
Wisent: A General Framework for Reliable Representation Identification and Representation Steering

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 Representation engineering is a powerful method for identifying and modifying
2 high-level concepts within the internal layers of large language models. Despite its
3 potential, real-life deployments of activation steering remain difficult. We present
4 Wisent, a flexible, open-source framework for monitoring and steering internal
5 activations of large language models. Practical applications of the framework show
6 XXX percent hallucination reduction, XXX percent improvement in coding ability
7 and deep personalization capabilities.

8 1 Introduction

9 Large language models, with billions of parameters and Internet-scale training dataset, have displayed
10 significant capabilities across a wide range of tasks, such as writing, coding or reasoning.

11 However, their internal mechanisms of generating the next token cannot be precisely explained, with
12 interactions between layers and parameters increasing in complexity as the size of these models
13 increases.

14 Experiments with representation engineering (also known as steering or activation steering) have
15 shown activation modification to be a powerful method of identifying and influencing high-level
16 concepts (representations) within the layers of an LLM. Despite strong empirical performance on
17 selected truthfulness, safety or personalization tasks, representation engineering methods lack a
18 universal formulation and a unifying framework for understanding the underlying phenomenon,
19 comparing methods and applying them to new problems.

20 We propose Wisent, a modular framework for analyzing the internal mechanisms within a large
21 language model and influencing them to improve performance and individual alignment. Wisent
22 surpasses state of the art performance in identifying particular behaviors

23 2 Representation Engineering Problem

24 We formulate the **Representation Engineering Problem** as the following:

25 For a given model M and a Representation

26 Basic primitives and definitions of key terms are outlined in Appendix A.

27 **3 Representation Reading**

28 **3.1 Classifier**

29 **3.2 Detection Handling Method**

30 **4 Representation Control**

31 **4.1 Classifier**

32 **References**

- 33 [1] David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba O. Alabi,
34 Xuanli He, Millicent Ochieng, Sara Hooker, et al. Irokobench: A new benchmark for african
35 languages in the age of large language models. *arXiv preprint arXiv:2406.03368*, 2024.
- 36 [2] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David
37 Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis
38 with large language models. *arXiv preprint arXiv:2108.07732*, 2021.
- 39 [3] Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim.
40 Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the
41 European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*,
42 pages 251–261, Valencia, Spain, 2017. Association for Computational Linguistics.
- 43 [4] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase
44 from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in
45 Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA, 2013. Association
46 for Computational Linguistics.
- 47 [5] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning
48 about physical commonsense in natural language. *arXiv preprint arXiv:1911.11641*, 2019.
- 49 [6] Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes
50 Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia
51 Specia, and Aleš Tamchyna. Findings of the 2014 workshop on statistical machine translation. In
52 *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore,
53 Maryland, USA, 2014. Association for Computational Linguistics.
- 54 [7] Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias
55 Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. Findings
56 of the 2016 conference on machine translation. In *Proceedings of the First Conference on
57 Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany, 2016.
58 Association for Computational Linguistics.
- 59 [8] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla
60 Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language
61 models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*,
62 2020.
- 63 [9] Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient
64 intent detection with dual sentence encoders. *arXiv preprint arXiv:2003.04807*, 2020.
- 65 [10] Federico Cassano, John Gouwar, Daniel Nguyen, Sydney Nguyen, Luna Phipps-Costin, Donald
66 Pinckney, Ming-Ho Yee, Yangtian Zi, Carolyn Jane Anderson, Molly Q. Feldman, Arjun Guha,
67 Michael Greenberg, and Abhinav Jangda. Multipl-e: A scalable and extensible approach to
68 benchmarking neural code generation. *arXiv preprint arXiv:2208.08227*, 2022.
- 69 [11] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. Semeval-
70 2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In
71 *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages
72 1–14, Vancouver, Canada, 2017. Association for Computational Linguistics.

- 73 [12] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto,
74 Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul
75 Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke
76 Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad
77 Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias
78 Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex
79 Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain,
80 William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra,
81 Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer,
82 Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech
83 Zaremba. Evaluating large language models trained on code. 2021.
- 84 [13] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and
85 Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions.
86 *arXiv preprint arXiv:1905.10044*, 2019.
- 87 [14] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick,
88 and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning
89 challenge. *arXiv preprint arXiv:1803.05457*, 2018.
- 90 [15] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christo-
91 pher Hesse, John Schulman, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, and
92 Jerry Tworek. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*,
93 2021.
- 94 [16] CodeParrot. Instructhumaneval, 2023.
- 95 [17] Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. The commitmentbank:
96 Investigating projection in naturally occurring discourse. In *Proceedings of Sinn und Bedeutung*
97 23, volume 2, pages 107–124, Bellaterra (Cerdanyola del Vallès), 2019. Universitat Autònoma
98 de Barcelona.
- 99 [18] Mingzhe Du, Anh Tuan Luu, Bin Ji, Liu Qian, and See-Kiong Ng. Mercury: A code efficiency
100 benchmark for code llms. *arXiv preprint arXiv:2402.07844*, 2024.
- 101 [19] Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, Kang Zhu, Minghao Liu,
102 Yiming Liang, and et al. Supergpqa: Scaling llm evaluation across 285 graduate disciplines.
103 *arXiv preprint arXiv:2502.14739*, 2025.
- 104 [20] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs.
105 *arXiv preprint arXiv:1903.00161*, 2019.
- 106 [21] Iker García-Ferrero and Begoña Altuna. Noticia: A clickbait article summarization dataset in
107 spanish. *arXiv preprint arXiv:2404.07611*, 2024.
- 108 [22] Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and
109 Noam Slonim. A large-scale dataset for argument quality ranking: Construction and analysis.
110 In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.
- 111 [23] Maxime Griot, Jean Vanderdonckt, Demet Yuksel, and Coralie Hemtinne. Pattern recognition
112 or medical knowledge? the problem with multiple-choice questions in medicine. In *Proceedings*
113 *of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*
114 *Papers)*, pages 5321–5341, Vienna, Austria, 2025. Association for Computational Linguistics.
- 115 [24] Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and
116 Preslav Nakov. Exams: A multi-subject high school examinations dataset for cross-lingual and
117 multilingual question answering. In *Proceedings of the 2020 Conference on Empirical Methods*
118 *in Natural Language Processing (EMNLP)*, pages 5427–5444. Association for Computational
119 Linguistics, 2020.
- 120 [25] Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo,
121 Collin Burns, Samir Puranik, Horace He, Dawn Song, and Jacob Steinhardt. Measuring coding
122 challenge competence with apps. *arXiv preprint arXiv:2105.09938*, 2021.

- 124 [26] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and
125 Jacob Steinhardt. Measuring massive multitask language understanding. In *Proceedings of the*
126 *International Conference on Learning Representations (ICLR)*, 2021.
- 127 [27] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn
128 Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset.
129 *arXiv preprint arXiv:2103.03874*, 2021.
- 130 [28] Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa
131 Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in*
132 *Neural Information Processing Systems*, volume 28, 2015.
- 133 [29] Srinivasan Iyer, Ioannis Konstas, Alvin Cheung, and Luke Zettlemoyer. Mapping language to
134 code in programmatic context. *arXiv preprint arXiv:1808.09588*, 2018.
- 135 [30] Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Ar-
136 mando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination
137 free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024.
- 138 [31] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. Triviaqa: A large scale
139 distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th*
140 *Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,
141 pages 1601–1611, Vancouver, Canada, 2017. Association for Computational Linguistics.
- 142 [32] Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and
143 Mikhail Burtsev. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack.
144 *arXiv preprint arXiv:2406.10149*, 2024.
- 145 [33] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris
146 Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion
147 Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav
148 Petrov. Natural questions: A benchmark for question answering research. *Transactions of the*
149 *Association for Computational Linguistics*, 7:452–466, 2019.
- 150 [34] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale
151 reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*, 2017.
- 152 [35] Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt,
153 Ryan A. Rossi, and Thien Huu Nguyen. Okapi: Instruction-tuned large language models
154 in multiple languages with reinforcement learning from human feedback. *arXiv preprint*
155 *arXiv:2307.16039*, 2023.
- 156 [36] Yuhang Lai, Chengxi Li, Yiming Wang, Tianyi Zhang, Ruiqi Zhong, Luke Zettlemoyer, Scott
157 Wen-tau Yih, Daniel Fried, Sida Wang, and Tao Yu. Ds-1000: A natural and reliable benchmark
158 for data science code generation. *arXiv preprint arXiv:2211.11501*, 2022.
- 159 [37] Ken Lang. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International*
160 *Conference on Machine Learning*, pages 331–339, 1995.
- 161 [38] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic
162 human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.
- 163 [39] Marco Lippi, Przemysław Pałka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Mick-
164 litz, Giovanni Sartor, and Paolo Torroni. Claudette: an automated detector of potentially unfair
165 clauses in online terms of service. In *Artificial Intelligence and Law*, volume 27, pages 117–139.
166 Springer, 2019.
- 167 [40] Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. Evaluating the
168 logical reasoning ability of chatgpt and gpt-4. *arXiv preprint arXiv:2304.03439*, 2023.
- 169 [41] Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by
170 chatgpt really correct? rigorous evaluation of large language models for code generation. In
171 *NeurIPS 2023 Datasets and Benchmarks Track*, 2023.

- 172 [42] Junnan Liu, Hongwei Liu, Linchen Xiao, Ziyi Wang, Kuikun Liu, Songyang Gao, Wenwei
173 Zhang, Songyang Zhang, and Kai Chen. Are your llms capable of stable reasoning? *arXiv*
174 preprint arXiv:2412.13147, 2024.
- 175 [43] Shuai Lu, Daya Guo, Shuo Ren, Junjie Huang, Alexey Svyatkovskiy, Ambrosio Blanco, Colin
176 Clement, Dawn Drain, Daxin Jiang, Duyu Tang, Ge Li, Lidong Zhou, Linjun Shou, Long
177 Zhou, Michele Tufano, Ming Gong, Ming Zhou, Nan Duan, Neel Sundaresan, Shao Kun Deng,
178 Shengyu Fu, and Shujie Liu. Codexglue: A machine learning benchmark dataset for code
179 understanding and generation. *arXiv preprint arXiv:2102.04664*, 2021.
- 180 [44] Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. Good debt or
181 bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for*
182 *Information Science and Technology*, 65(4):782–796, 2014.
- 183 [45] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large
184 annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330,
185 1993.
- 186 [46] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture
187 models. *arXiv preprint arXiv:1609.07843*, 2016.
- 188 [47] Shen-Yun Miao, Chao-Chun Liang, and Keh-Yih Su. A diverse corpus for evaluating and
189 developing english math word problem solvers. *arXiv preprint arXiv:2106.15772*, 2021.
- 190 [48] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor
191 conduct electricity? a new dataset for open book question answering. In *Proceedings of the*
192 *2018 Conference on Empirical Methods in Natural Language Processing*. Association for
193 Computational Linguistics, 2018.
- 194 [49] Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. Ethos: an
195 online hate speech detection dataset. *Complex & Intelligent Systems*, 8:4663–4678, 2022.
- 196 [50] Shahriar Kabir Nahin, Rabindra Nath Nandi, Sagor Sarker, Quazi Sarwar Muhtaseem, Md Kow-
197 sher, Apu Chandraw Shill, Md Ibrahim, Mehadi Hasan Menon, Tareq Al Muntasir, and Firoj
198 Alam. Titullms: A family of bangla llms with comprehensive benchmarking. *arXiv preprint*
199 arXiv:2502.11187, 2025.
- 200 [51] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don’t give me the details, just the sum-
201 mary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of*
202 *the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807,
203 Brussels, Belgium, 2018. Association for Computational Linguistics.
- 204 [52] Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, and et al. Humanity’s last exam. *arXiv*
205 preprint arXiv:2501.14249, 2025.
- 206 [53] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don’t know: Unanswerable
207 questions for squad. *arXiv preprint arXiv:1806.03822*, 2018.
- 208 [54] Siva Reddy, Danqi Chen, and Christopher D. Manning. Coqa: A conversational question
209 answering challenge. *arXiv preprint arXiv:1808.07042*, 2019.
- 210 [55] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien
211 Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a
212 benchmark. *arXiv preprint arXiv:2311.12022*, 2023.
- 213 [56] Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. Choice of plausible
214 alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium on*
215 *Logical Formalizations of Commonsense Reasoning*, Stanford, CA, 2011.
- 216 [57] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An
217 adversarial winograd schema challenge at scale. *arXiv preprint arXiv:1907.10641*, 2019.

- 218 [58] Tim Schopf, Daniel Braun, and Florian Matthes. Evaluating unsupervised text classification:
219 Zero-shot and similarity-based approaches. In *Proceedings of the 2022 6th International
220 Conference on Natural Language Processing and Information Retrieval (NLPiR)*, pages 6–15,
221 Bangkok, Thailand, 2022. ACM.
- 222 [59] Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui,
223 Daniel Vila-Suero, Peerat Limkonchotiwat, et al. Global mmlu: Understanding and addressing
224 cultural and linguistic biases in multilingual evaluation. *arXiv preprint arXiv:2412.03304*, 2024.
- 225 [60] Don Tuggener, Pius von Däniken, Thomas Peetz, and Mark Cieliebak. Ledgar: A large-scale
226 multi-label corpus for text classification of legal provisions in contracts. In *Proceedings of the
227 Twelfth Language Resources and Evaluation Conference*, pages 1235–1241, Marseille, France,
228 2020. European Language Resources Association.
- 229 [61] Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi, Rodrigo Agerri, and Aitor Soroa. Basqueglue:
230 A natural language understanding benchmark for basque. In *Proceedings of the Thirteenth
231 Language Resources and Evaluation Conference*, pages 1603–1612, Marseille, France, 2022.
232 European Language Resources Association.
- 233 [62] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill,
234 Omer Levy, and Samuel R. Bowman. Super glue: A stickier benchmark for general-purpose
235 language understanding systems. In *Advances in Neural Information Processing Systems*,
236 volume 32, 2019.
- 237 [63] Shiqi Wang, Zheng Li, Haifeng Qian, Chenghao Yang, Zijian Wang, Mingyue Shang, Varun
238 Kumar, Samson Tan, Baishakhi Ray, Parminder Bhatia, Ramesh Nallapati, Murali Krishna
239 Ramanathan, Dan Roth, and Bing Xiang. Recode: Robustness evaluation of code generation
240 models. *arXiv preprint arXiv:2212.10264*, 2022.
- 241 [64] Yiming Wang, Pei Zhang, Jialong Tang, Haoran Wei, Baosong Yang, Rui Wang, Chenshu Sun,
242 Feitong Sun, Jiran Zhang, Junxuan Wu, Qiqian Cang, Yichang Zhang, Fei Huang, Junyang Lin,
243 Fei Huang, and Jingren Zhou. Polymath: Evaluating mathematical reasoning in multilingual
244 contexts. *arXiv preprint arXiv:2504.18428*, 2025.
- 245 [65] Pengcheng Yin, Bowen Deng, Edgar Chen, Bogdan Vasilescu, and Graham Neubig. Learning
246 to mine aligned code and natural language pairs from stack overflow. In *Proceedings of the 15th
247 IEEE/ACM International Conference on Mining Software Repositories (MSR)*, pages 476–486.
248 IEEE, 2018.
- 249 [66] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. Swag: A large-scale adversarial
250 dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on
251 Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium, 2018.
252 Association for Computational Linguistics.
- 253 [67] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a
254 machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.
- 255 [68] Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme.
256 Record: Bridging the gap between human and machine commonsense reading comprehension.
257 *arXiv preprint arXiv:1810.12885*, 2018.
- 258 [69] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text
259 classification. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

260 **A Wisent Primitives**

261 **A.1 Model**

262 **A.2 Contrastive Pair**

263 **A.3 Activations**

264 **A.4 Activation Collection Method**

265 **A.5 Additional Utilities**

266 **B Representation Reading Functionalities**

267 **B.1 Classifier**

268 **B.2 Detection Handling Method**

269 **C Representation Control Functionalities**

270 **D Ablation**

271 **A All supported benchmarks**

272 This section enumerates all benchmarks used in our study, the task traits, the evaluation protocol, and
273 the contrastive pair generation method applied to produce minimally perturbed negative targets. We
274 first merged the *coding* and *mathematics* benchmark lists you provided and then appended them to
275 the original master list.

276 **Contrastive pair generation methods (definitions)**

277 **Reading Comprehension Abstention Swap** [RC-Abstain] For extractive/open-domain RC: positive
278 is the gold span; negative is an abstention (e.g., “Not provided in the text.”). If gold is
279 *No answer*, the negative is a confident but wrong claim.

280 **Conversational Reading Comprehension Abstention** [ConvRC-Abstain] As RC-Abstain, but
281 with dialogue context (CoQA). Negatives are generic abstentions; yes/no items are flipped
282 when applicable.

283 **Language Modeling Corrupted Continuation** [LM-CorruptCont] Language modeling: positive
284 is the true continuation; negative is a corrupted continuation (local shuffles/randomization)
285 to break coherence.

286 **Generic answer** [Generic] Negative is some generic answer which is incorrect.

287 **Letter shuffling** [L-Shuff] Negative is created by shuffling letters of positive.

288 **Two-Choice Flip** [2C-Flip] Two-option tasks (PIQA, COPA, WinoGrande, CB): negative is simply
289 the other option.

290 **Multichoice First Distractor** [MC-FirstDistr] Multi-choice tasks: negative = the first incorrect
291 option in the provided order (deterministic).

292 **Multichoice Random Distractor** [MC-RandDistr] Multi-choice tasks: negative = a randomly cho-
293 sen incorrect option from the same set (used for GPQA).

294 **Multichoice Letter Swap** [MC-LetterSwap] Multi-choice tasks scored over option letters (Truth-
295 fulQA MC1/MC2): negative = the first incorrect letter.

296 **Exact Match Partial Mask** [EM-PartialMask] Exact-match free-form answers (HLE-EM): nega-
297 tive is the gold text with partial token masking (approximately 1/3 words, or partial masking
298 for single-word answers).

299 **Keyword-Preserving Token Deletion** [KP-Del] Coding tasks: negative program created by delet-
300 ing non-keyword tokens while preserving syntax-critical keywords; aims to remain plausible
301 but fail unit tests.

- 302 **Summary Content-Word Drop** [Summ-WordDrop] Code-to-text summarization: negative descrip-
 303 tion formed by dropping content words (nouns/verbs) while keeping scaffolding words to
 304 preserve superficial form.
- 305 **Numeric Offset (+1) Perturbation** [Num+1] Math QA: negative is the correct numeric answer
 306 offset by a small integer (typically +1); for non-integer answers, apply the minimal unit
 307 offset.
- 308 **Evaluation types (definitions)**
- 309 **Log-likelihood option scoring** [LL] The model scores each provided option/target by conditional
 310 log-probability given the prompt. Metrics typically compute accuracy over the highest-
 311 likelihood choice (MC tasks) or compare likelihoods of gold vs. negative targets.
- 312 **Text generation string matching** [TG] The model generates free-form text (or a number), which
 313 is then judged by task-specific metrics (e.g., exact match on numerical value for
 314 GSM8K/MATH; span/string matching for RC tasks; structured checks for DROP). Used
 315 also for CoT/generative GPQA variants and HLE-Exact-Match.
- 316 **Perplexity (language modeling)** [PPL] The model’s next-token distribution is evaluated over a
 317 reference text to compute Perplexity (lower is better). Used for language-modeling corpora
 318 like WikiText.
- 319 **Code execution against unit tests** [CE] The model generates code, which is executed in a sandbox
 320 against unit tests provided by a dataset (e.g., pass@1). Applies to HumanEval/MBPP/APPS,
 321 MultiPL-E, DS-1000, LiveCodeBench, etc.

Table 1: Benchmarks (short names), evaluation abbreviations, contrastive
 method (short), and traits. Versions merged where applicable.

Benchmark	Eval	Method [CM]	Traits
DROP [20]	[TG]	RC-Abstain	reading comprehension
ReCoRD [68]	[TG]	RC-Abstain	reading comprehension
SQuAD2 [53]	[TG]	RC-Abstain	reading comprehension
WebQuestions [4]	[TG]	RC-Abstain	factual QA
Natural Questions [33]	[TG]	RC-Abstain	factual QA
TriviaQA [31]	[TG]	RC-Abstain	factual QA
CoQA [54]	[TG]	ConvRC-Abstain	conversational RC
BoolQ [13]	[LL]	2C-Flip	boolean RC
WinoGrande [57]	[LL]	2C-Flip	commonsense
PIQA [5]	[LL]	2C-Flip	commonsense
COPA [56]	[LL]	2C-Flip	causal reasoning
HellaSwag [67]	[LL]	MC-FirstDistr	commonsense
SWAG [66]	[LL]	MC-FirstDistr	commonsense
OpenBookQA [48]	[LL]	MC-FirstDistr	science MCQ
ARC [14]	[LL]	MC-FirstDistr	science MCQ
RACE [34]	[LL]	MC-FirstDistr	RC (MC)
MMLU [26]	[LL]	MC-FirstDistr	multi-subject exams
GPQA [55]	[LL]/[TG]	MC-RandDistr	expert STEM exams
SuperGPQA [19]	[LL]	MC-FirstDistr	expert STEM exams
HLE [52]	[TG]/[LL]	EM-PartialMask; MC-FirstDistr	expert exams
GSM8K [15]	[TG]	Num+1	mathematics
ASDiv [47]	[TG]	Num+1	mathematics
Arithmetic [8]	[TG]	Num+1	mathematics
MATH [27]	[TG]	Num+1	mathematics (contest)
MATH-500 [27]	[TG]	Num+1	mathematics (contest)
AIME []	[TG]	Num+1	mathematics (contest)

Benchmark	Eval	Method [CM]	Traits
HMMT []	[TG]	Num+1	mathematics (contest)
PolyMath [64]	[TG]	Num+1	mathematics (multiling.)
LiveMathBench [42]	[TG]	Num+1	mathematics (EN/ZH)
MBPP [2]	[CE]	KP-Del	coding (Python)
HumanEval [12]	[CE]	KP-Del	coding (Python)
CoNaLa [65]	[CE]	KP-Del	coding (Python)
CONCODE [29]	[CE]	KP-Del	coding (Java)
Mercury [18]	[CE]	KP-Del	coding (multi-language)
HumanEval+ [41]	[CE]	KP-Del	coding (Python)
InstructHumanEval [16]	[CE]	KP-Del	coding (Python)
MBPP+ [41]	[CE]	KP-Del	coding (Python)
APPS [25]	[CE]	KP-Del	coding (Python)
DS-1000 [36]	[CE]	KP-Del	coding (Python)
MultiPL-E [10]	[CE]	KP-Del	coding (multi-language)
CodeXGLUE [43]	[TG]	Summ-WordDrop	coding (code-to-text)
ReCode [63]	[CE]	KP-Del	coding (Python)
LiveCodeBench [30]	[CE]	KP-Del	coding (Python)
TruthfulQA [38]	[LL]	MC-LetterSwap	truthfulness
CB [17]	[LL]	2C-Flip	NLI
WikiText (2/103) [46]	[PPL]	LM-CorruptCont	language modeling

322 **Category legend**

- RC/ODQA
- Multi-choice Reasoning
- Exams & Knowledge Tests
- Mathematics
- Coding
- Other (Truthfulness/NLI/LM)

Abbreviation legend

- | | | | |
|-------|--------------------------------|----------------|------------------------------|
| [LL] | Log-likelihood option scoring | RC-Abstain | RC abstention swap |
| [TG] | Text generation (string match) | ConvRC-Abstain | Conversational RC abstention |
| [PPL] | Perplexity (LM) | LM-CorruptCont | LM corrupted continuation |
| [CE] | Code execution vs. unit tests | 2C-Flip | Two-choice flip |
| | | MC-FirstDistr | First distractor (MC) |
| | | MC-RandDistr | Random distractor (MC) |
| | | MC-LetterSwap | Letter swap (MC) |
| | | Bool-Flip | Boolean flip |
| | | EM-PartialMask | Exact-match partial mask |
| | | KP-Del | Keyword-preserving deletion |
| | | Summ-WordDrop | Summary word drop |
| | | Num+1 | Numeric offset (+1) |

Method [CM] codes

Table 2: Benchmarks (short names), evaluation abbreviations, contrastive method (short), and traits. Versions merged where applicable.

Benchmark	Eval	Method [CM]	Traits
20_newsgroups [37]	[TG]	MC-FirstDistr	reasoning
ag_news [69]	[TG]	MC-FirstDistr	reasoning
argument_topic [22]	[TG]	MC-FirstDistr	reasoning
banking77 [9]	[TG]	MC-FirstDistr	reasoning
boolq [13]	[LL]	2C-Flip	reasoning
boolq-seq2seq [13]	[TG]	2C-Flip	reasoning
cb [17]	[LL]	MC-FirstDistr	reasoning
claim stance topic [3]	[TG]	MC-FirstDistr	reasoning
cnn dailymail [28]	[TG]	Generic	reasoning
dpedia 14 [69]	[TG]	MC-FirstDistr	reasoning
ethos binary [49]	[TG]	MC-FirstDistr	reasoning
financial tweets [44]	[TG]	MC-FirstDistr	reasoning
squadv2 [53]	[TG]	RC-Abstain	reasoning

Benchmark	Eval	Method [CM]	Traits
logieval [40]	[TG]	MC-FirstDistr	reasoning
ledgar [60]	[TG]	MC-FirstDistr	reasoning
logieval [40]	[TG]	MC-FirstDistr	reasoning
penn treebank [45]	[PPL]	LM-CorruptCont	reasoning
medical abstracts [58]	[TG]	MC-FirstDistr	reasoning
unfair tos [39]	[TG]	LM-CorruptCont	reasoning
record [68]	[LL]	MC-FirstDistr	reasoning
stsbs [11]	[TG]	2C-Flip	reasoning
sglue-rte [62]	[LL]	2C-Flip	reasoning
xsum [51]	[TG]	Generic	reasoning
yashoo answers topics [69]	[TG]	MC-FirstDistr	reasoning

Table 3: Benchmarks (short names), evaluation abbreviations, contrastive method (short), and traits. Versions merged where applicable.

Benchmark	Eval	Method [CM]	Traits
afrimgsm direct amh [1]	[TG]	Num+1	mathematics
aime []	[TG]	Num+1	mathematics
aime2024 []	[TG]	Num+1	mathematics
aime2025 []	[TG]	Num+1	mathematics
gsm [15]	[TG]	Num+1	mathematics
hmmt []	[TG]	Num+1	mathematics
math [27]	[TG]	Num+1	mathematics
math500 [27]	[TG]	Num+1	mathematics
polymath [64]	[TG]	Num+1	mathematics
livemathbench [42]	[TG]	Num+1	mathematics

Table 4: Benchmarks (short names), evaluation abbreviations, contrastive method (short), and traits. Versions merged where applicable.

Benchmark	Eval	Method [CM]	Traits
conala [65]	[TG]	L-Shuff	coding

Table 5: Benchmarks (short names), evaluation abbreviations, contrastive method (short), and traits. Versions merged where applicable.

Benchmark	Eval	Method [CM]	Traits
afrimmlu direct amh [1]	[LL]	MC-FirstDistr	multilingual
afrixnli en direct amh [1]	[TG]	MC-FirstDistr	multilingual
arabic exams [24]	[LL]	MC-FirstDistr	multilingual
bangla mmlu [50]	[LL]	MC-FirstDistr	multilingual
basque glue [61]	[LL]	2C-Flip	multilingual
copa [56]	[LL]	2C-Flip	multilingual
global mmlu [59]	[LL]	MC-FirstDistr	multilingual
m mmlu [35]	[LL]	MC-FirstDistr	multilingual
m mmlu [35]	[LL]	2C-Flip	multilingual
noticia [21]	[LL]	Generic	multilingual
phrases ca-va []	[TG]	Generic	multilingual
wmt14 [6]	[TG]	L-Shuff	multilingual

Benchmark	Eval	Method [CM]	Traits
wmt16 [7]	[TG]	L-Shuff	multilingual

Table 6: Benchmarks (short names), evaluation abbreviations, contrastive method (short), and traits. Versions merged where applicable.

Benchmark	Eval	Method [CM]	Traits
babilong [32]	[LL]	MC-RandDistr	longcontext

Table 7: Benchmarks (short names), evaluation abbreviations, contrastive method (short), and traits. Versions merged where applicable.

Benchmark	Eval	Method [CM]	Traits
glianorex [23]	[LL]	MC-FirstDistr	medical

Table 8: Benchmarks (short names), evaluation abbreviations, contrastive method (short), and traits. Versions merged where applicable.

Benchmark	Eval	Method [CM]	Traits
wikitext103 [46]	[PPL]	LM-CorruptCont	general knowledge

³²³ **B Per-Task Results**

³²⁴ **C Detailed Classification Results**

³²⁵ **D Benchmark-Aided Steering Results**

³²⁶ **E Optimal Sample Size Calculations**

³²⁷ **F Fully Synthetic Generation**

³²⁸ **G Agentic Capabilities**