| Task | *Chinchilla* | *Gopher* | Task | *Chinchilla* | *Gopher* |
|---|---|---|---|---|---|
| abstract_algebra | 31.0 | 25.0 | anatomy | 70.4 | 56.3 |
| astronomy | 73.0 | 65.8 | business_ethics | 72.0 | 70.0 |
| clinical_knowledge | 75.1 | 67.2 | college_biology | 79.9 | 70.8 |
| college_chemistry | 51.0 | 45.0 | college_computer_science | 51.0 | 49.0 |
| college_mathematics | 32.0 | 37.0 | college_medicine | 66.5 | 60.1 |
| college_physics | 46.1 | 34.3 | computer_security | 76.0 | 65.0 |
| conceptual_physics | 67.2 | 49.4 | econometrics | 38.6 | 43.0 |
| electrical_engineering | 62.1 | 60.0 | elementary_mathematics | 41.5 | 33.6 |
| formal_logic | 33.3 | 35.7 | global_facts | 39.0 | 38.0 |
| high_school_biology | 80.3 | 71.3 | high_school_chemistry | 58.1 | 47.8 |
| high_school_computer_science | 58.0 | 54.0 | high_school_european_history | 78.8 | 72.1 |
| high_school_geography | 86.4 | 76.8 | high_school_gov_and_politics | 91.2 | 83.9 |
| high_school_macroeconomics | 70.5 | 65.1 | high_school_mathematics | 31.9 | 23.7 |
| high_school_microeconomics | 77.7 | 66.4 | high_school_physics | 36.4 | 33.8 |
| high_school_psychology | 86.6 | 81.8 | high_school_statistics | 58.8 | 50.0 |
| high_school_us_history | 83.3 | 78.9 | high_school_world_history | 85.2 | 75.1 |
| human_aging | 77.6 | 66.4 | human_sexuality | 86.3 | 67.2 |
| international_law | 90.9 | 77.7 | jurisprudence | 79.6 | 71.3 |
| logical_fallacies | 80.4 | 72.4 | machine_learning | 41.1 | 41.1 |
| management | 82.5 | 77.7 | marketing | 89.7 | 83.3 |
| medical_genetics | 69.0 | 69.0 | miscellaneous | 84.5 | 75.7 |
| moral_disputes | 77.5 | 66.8 | moral_scenarios | 36.5 | 40.2 |
| nutrition | 77.1 | 69.9 | philosophy | 79.4 | 68.8 |
| prehistory | 81.2 | 67.6 | professional_accounting | 52.1 | 44.3 |
| professional_law | 56.5 | 44.5 | professional_medicine | 75.4 | 64.0 |
| professional_psychology | 75.7 | 68.1 | public_relations | 73.6 | 71.8 |
| security_studies | 75.9 | 64.9 | sociology | 91.0 | 84.1 |
| us_foreign_policy | 92.0 | 81.0 | virology | 53.6 | 47.0 |
| world_religions | 87.7 | 84.2 | | | |

Table A6 | ***Chinchilla* MMLU results.** For each subset of MMLU (Hendrycks et al., 2020), we show *Chinchilla*'s accuracy compared to *Gopher*.

| **Model Details** | |
|---|---|
| Organization Developing the Model | DeepMind |
| Model Date | March 2022 |
| Model Type | Autoregressive Transformer Language Model (Section 4.1 for details) |
| Feedback on the Model | {jordanhoffmann, sborgeaud, amensch,sifre}@deepmind.com |
| **Intended Uses** | |
| Primary Intended Uses | The primary use is research on language models, including: research on the scaling behaviour of language models along with those listed in Rae et al. (2021). |

| | |
|---|---|
| Primary Intended Users | DeepMind researchers. We will not make this model available publicly. |
| Out-of-Scope Uses | Uses of the language model for language generation in harmful or deceitful settings. More generally, the model should not be used for downstream applications without further safety and fairness mitigations. |
| **Factors** | |
| Card Prompts – Relevant Factor | Relevant factors include which language is used. Our model is trained on English data. Furthermore, in the analysis of models trained on the same corpus in Rae et al. (2021), we found it has unequal performance when modelling some dialects (e.g., African American English). Our model is designed for research. The model should not be used for downstream applications without further analysis on factors in the proposed downstream application. |
| Card Prompts – Evaluation Factors | See the results in Rae et al. (2021) which analyzes models trained on the same text corpus. |
| **Metrics** | |
| Model Performance Measures | • Perplexity and bits per byte on language modelling datasets<br>• Accuracy on completion tasks, reading comprehension, MMLU, BIG-bench and fact checking.<br>• Exact match accuracy for question answering.<br>• Generation toxicity from Real Toxicity Prompts (RTP) alongside toxicity classification accuracy.<br>• Gender and occupation bias. Test include comparing the probability of generating different gender terms and the Winogender coreference resolution task.<br><br>We principally focus on *Chinchilla*'s performance compared to *Gopher* on text likelihood prediction. |
| Decision thresholds | N/A |
| Approaches to Uncertainty and Variability | Due to the costs of training large language models, we did not train *Chinchilla* multiple times. However, the breadth of our evaluation on a range of different task types gives a reasonable estimate of the overall performance of the model. Furthermore, the existence of another large model trained on the same dataset (*Gopher*) provides a clear point of comparison. |
| **Evaluation Data** | |

| | |
|---|---|
| Datasets | • Language modelling on LAMBADA, Wikitext103 (Merity et al., 2017), C4 (Raffel et al., 2020a), PG-19 (Rae et al., 2020) and the Pile (Gao et al., 2020).<br>• Language understanding, real world knowledge, mathematical and logical reasoning on the Massive Multitask Language Understanding (MMLU) benchmark (Hendrycks et al., 2020) and on the "Beyond the Imitation Game Benchmark" (BIG-bench) (BIG-bench collaboration, 2021).<br>• Question answering (closed book) on Natural Questions (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017).<br>• Reading comprehension on RACE (Lai et al., 2017)<br>• Common sense understanding on HellaSwag (Zellers et al., 2019), PIQA (Bisk et al., 2020), Winogrande (Sakaguchi et al., 2020), SIQA (Sap et al., 2019), BoolQ (Clark et al., 2019), and TruthfulQA (Lin et al., 2021). |
| Motivation | We chose evaluations from Rae et al. (2021) to allow us to most directly compare to *Gopher*. |
| Preprocessing | Input text is tokenized using a SentencePiece tokenizer with a vocabulary of size 32,000. Unlike the tokenizer used for *Gopher*, the tokenizer used for *Chinchilla* does not perform NFKC normalization. |
| **Training Data** | |
| The same dataset is used as in Rae et al. (2021). Differences in sampling are shown in Table A1. | |
| **Quantitative Analyses** | |
| Unitary Results | Section 4.2 gives a detailed description of our analysis. Main take-aways include:<br>• Our model is capable of outputting toxic language as measured by the PerspectiveAPI. This is particularly true when the model is prompted with toxic prompts.<br>• Gender: Our model emulates stereotypes found in our dataset, with occupations such as "dietician" and "receptionist" being more associated with women and "carpenter" and "sheriff" being more associated with men.<br>• Race/religion/country sentiment: Prompting our model to discuss some groups leads to sentences with lower or higher sentiment, likely reflecting text in our dataset. |