed that improve LLMs outputs without training data, fine-tuning or reinforcement learning, which includes, among others, self-consistency Wang et al. (2022), debating LLMs Du et al. (2023), and self-refinement Madaan et al. (2024). Research by Huang et al. Huang et al. (2023) demonstrates that self-consistency offers competitive results while being more computationally efficient compared to other methods. Therefore, in this work, self-consistency is used to improve the quality of answers from a LLM. As utilized in our approach, self-consistency can be defined as:

Algorithm 4 Retrieval of ontological context

procedure PROCESSQUERY(query)
  $recognizedConcepts \leftarrow$ `RecognizeConcepts`$(query)$
  $output \leftarrow \emptyset$
  for each $ontology, concepts$ in $recognizedConcepts$ **do**
    for each $concept$ in $concepts$ **do**
      $context \leftarrow$ `GetOntologicalContext`$(concept, ontology)$
      $output[ontology][concept] \leftarrow context$
  return $output$
procedure RECOGNIZECONCEPTS(text)
  $doc \leftarrow$ `NLP`$(text)$
  $recognizedConcepts \leftarrow \emptyset$
  for each $token$ in doc **do**
    if $token$ matches any ontology pattern **then**
      $concept \leftarrow token.text$
      $ontology \leftarrow$ `DetermineOntology`$(concept)$
      $recognizedConcepts[ontology].add(concept)$
  return $recognizedConcepts$
procedure GETONTOLOGICALCONTEXT(concept, ontology)
  $class \leftarrow ontology.search(label = concept)$
  $context \leftarrow \{$
    $"label"$: $class.label,$
    $"definition"$: $class.definition,$
    $"parents"$: $class.superclasses(),$
    $"children"$: $class.subclasses()$
  $\}$ **return** $context$

Definition A.1 *Let $a_1, a_2, ..., a_n \in \mathbb{A}$ be the answers to a given prompt p generated by a LLM, and r_i the set of tokens generated before the answer a_i.*

Self-Consistency *(SC) applies a marginalization over r_i by taking the majority vote of the answers a_i, i.e. $a = \arg\max_{a_i} \sum_{j=1}^{n} \mathbb{1}(a_i = a_j)$, thus giving as a final answer the most "consistent" answer generated by the LLM.*

It is important to note that self-consistency was initially proposed to enhance Chain of Thought (CoT) reasoning Wei et al. (2022) in LLMs Wang et al. (2022), to improve performance on generalized problem-solving tasks. In our work, we leverage the generalizability of self-consistency to improve the quality of our knowledge schemas reconstruction.

A.4.2 VOCABULARY EXTRACTION

After each iteration with the LLM, when it has extracted a list of concepts, a verification step is performed that consists of performing a string search of each of the list terms, in the original sentence. Terms pass this filter only if they are contained in the original sentence. With this process, we terms that originate as a result of hallucinations from the LLM used.

A.4.3 CATEGORIES GENERATION

During the *refinement* step, the LLM is prompted to curate a list of the most frequent categories extracted from the previous step. SC is applied here by generating many answers from the same prompt, and taking the majority vote of the categories extracted. While this provides a more robust list of categories, it is important to note that the correctness of an ontology is dependent on the downstream application it is intended for. Therefore, human involvement may be required in this step to select or exclude certain categories in order to align it with the downstream application. The final list of categories is then used as a seed for extracting the entire taxonomy, making it crucial to ensure the list is of high quality.

In the case of SACs ontology, the generated list of categories, obtained by majority voting was: *Characterization, Physical properties, Synthesis methods, Reaction mechanisms, Structure, Applications, Reactions* and *Support*. The manual curation performed in this step involved selecting the following additional categories from the pool of generated categories, so as to make the ontology more aligned with our chemistry knowledge: *Catalytic performance, Preparation methods, Theory and modelling,* and *Materials*.

A.4.4 Algorithm for Taxonomy Generation

Algorithm 5 Iterative and Incremental Top-Down Taxonomy Generation

Input: Papers $\mathcal{P}$, Vocabulary $\mathcal{V}$, Initial Taxonomy $\mathcal{T}^{(0)}$
Output: Reconstructed Taxonomy after K iterations $\mathcal{T}^{(K)}$

for $k = 1, \ldots, K$ **do**
  $\mathcal{T}^{(k)} \leftarrow \mathcal{T}^{(k-1)}$
  for $P_i \in \mathcal{P}$ **do**
    $R_i \leftarrow$ query_relationships$(P_i, V_i, \mathcal{T}^{(k)})$
    for $(s,t) \in R_i$ **do**
      if is_valid$((s,t), \mathcal{T}^{(k)})$ **then**
        $\mathcal{T}^{(k)} \leftarrow \mathcal{T}^{(k)} \cup \{(s,t)\}$
return $\mathcal{T}^{(K)}$

Where,

- `query_relationships`: Extracts *isA* relationships (s,t) from paper P_i, where $s \in \mathcal{C}(\mathcal{T}^{(k)})$ is a term in the current taxonomy $\mathcal{T}^{(k)}$, and $t \in V_i$. This function aims to place each term into the existing taxonomy, potentially returning multiple relationships per term.
- `is_valid`: Ensures no loops are created in the taxonomy when inserting a new relationship.

In our implementation, `query_relationships` utilizes an LLM prompted with the paper content, the current taxonomy terms, and the vocabulary to be queried. An example prompt and response can be found in Appendix A.6. To enhance the quality of the generated taxonomy and reduce hallucinations, SC is applied in this step by generating multiple answers from the same prompt and taking the majority voting as the final answer.

A.4.5 Expert evaluation

In order to evaluate the quality of the generated ontology, a panel of two experts was assembled to assess the taxonomical relationships. The experts were tasked with randomly sampling relationships from various iterations of the ontology and determining whether each sampled relationship was correct according to the context provided for such relationship, in this case, the corresponding paper. According to the experts, on average at least 64.5% of the sampled relationships were considered correct. While this indicates a majority of accurate relationships, it also suggests room for improvement in the ontology generation process. Upon analysis of the incorrect relationships, the experts identified as potential improvements the removal of semantically similar concepts, which might appear repeated in different parts of the structure, and the need to provide a more specific context for the relationships, in order to reduce ambiguity.

A.4.6 SACs ontology example

To provide a concrete example of how the ontology is able to capture meaningful relationships, below two examples are provided corresponding to the *synthesis methods* (left) and *CO2 reduction reactions* (right) branches for both the ontologies generated with Claude 3.5 Sonnet and Llama3.1:70b. Here it can be seen that both ontologies are able to capture meaningful synthesis methods for SACs that appear in the literature. It can be seen that, generally there is an agreement in the synthesis methods identified in both ontologies. It can be highlighted, however, that the Llama-generated ontology contains a larger number of false-positive synthesis methods (e.g. *Methodology, Synthesis, Strategies*), which explains the larger number of terms included in this ontology. Regarding the *CO2 reduction* branch, one can notice that each ontology contains semantically similar terms (e.g. *Carbon dioxide reduction reaction* and *CO2 reduction reaction*). While this does not affect the downstream performance of the ontology, it creates unnecessary redundancies in the structure. Additionally, it can be seen that, in the Llama-generated ontology, *CO2 reduction* has not been classified as a separate branch, but instead, it is contained inside the *Reactions* branch, without this being necessarily incorrect. Finally, as it happened with the *synthesis methods* branch, the Llama-generated ontology contains evident false-positives (e.g. *CO2 molecules, dioxide*), which did not appear in the Claude-generated ontology.

Example SACs Ontology (Claude 3.5 Sonnet)

```
Thing
└── Synthesis methods
    ├── Catalyst synthetic strategies
    ├── Two-step approach
    ├── Ni-TAPc anchoring strategies
    ├── Pyrolysis procedure
    ├── Bimodal template based synthesis strategies
    ├── Multistep pyrolysis process
    ├── Multistep pyrolysis method
    ├── Wet chemistry methods
    ├── Pyrolysis
    ├── Atomic layer deposition
    ├── Pyrolysis process
    ├── NH3 atmosphere annealing
    ├── Co precipitation
    ├── Annealing
    ├── Lyophilization
    ├── Galvanic replacement reaction
    ├── Synthetic process
    ├── Incipient wetness impregnation
    ├── Synthesis approach
    ├── Silica templating
    ├── Synthetic approaches
    ├── Synthesis
    ├── Synthesis condition
    ├── Heteroatom doped
    ├── Reduction temperature
    ├── Hydrothermal ethanol reduction method
    ├── High-temperature pyrolysis
    ├── Immobilization via functional group
    ├── Dendrimer encapsulation
    ├── Hydrothermal treatment
    ├── Impregnation methods
    ├── Wet impregnation
    ├── Sol-gel approach
    ├── Self-assembly route
    ├── Synthetic strategies
    └── High-temperature self-assembly route
```

```
Thing
└── Reactions
    └── CO2 reduction
        ├── Electrochemical carbon dioxide reduction
        ├── Carbon dioxide reduction reaction
        ├── CO2 reduction reaction (CO2RR)
        ├── Electrochemical CO2-to-CO conversion
        ├── Electrochemical CO2 reduction reaction (CO2RR)
        ├── CO2 conversion
        ├── eCO2RR
        ├── CO2 electroreduction
        ├── Photocatalytic CO2 conversion
        ├── Photocatalytic CO2 reduction reaction
        ├── CO2 to CO conversion
        ├── Photocatalytic reduction
        ├── CO2 photoreduction
        ├── Catalytic CO2 conversion
        ├── CO2 hydrogenation
        └── Electroreduction
```

Example SACs Ontology (Llama 3.1:70b)

```
Thing
└── Synthesis methods
    ├── Catalyst synthetic strategies
    ├── Nanoconfined ILs strategy
    ├── Solid liquid interface engineering
    ├── Confinement
    ├── Synthesis
    ├── Strategies
    ├── Postprocessing solution treatments
    ├── Acidic leaching
    ├── Sol-gel approach
    ├── Incipient wetness impregnation
    ├── Annealing
    ├── Lyophilization
    ├── Galvanic replacement reaction
    ├── Atomic layer deposition
    ├── Co-precipitation
    ├── Alloying
    ├── Synthetic process
    ├── NH3 atmosphere annealing
    ├── Hydrothermal treatment
    ├── Oxychlorination
    ├── Iodo hydrocarbon treatment
    ├── NO/CO treatment
    ├── Dendrimer encapsulation
    ├── Repetitive oxidation and reduction
    ├── Immobilization via functional group
    ├── Pyrolysis procedure
    ├── Bimodal template based synthesis strategies
    │
```

```
Thing
└── Reactions
    ├── CO2 molecules
    ├── CO2 reduction
    ├── Electrochemical CO2 reduction reaction (CO2RR)
    ├── Carbon dioxide
    ├── CO2 emissions
    ├── CO2 reduction reaction (CO2RR)
    ├── Anthropogenic CO2 emissions
    ├── Carbon dioxide reduction reaction
    ├── Electrochemical carbon dioxide reduction
    ├── Photocatalytic CO2 reduction reaction
    ├── CO2 to CO conversion
    ├── dioxide
    ├── eCO2RR
    ├── CO2 electroreduction
    ├── CO2 photoreduction
    ├── CO2 conversion
    ├── CO2 activation
    ├── Electrochemical CO2 to CO conversion
    ├── <remaining omitted for clarity>
```

```
├── Calcination
├── Wet impregnation
├── Reduction
├── Impregnation methods
├── Rational design
├── Two-step approach
├── Ni-TAPc anchoring strategies
├── High-temperature aging
├── Synthetic strategies
├── Scale up flexibility
├── Low cost
├── Self-assembly route
├── High-temperature self-assembly route
├── Aging treatment
├── Synthesis approach
├── Acid wash steps
├── Silica templating
├── Sacrificial Zn based metal organic framework
├── Synthetic approaches
├── NaOH etching
├── Pyrolysis
├── Rational identification
├── Synthesis condition
├── High temperature pyrolysis
├── Hydrothermal ethanol reduction method
├── Catalyst design
├── Ionic exchange
├── Modulation
├── Composition evolution
├── Metal salt
├── Structure performance relationships
├── o phenylenediamine
├── Ultrahigh vacuum surface science procedures
├── Multistep pyrolysis process
├── Wet-chemistry methods
├── Multistep pyrolysis method
├── Physical techniques
├── Mass-selected soft-landing
├── Pyrolysis process
├── Atom beams
├── Growth mechanism
├── Post treatment processes
├── Reconnaissance study
├── Ketoamine condensation reaction
├── Multiscale tuning
├── Ni salts
├── Methodology
├── Ni precursor
├── Formation mechanism
├── Distribution
└── Metal precursor
```

A.5 SACBench: Benchmark for SAC Synthesis Procedures

SACBench is a comprehensive benchmark designed to evaluate the performance of systems that generate experimental procedures for the synthesis of Single-Atom Catalysts (SACs). The benchmark consists of 50 input-output pairs, where the input specifies a desired SAC and the output is the correct synthesis procedure.

The evaluation metrics used aim to assess the validity and correctness of a generated synthesis suggestion, in chemically meaningful terms.

Some metrics include:

1. Procedure Accuracy: Measures the overall correctness of the generated procedure.
2. Procedure Completeness: Assesses how comprehensive the generated procedure is compared to the reference.
3. Procedure Order: Evaluates the correct sequencing of steps in the generated procedure.
4. Chemical Identification: Includes recall, precision, F1 score, and accuracy for identifying correct chemicals in the procedure.
5. Metal Identification: Measures recall, precision, F1 score, and accuracy for correctly identifying the metal component of the SAC.
6. Support Identification: Evaluates recall, precision, F1 score, and accuracy for correctly identifying the support material in the SAC synthesis.

Figure 6 shows some general statistics about the test dataset, and the co occurrences between different variables.

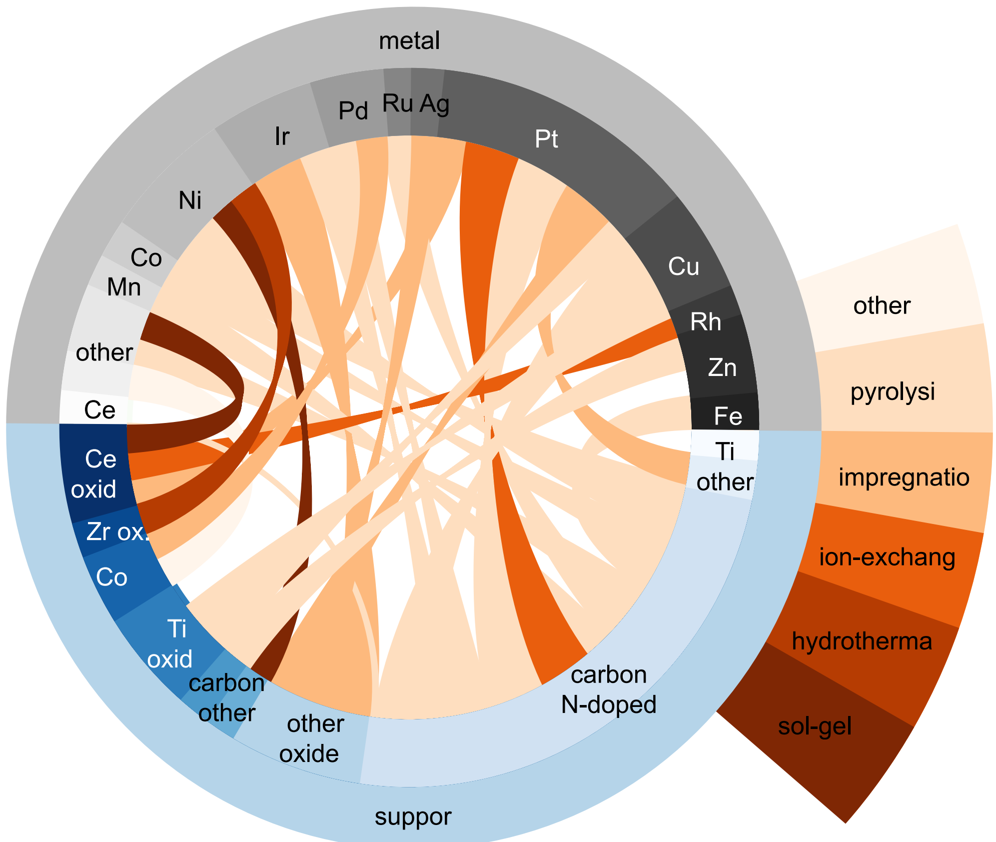

Figure 6: Descriptive statistics of the benchmark created for this work.

A.5.1 SAC Research Papers Corpus

The corpus of 500 recent research papers on Single-Atom Catalysts (SACs) used for ontology generation includes publications from top journals in catalysis and materials science from the past 5 years. The papers cover various aspects of SACs, including synthesis methods, characterization techniques, and applications.

The research papers were obtained from Wiley Journals through Wiley's official API (Wiley-API (2024)).

A.6 Prompt Example

Here's an example prompt used in the `query_relationships` function for taxonomy extraction:

```
Given the following paper content, current taxonomy terms, and vocabulary
    ↪  to be queried, please identify 'isA' relationships between terms
    ↪ in the vocabulary and terms in the current taxonomy. Ensure that
    ↪ each relationship is supported by evidence from the paper content.

Paper content:
In the field of catalysis, single-atom catalysts represent a specialized
    ↪ form of catalysts, emerging from the parent concept of a catalyst
    ↪ but with isolated active sites at the atomic level. Their creation
```

```
    ↪ often involves various synthesis methods, with wet impregnation
    ↪ being a common technique to distribute the active metal atoms
    ↪ evenly on a support. Once synthesized, these catalysts can be
    ↪ characterized using X-ray absorption spectroscopy.

Current taxonomy terms:
- Reactions
- Catalyst
- Materials
- Synthesis method
- Characterization technique
- Preparation method

Vocabulary to be queried:
- Single-atom catalyst
- Wet impregnation
- X-ray absorption spectroscopy

Please format your response as a list of relationships in the form (
    ↪ parent_term, child_term), where parent_term is from the current
    ↪ taxonomy and child_term is from the vocabulary to be queried."
```

Listing 1: Prompt Example

```
Here is the list of relationships:

(Catalyst, Single-atom catalyst)
(Synthesis method, Wet impregnation)
(Characterization technique, X-ray absorption spectroscopy)
```

Listing 2: Response Example

A.7 Downstream evaluation of Ontologies

Evaluating the quality of generated ontologies requires either careful expert evaluation, typically involving committees of experts in the field Keet (2018), or downstream applications that use them as an integral part of the pipeline and provide quantitative result of some sort.

In our work, we opt for the downstream application on SAC Synthesis to compare two SAC ontologies generated with OntoGen, using LLMs of different capacity, namely Claude-3.5-Sonnet, and Llama-3.1-70B. We compare two variants of OntoRAG-simple: with and without a Translation Module. Additionally we include the results of the ZeroShot and CoT baselines for comparison. All the results in Tables 4 to 6 are results with gpt-4o-mini as LLM. The metrics used are defined in Appendix A.5.

Table 4: ZeroShot (Baseline)

	Procedure			Chemicals	Metal	Support
ontology	completeness	order	accuracy	accuracy	accuracy	accuracy
Claude	0.725011	0.400722	0.055564	0.130818	0.490196	0.549020
Llama	0.725011	0.400722	0.055564	0.130818	0.490196	0.549020

Table 5: CoT (Baseline)

	procedure			chemicals	metal	support
ontology	completeness	order	accuracy	accuracy	accuracy	accuracy
Claude	0.570561	0.321268	0.048420	0.141569	0.578431	0.490196
Llama	0.570561	0.321268	0.048420	0.141569	0.578431	0.490196

A.8 SACBench results & analysis

Table 6: OntoRAG-simple

ontology	procedure completeness	procedure order	procedure accuracy	chemicals accuracy	metal accuracy	support accuracy
Claude	**0.577304**	0.330314	0.044630	0.130324	**0.607843**	**0.490196**
Llama	0.536076	0.337008	0.038061	**0.138353**	0.509804	0.431373

Table 7: OntoRAG-simple-tm

ontology	procedure completeness	procedure order	procedure accuracy	chemicals accuracy	metal accuracy	support accuracy
Claude	**0.613198**	0.364592	0.049093	0.132388	**0.705882**	0.568627
Llama	0.593519	0.369136	0.049502	0.138899	0.647059	**0.607843**

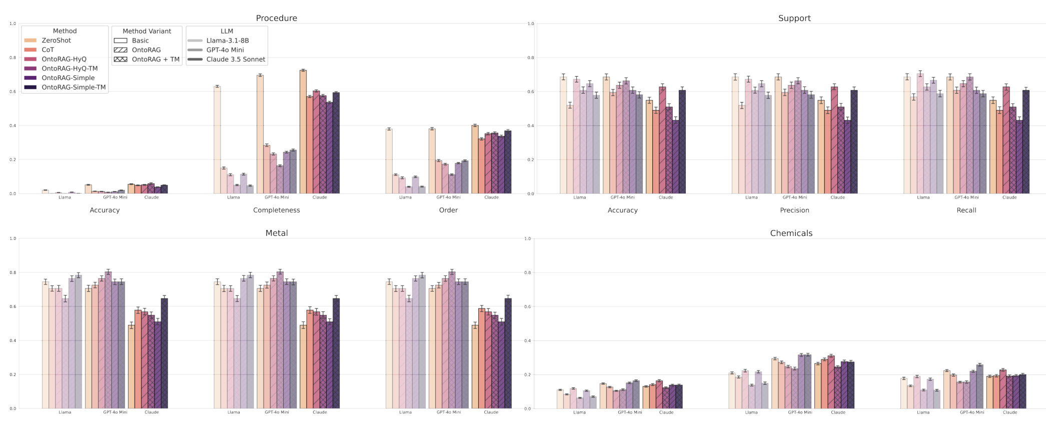

Figure 7: Complete results of multiple methods, and LLMs, on multiple metrics of the SACBench benchmark.

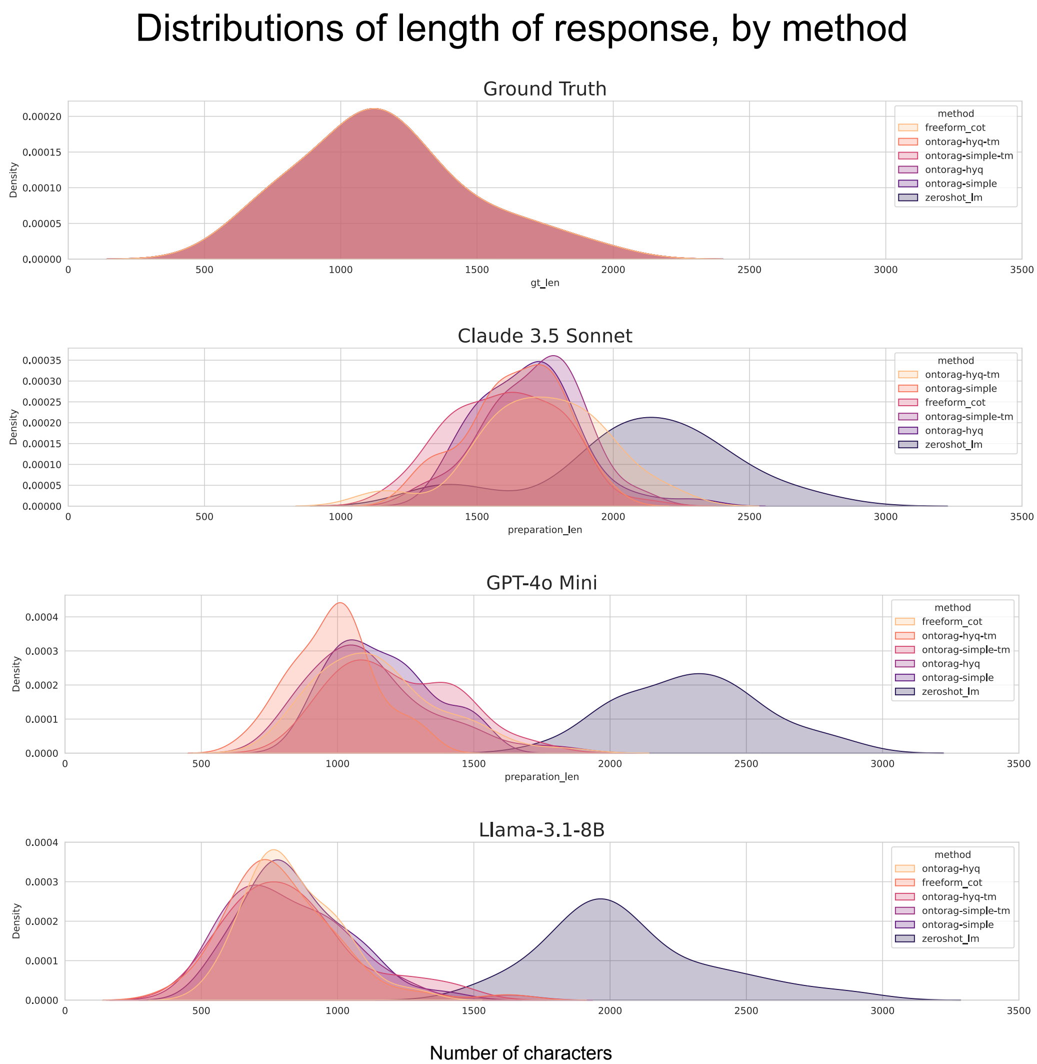

Figure 8: Distribution of response length for each LLM, by method. The plot shows a clear difference between the ZeroShot responses as compared to the rest of the methods.