SKILL-MIX: A Flexible and Expandable Family of Evaluations for AI Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

With LLMs shifting their role from statistical modeling of language to serving as general-purpose AI agents, how should LLM evaluations change? Arguably, a key ability of an AI agent is to flexibly combine, as needed, the basic skills it has learned. This capability to combine skills plays an important role in (human) pedagogy and also in a recent paper on emergence phenomena (Arora & Goyal, 2023). A new evaluation, SKILL-MIX , is introduced to measure this capability. Using a list of N skills the evaluator repeatedly picks random subsets of k skills and asks the LLM to produce text combining that subset of skills. Since the number of subsets grows like N^k, for even modest k this evaluation will, with high probability, require the LLM to produce text it has not seen in the training set. The paper develops a methodology for (a) designing and administering such an evaluation, and (b) automatic grading (plus spot-checking by humans) of the results using GPT-4 as well as the open LLaMA-2 70B model.

Administering a version of SKILL-MIX to popular chatbots gave results that, while generally in line with prior expectations, contained surprises. Sizeable differences exist among model capabilities —including suspected cases of "cramming for the leaderboard"— that are not captured by their ranking on popular LLM leaderboards. Our methodology can flexibly change to future models and model capabilities, by expanding the set of skills being tested and increasing k. By publicly releasing the SKILL-MIX methodology, we hope it may grow into an eco-system of open evaluations for AI capabilities, including in multi-modal settings. These may serve as more trustworthy gauges of model capabilities than current leaderboards.

1 Introduction

As LLMs shift roles from mere statistical models of language to fairly general-purpose AI agents, the inadequacy of existing evaluations—even those introduced within the past year—has become clear. Leading models routinely score over 90% on many evaluations (OpenAI, 2023). LLMs continue to struggle with evaluations involving quantitative, scientific, and academic reasoning, thus making these evaluations popular for leaderboards. As LLMs get better, one could also go to harder exam questions from higher academic levels. But soon it will become questionable if, say, PhD-level physics knowledge should be a good measure of general-purpose intelligence. (All authors of this paper would fail!)

A more pertinent issue that plagues evaluations derived from human tests is *training-set contamination*, i.e., examples very similar to the ones on the evaluation ending up in the training corpus of the model (OpenAI, 2023; Li, 2023). This is hard to detect given the size of training corpora and the fact that they are unreleased (including for some public models such as LLaMA-2 (Touvron et al., 2023)). The contamination issue especially bedevils evaluations based upon human exams (whose difficulty is tied to being time-limited and closed-book) since models are now being trained on technical textbooks as

Submitted to the Workshop on Distribution Shifts (DistShift@NeurIPS 2023). Do not distribute.

well as course materials.[1] A variant of the contamination issue is “cramming for the leaderboard". It is possible to deliberately train a model on data similar to those used in the leaderboard evaluations. Copycat datasets are easy to generate from a small number of examples using existing strong models. If “cramming” happens during pre-training, it becomes hard to detect. If it happens during fine-tuning, it may be detectable if it ends up harming general-purpose language skills.

Yet another issue arising from the secrecy of the training corpus is that it is difficult to verify how original the model’s text productions truly are. In a recent interview (Hinton & Ng), Hinton suggested that a significant hurdle in current discussion of AI risk is absence of agreement among experts on whether or not models have already gone past “stochastic parrots” behavior (Bender et al., 2021)—i.e., whether they are able to actually understand the world, or at a minimum produce novel thoughts or reasoning that they did not see in the training corpus.

Those issues can be naturally addressed by proposing evaluations carrying some type of *distribution shift*, making the test prompts out-of-distribution of the training corpus, but still reasonably measuring the capability of LLMs as general-purpose AI agents.

Desiderata for next-generation evaluations To sum up, we need evaluations that are: (a) clearly relevant to general-purpose intelligence and language understanding; (b) easy for humans to design and administer, including with academic-level resources; (c) resistant to training-set contamination and “cramming for the leaderboard;” (d) capable of revealing novelty along the lines sought by Hinton (Hinton & Ng); (e) easy to grade at scale (while allowing human spot-checking); (f) easily upgradeable into harder evaluations in the future as models get stronger; and (g) comprehensible at some level for the general public, including with respect to points (c) and (d).

1.1 SKILL-MIX

Our SKILL-MIX evaluation tests the model’s ability to produce sensible text satisfying natural constraints. It starts with a set of N skills (an example is: “hyperbole”) that every LLM could reasonably be expected to have encountered in its training set —say, because each skill has a Wikipedia entry. SKILL-MIX also uses a list of T topics that have low, but non-negligible, probability in any reasonable training corpus –e.g., *“sushi,” “ballroom dancing.”* SKILL-MIX (k) consists of randomly picking a subset of k skills out of N, and a random topic out of T topics. The chat agent is then asked to produce a short piece of text (say, 2 sentences) in the context of the selected topic and illustrate all k selected skills. This evaluation can be easily administered using any set of skills and questions and any k. It focuses on *highly constrained* text generation, whose difficulty intuitively increases with k[2]. Grading our evaluation is humanly possible with modest budgets. For convenience, we chose to grade using GPT-4 (OpenAI, 2023)) and LLaMA-2-70B-Chat (Touvron et al., 2023), after which the authors spot-checked the grading.

Resistance to dataset contamination It feels intuitive that the hardness of SKILL-MIX increases with k. A simple calculation supports this intuition. Given a list of N skills, there are $\binom{N}{k}$ ways to choose the subset of k skills. For $N = 1000$, this quantity exceeds 10^{10} when $k = 4$, and 10^{12} when $k = 5$. Furthermore, we selected the topics to be fairly rare —e.g., the word “sushi” has a unigram probability of 10^{-7} in Google n-grams (Google, 2012). Thus, the chance that the training corpus contains a short piece of text on the chosen topic that exhibits the chosen set of k skills becomes very small, even for $k = 4$. Furthermore, it is also rare *a priori* to find fragments of text on the web consisting of just two sentences that condense four linguistic skills in them. All this math, even if rough, suggests some resistance to training-set contamination — it at least rules out with high probability that the training set contained an *exact* solution to the question. Interestingly, SKILL-MIX with $k = 4$ proved hard (but solvable) for the authors. GPT-4 performed strongly on $k = 4$ and reasonably on $k = 5$. Many weaker models struggled even for $k = 3$.

[1]Studies have revealed significant anomalies on GPT-4’s performance on exam questions (Narayan & Kapoor).

[2]The authors of this paper took an average of more than 7 minutes to answer SKILL-MIX (4). Imagine having to write two sentences on the topic “sushi” that demonstrate hyperbole, equivocation, ad hominem attack, and fallacy of division. One has to comb through imaginary situations involving sushi, short enough to describe in a couple of sentences, and yet involving the four skills. Sometimes it is easier to first use 3 or 4 sentences and then shorten it. GPT-4’s answer: *“Your dislike for this award-winning sushi spot proves you have no palate, so your criticism is meaningless. After all, if the restaurant is world-class, then this California roll must be a culinary masterpiece, elevating my life to unparalleled heights of flavor.”* Note that GPT-4 left out equivocation, which it was able to rectify when prompted.

Table 1: **Performance of various instruction-tuned student (generating) models on SKILL-MIX (k) graded by GPT-4.** Ratio of Full Marks/Ratio of All Skills/Skill Fraction are reported for each student model at $k = 2, 3, 4$. Evaluations on $k = 5, 6$ are skipped if the Ratio of Full Marks drops below 0.2 with smaller k. Details on prompts can be found in Appendix C, D, and F. See Table 4 for additional metrics.

Student (generator)	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$
LLaMA-2-7B-Chat	.03/.10/.40	.00/.00/.27	.00/.00/.27	-/-/-	-/-/-
LLaMA-2-13B-Chat	.20/.40/.42	.03/.03/.34	.03/.03/.44	-/-/-	-/-/-
LLaMA-2-70B-Chat	.30/.30/.63	.03/.07/.44	.00/.03/.45	-/-/-	-/-/-
GPT-3.5-turbo	.60/.67/.78	.20/.23/.56	.07/.10/.58	-/-/-	-/-/-
GPT-4	.90/.93/.95	.73/.73/.91	.43/.43/.86	.30/.33/.84	.17/.17/.82
Mistral-7B-Instruct-v0.1	.10/.33/.35	.00/.03/.38	.00/.07/.24	-/-/-	-/-/-
Qwen-14B-Chat	.27/.30/.50	.03/.10/.40	.00/.00/.40	-/-/-	-/-/-
Xwin-LM-70B-V0.1	.43/.53/.68	.23/.40/.66	.20/.37/.60	.03/.13/.56	-/-/-
Falcon-180B-Chat	.27/.33/.53	.00/.03/.44	.03/.07/.38	-/-/-	-/-/-

Comparison with evaluations with natural prompts Popular evaluations (Li et al., 2023; Zheng et al., 2023; Zellers et al., 2019a; Mihaylov et al., 2018) for LLM usually consist of prompts extracted or bootstrapped from online or human-written corpus. As an example, AlpacaEval (Li et al., 2023) is a popular evaluation (with an accompanying leaderboard) for text generation. Similar to SKILL-MIX , it uses GPT-4 for evaluation, checking how often the model's generated output "wins" against that of DaVinci003 on a dataset of prompts. AlpacaEval presents a good case study on the difficulties of creating good evaluations by GPT-4. The primary difference between AlpacaEval and SKILL-MIX is that AlpacaEval is on a natural distribution of prompts. However, the natural distribution has a long tail, which is not well-represented in datasets of tens of thousands of examples, and whose structure is probably due to compositionality. The prompts of SKILL-MIX are constrained by asking LLM to output a combination of skills, creating a distribution shift to the natural prompts. The number of prompts in SKILL-MIX can be huge as mentioned before, and they are all arguably natural, allowing SKILL-MIX to capture the untested compositional structure of natural language.

Designed to give small open models a fighting chance in early 2023, AlpacaEval shows signs of saturation a mere 6 months later. Even 13B models now win against DaVinci003 with probability exceeding 90%, and recently Xwin-LM-70B-V0.1 (built on the LLaMA-2 base models) climbed to the top, pushing GPT-4, by a hair, to second place. Have LLM capabilities truly progressed to this extent within 6 months? We find that the new champion scores well on SKILL-MIX , and noticeably better than LLaMA-2-70B-Chat, but it is handily beaten by GPT-4. In general, most models now get similar scores on AlpacaEval, but show widely varying performances on SKILL-MIX , which suggests that AlpacaEval has lost its discriminative ability. SKILL-MIX avoids the shortcomings of AlpacaEval by evaluating on constrained text generation, and using k to adjust the difficulty. For further discussion on SKILL-MIX with respect to prior work, see Appendix A.

Evidence of cramming for the leaderboard Hugging Face's Open LLM leaderboard (Beeching et al., 2023), which is based upon EleutherAI's evaluation harness (Gao et al., 2021), is seen as a proving ground for open LLMs. Many models currently at the top of the leaderboard are LLaMA-2 derivatives, and are ranked much higher than the corresponding LLaMA-2 model. However, these models perform poorly on SKILL-MIX and worse than LLaMA-2-70B-Chat, suggestive of cramming that wiped away general-purpose text skills they had started with (see Section 2). The recent Falcon-180B-Chat (Almazrouei et al., 2023) also places higher on the leaderboard than LLaMA-2-70B-Chat, and has been claimed to have capabilities between GPT-3.5-turbo and GPT-4 (OpenAI, 2023) based upon this ranking. Yet, it fares worse than LLaMA-2-70B-Chat on SKILL-MIX .

2 Experimental Results

In this section, we test various instruction-tuned models on their performance on SKILL-MIX for various k. (For details regarding our procedure, please see Appendix C and D.) For convenience, we use *saturation point* to denote the value of k at which a model's score in SKILL-MIX drops off. From the experimental results, we answer the following questions: (1) What is the effect of increasing k on SKILL-MIX performance? (2) What is the relationship between model scale and saturation point, especially among LLaMA-2 models?

For each SKILL-MIX (k), we evaluate all models on 30 (k skills, 1 topic) combinations. We provide each specific combination of k skills to the (Student) model on three instances (see Figure 2). Each of the three generated texts is also graded three times (to reduce the randomness caused by the Grader), in total creating nine grading results for each (k skills, 1 topic) (see Figure 3).

Metrics Each generated text can receive up to $k + 3$ points: 1 point for each correctly illustrated skill, 1 point for sticking to the topic, 1 point for coherence / making sense, and 1 point for having at most $k - 1$ sentence. The points are then converted into various metrics of interest (Appendix G), including: (1) *Ratio*[3] *of Full Marks*: 1 if all $k + 3$ points are earned; (2) *Ratio of All Skills*: 1 if k points are awarded for skills and ≥ 2 points for remaining criteria; and (3) *Skill Fraction*: $\frac{c}{k}$ if c points awarded for k skills and 3 points for remaining criteria. We then take the maximum value of the metrics among the 3 generations for a given (k skill, 1 topic) combination, and average the maximum value across the 30 combinations. The results graded by GPT-4 are shown in Table 1.

Increasing k degrades SKILL-MIX performance We observe that the ratio of full marks and the ratio of all skills can decrease dramatically when k increases. With the exception of GPT-4, GPT-3.5-turbo and Xwin-LM-70B-V0.1 (Xwin-LM Team, 2023), all models saturate on or before $k = 3$ with GPT-4 grading. Amongst the small models, LLaMA-2-7B-Chat and Mistral-7B-Instruct-v0.1 (Mistral AI Team, 2023) saturate at $k = 2$. As reported in Table 5, LLaMA-2 grading is more generous, and the saturation point is delayed by 1.

Relationship between model scale and saturation point We find that as capacity increases on LLaMA-2, so does the saturation point. Observe that for LLaMA-2-7B-Chat, LLaMA-2-13B-Chat, LLaMA-2-70B-Chat, the saturation points (of GPT-4 grading) are $k = 2$, 3, and 3, respectively. Additionally, for any fixed k and metric type, higher model capacity corresponds to a better score amongst the LLaMA-2 model family. However, these observations do not necessarily hold true for models from different families. For example, Falcon-180B-Chat has more model parameters than Xwin-LM-70B-V0.1, yet the saturation point of Xwin-LM-70B-V0.1 is higher than that of Falcon-180B-Chat, and Xwin-LM-70B-V0.1 also outperforms Falcon-180B-Chat across all metrics for $k = 2, 3, 4$.

A deviation from model rankings on popular LLM leaderboards Recent models (i.e., Falcon-180B-Chat, Xwin-LM-70B-V0.1, Qwen-14B-Chat (Bai et al., 2023), Mistral-7B-Instruct-v0.1) are often introduced with their performance evaluated on AlpacaEval or Hugging Face's Open LLM Leaderboard (which contains ARC (Clark et al., 2018), HellaSwag (Zellers et al., 2019b), MMLU (Hendrycks et al., 2020), and TruthfulQA (Lin et al., 2021)), along with a comparison to the LLaMA-2 and GPT families. We find that their superior performance on those evaluations may not extend to SKILL-MIX (See Appendix G for more details and discussion). Differences between model ranking on popular LLM leaderboards vs. SKILL-MIX provide evidence of "cramming for the leaderboard", further validating that SKILL-MIX is a good evaluation benchmark.

3 Conclusions and Takeaways

SKILL-MIX is an attempt to evaluate general-purpose language capabilities, specifically, a particular sort of compositional generalization. It tests the model's ability to create text on-demand on a given topic and with a given combination of well-known skills. A key idea is that the skills and the topic are chosen randomly from a big list, which itself can be expanded in many ways to give new evaluations.

Section 2 showed that the performance of proprietary models on SKILL-MIX (k) generally accords with popular perceptions of their quality. In line with Arora & Goyal (2023), the results also show that when created by competent teams, larger models achieve a higher saturation point than smaller models (e.g., the three LLaMA-2 models) The disappointing performance of Falcon-180B-Chat was an exception. Several open models show signs of being over-trained for leaderboards at the expense of general-purpose language capabilities. Since current leaderboards show signs of losing their discriminative power, in Appendix H, we sketch a vision for an ecosystem of independent SKILL-MIX evaluations specializing in different sets of skills and topics that could provide trusted estimates of model capabilities. Soon, all proprietary models will be multi-modal, and more powerful. We hope to design a multi-modal version of SKILL-MIX for them.

[3] This is called "ratio" because the metric is later averaged over the 30 combinations, even though this metric is 0 and 1 for a single generation.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Maitha Alhammadi, Mazzotta Daniele, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. The falcon series of language models: Towards open frontier models. 2023.

Lorin W. Anderson and David R. Krathwohl (eds.). *A Taxonomy for Learning, Teaching, and Assessing. A Revision of Bloom's Taxonomy of Educational Objectives*. Allyn & Bacon, New York, 2 edition, December 2001. ISBN 978-0801319037.

Sanjeev Arora and Anirudh Goyal. A theory for emergence of complex skills in language models. *arXiv preprint arXiv:2307.15936*, 2023.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. 2023.

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. Open llm leaderboard. `https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard`, 2023.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

B. S. Bloom, M. B. Engelhart, E. J. Furst, W. H. Hill, and D. R. Krathwohl. *Taxonomy of educational objectives. The classification of educational goals. Handbook 1: Cognitive domain*. Longmans Green, New York, 1956.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Mayee F Chen, Nicholas Roberts, Kush Bhatia, Jue Wang, Ce Zhang, Frederic Sala, and Christopher Ré. Skill-it! a data-driven skills framework for understanding and training language models. *arXiv e-prints*, pp. arXiv–2307, 2023.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Louise Cummings. Pragmatics: a multidisciplinary perspective. In *Edinburgh University Press*, 2005.

Mary Forehand. *Emerging perspectives on learning, teaching, and technology.*, chapter Bloom's taxonomy: Original and revised. 2005.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, September 2021. URL `https://doi.org/10.5281/zenodo.5371628`.

Google. Google ngram viewer. http://books.google.com/ngrams/datasets, 2012. URL `http://books.google.com/ngrams/datasets`.

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.

G Hinton and A Ng. A conversation about ai risk. URL `https://twitter.com/AndrewYNg/status/1667920020587020290?lang=en`.

David Kelly. The art of reasoning. an introduction to critical and logical thinking. In *WW Norton & Company Inc*, 2014.

Kenneth R Koedinger, Elizabeth A McLaughlin, and John C Stamper. Automated student model improvement. *International Educational Data Mining Society*, 2012.

Kenneth R Koedinger, Paulo F Carvalho, Ran Liu, and Elizabeth A McLaughlin. An astonishing regularity in student learning rate. *Proceedings of the National Academy of Sciences*, 120(13): e2221311120, 2023.

Nan Li, Eliane Stampfer, William Cohen, and Kenneth Koedinger. General and efficient cognitive model discovery using a simulated student. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35, 2013.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`, 2023.

Yucheng Li. Estimating contamination via perplexity: Quantifying memorisation in language model evaluation, 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering, 2018.

Mistral AI Team. Mistral 7b, 9 2023. URL `https://mistral.ai/news/announcing-mistral-7b/`.

Erik T Mueller. *Commonsense reasoning: an event calculus based approach*. Morgan Kaufmann, 2014.

Arvind Narayan and Sayash Kapoor. Gpt-4 and professional benchmarks: the wrong answer to the wrong question. URL `https://www.aisnakeoil.com/p/gpt-4-and-professional-benchmarks`.

Shaun Nichols and Stephen P Stich. Mindreading. an integrated account of pretence, self-awareness, and understanding other's minds. In *Oxford University Press*, 2003.

OpenAI. Gpt-4 technical report, 2023.

Karl Pertsch, Youngwoon Lee, and Joseph Lim. Accelerating reinforcement learning with learned skill priors. In *Conference on robot learning*, pp. 188–204. PMLR, 2021.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. Large language models are not zero-shot communicators. *NeuRIPS*, 2023.

Philipp Schmid, Omar Sanseviero, Pedro Cuenca, Leandro von Werra, and Julien Launay. Spread your wings: Falcon 180b is here, 9 2023. URL `https://huggingface.co/blog/falcon-180b`.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Xwin-LM Team. Xwin-lm, 9 2023. URL `https://github.com/Xwin-LM/Xwin-LM`.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *CoRR*, abs/1905.07830, 2019a. URL `http://arxiv.org/abs/1905.07830`.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence?, 2019b.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

A Prior work

Arora & Goyal (2023) gives a mathematical model for skill emergence via LLM scaling. (In principle it applies to non-text data as well, since it makes almost no assumptions about what "text" or "skills" are.) The key assumption is that pieces of text involve random combinations of skills, and then reductions in cross-entropy (which can happen on an arbitrary subset of text pieces) are shown to imply improvements in both individual skills and combinations of skills. A key implication is that the trained model may show competence on k-tuples of skills even though this k-tuple of skills was not demonstrated in the training set. (Note that individual skills were demonstrated, as were some random k-tuples, but most k-tuples were not demonstrated.) It is important to note that the theory did not need to specify what it means to "combine" skills.

Skill-It The goal of Skill-It (Chen et al., 2023) is to select an optimal subset of the training data such that an LLM trained on this subset will perform well on downstream tasks. Chen et al. (2023) utilize the notion of *skill ordering* to construct this subset. Skill ordering refers to the natural notion that learning "simpler" skills first can make learning "difficult" skills later easier for the learner. Chen et al. (2023) define an *ordered skill set* as "a collection of skills with a directed skills graph that is neither complete nor empty, where an edge from a prerequisite skill to a skill exists if the amount of training it takes to learn the skill can be reduced if the prerequisite skill is also learned."

Skills in Reinforcement Learning Reinforcement learning also has a notion of "skills," which is distinct from the notion that we use (Arora & Goyal, 2023), but bears some similarities to the notion of skill used in Chen et al. (2023). In particular, Pertsch et al. (2021) aims to learn a *skill prior*, i.e., a distribution over skills, such that a model trained to develop these skills will later perform well on downstream tasks.

Prior evaluations Prior work has emerged which evaluates LLMs on particular skills, in a different sense than we do. Ruis et al. (2023) finds that LLMs do not do well on implicature.

Pedagogy work Another important area where skills have been previously studied is that of human skill learning in pedagogy. Koedinger et al. (2023) develops a cognitive and statistical model of skill acquisition with the goal of understanding why/if some students learn faster than others. Koedinger et al. (2012) presents an algorithm to discover cognitive models, which are essentially skill models. Li et al. (2013) uses a computational model of student learning (a simulated student) in order to discover cognitive models (i.e., "learn the skills").

B List of Skills and Topics

B.1 Picking Skills

We obtained a set of 101 language skills and a set of 100 topics for SKILL-MIX evaluation. Since the goal of our evaluation is to test general-purpose text generation capability rather than the ability of the particular skills and topics, we only release 10 skills and 10 topics randomly sampled from the two sets to avoid potential "cramming" for SKILL-MIX . The randomly sampled skills and topics appear in (see Tables 2 and 3).

We curated the topic list by first sampling a large list of topics (e.g. using Reddit forums as inspiration), and then narrowing the list down to 100 topics based on the unigram frequency of the topic (and all related synonyms) on Google Ngram viewer (Google, 2012). To earn a spot on our list, a topic's average unigram had to be around 10^{-6} —low (further reducing the likelihood that the model had seen the k skills demonstrated in the context of the topic), and yet still ensuring good coverage even in 100B-sized corpora. The list of basic skills was designed to contain language skills which have a Wikipedia entry or listing (and thus known to every LLM), and whose definition the average person could understand. We started with a longer list of skills taken from textbooks on logical reasoning, rhetoric, theory of mind, (Kelly., 2014; Cummings, 2005; Nichols & Stich, 2003; Mueller, 2014). We tried to eliminate skills that either were too specialized —and thus difficult to apply in the context of the fairly narrow topics already chosen— or difficult to combine with other skills. (Some examples of discarded skills appear in the Appendix B.2.) For each skill, we created a description and an illustrative example of its usage —these were taken from either a textbook or Wikipedia, though occasionally, we modified them to make them clearer or more concise.

B.2 Skill choosing process

The list of ≈ 100 skills used to draw tuples of k skills was manually curated, and designed to include skills the average person could understand which were common enough that they would appear on Wikipedia (and hence in the model's pre-training data). We pick skills from standard textbooks on logic (Kelly., 2014), pragmatics (Cummings, 2005) and theory of mind (Nichols & Stich, 2003). We also pick literary skills from Wikipedia. Not all skills were considered suitable for our dataset. Because we want to evaluate the ability of models to compose skills, we eliminate skills that are trivially present in almost any piece of text. Below is an example of a skill that was eliminated from our dataset because it is so common in the English language that the model may "accidentally" generate it, thereby falsely appearing as though it is able to combine this skill with other skills.

```
Skill: Pied Piping

Definition: A syntax phenomenon whereby a given expression brings along an accompanying phrase when
it is moved.

Ex: She bought ''that house''. ''Which house'' did she buy? The preceding is an example of pied piping
because the word ''Which'' brings along the word ''house'' in the ''Wh...'' clause.
```

For similar reasons, we also eliminate skills that inherently compose poorly with other skills. An example is given below:

```
Skill: situational irony

Ex: A firehouse burning down is situational irony, as one would not expect a place that puts out
fires to burn.
```

Finally, some skills were not included due to being hard, even for a human grader, to grade whether the skill was present.

Definitions for skills vary between different sources and textbooks. Since the model was pre-trained on Wikipedia, we prefer the Wikipedia definition over other sources when available.

Skill examples were scraped from the internet, but chosen by human evaluators to be short and unambiguous.

B.3 List of Skills and Topics

Here we release a random sample of 10% of the skill list and topic list. We do not release the full lists to avoid potential "cramming" for SKILL-MIX .

Table 2: 10 randomly sampled skills from the 101 skills we used in SKILL-MIX evaluation.

Category	**Skill**	**Definition**	**Example**
reasoning	self serving bias	A cognitive or perceptual process that is distorted by the need to maintain and enhance one's self esteem.	"If I do well on the exam, it's because of my academic prowess and hard work. If I do poorly, it's because the course was poorly taught, and the exam was poorly proctored."
rhetorical	accident (fallacy)	an informal fallacy and a deductively valid but unsound argument occurring in a statistical syllogism (an argument based on a generalization) when an exception to a rule of thumb is ignored.	Cutting people with knives is a crime. Surgeons cut people with knives. Surgeons are criminals.

rhetorical	complex question (loaded question with implicit assumption)	A question that is loaded with an implicit assumption.	“Why are you lying to me?” is a question that presupposes you are lying to me. Any answer you give will force you to agree you are lying.
rhetorical	red herring	Introducing irrelevant points to detract attention from a question.	A member of the press asks the president why they voted to expand a welfare program. The president responds, “The strength of America is the strength of its communities, and I am proud to make our communities better places.”
literary	metaphor	a figure of speech that, for rhetorical effect, directly refers to one thing by mentioning another.	“All the world’s a stage, And all the men and women merely players” is a metaphor because it’s a comparison without using “like” or “as.”
logical	spatial reasoning	The capacity to reason about the spatial relationships between objects.	The key fit into the box. Using spatial reasoning, one can deduce that the width of the key was smaller than the width of the box.
logical	modus ponens	A syllogism that is of the form “If P then Q. P. Hence Q.”	“If today is Tuesday, then John will go to work. Today is Tuesday. Therefore, John will go to work.”
logical	statistical syllogism	A syllogism that argues, using inductive reasoning, from a generalization true for the most part to a particular case.	“Almost all people are taller than 26 inches. Gareth is a person. Therefore, Gareth is taller than 26 inches.”
theory of the mind	emotional self regulation	a complex process that involves initiating, inhibiting, or modulating one’s state or behavior in a given situation.	Examples of emotional self regulation include meditating, pausing to collect oneself before speaking, and practicing stress management.
physical knowledge	folk physics (common knowledge physics)	The untrained human perception of basic physical phenomena.	“If I roll the pen off of the table, it will fall to the floor.”

C Design of SKILL-MIX

C.1 Picking the skills

We obtained a set of 101 language skills and a set of 100 topics for SKILL-MIX evaluation. Since the goal of our evaluation is to test general-purpose text generation capability rather than the ability of the particular skills and topics, we only release 10 skills and 10 topics randomly sampled from the two sets to avoid potential “cramming” for SKILL-MIX . The randomly sampled skills and topics appear in Appendix B (see Tables 2 and 3).

We curated the topic list by first sampling a large list of topics (e.g. using Reddit forums as inspiration), and then narrowing the list down to 100 topics based on the unigram frequency of the topic (and all related synonyms) on Google Ngram viewer (Google, 2012). To earn a spot on our list, a topic’s average unigram had to be around 10^{-6} —low (further reducing the likelihood that the model had

Table 3: 10 random sampled topics from 100 topics used in SKILL-MIX evaluation.

Topic
Sewing
Dueling
The Ottoman Empire
Triathlons
Beekeeping
Survivalism
Guerilla warfare
Gardening
Knots
Urbanism

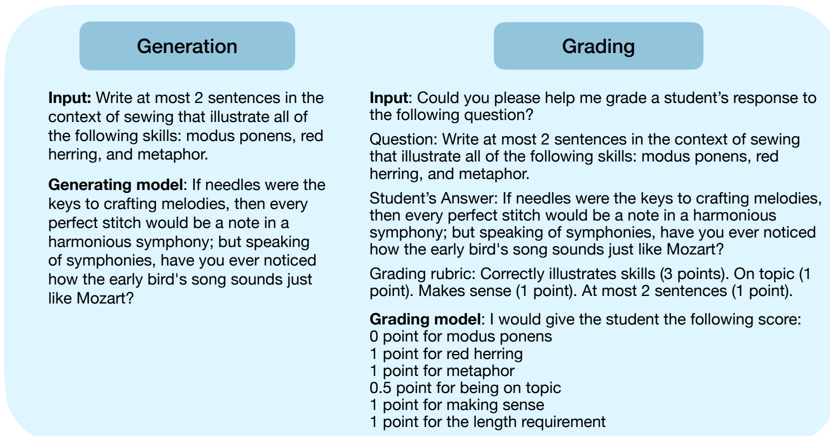

Figure 1: **Left: Simplified depiction (with simplified prompt) of the generation stage of our evaluation.** The full prompt appears in Appendix F.1.1. The generating model is given a topic (sewing) as well as skills (modus ponens, red hearing, metaphor) and asked to generate text demonstrating the skills. The full prompt contains skill definitions and examples, which can be found in Appendix Table 2. **Right: Simplified depiction (with simplified prompt) of the grading stage of our evaluation.** The grading model (not necessarily the same as the generating model) is given the generating model output and grading instructions, and returns pointwise grading. The full grading prompt can be found in Appendix F.1.2.

seen the k skills demonstrated in the context of the topic), and yet still ensuring good coverage even in 100B-sized corpora. The list of basic skills was designed to contain language skills which have a Wikipedia entry or listing (and thus known to every LLM), and whose definition the average person could understand. We started with a longer list of skills taken from textbooks on logical reasoning, rhetoric, theory of mind, (Kelly., 2014; Cummings, 2005; Nichols & Stich, 2003; Mueller, 2014). We tried to eliminate skills that either were too specialized —and thus difficult to apply in the context of the fairly narrow topics already chosen— or difficult to combine with other skills. (Some examples of discarded skills appear in the Appendix B.2.) For each skill, we created a description and an illustrative example of its usage —these were taken from either a textbook or Wikipedia, though occasionally, we modified them to make them clearer or more concise.

C.2 Procedure

Our evaluation is roughly broken down into two parts (see Figure 2). In the first part, we conduct **generation**, where a (Student) language model is given a set of k skills and a topic, and asked to generate some text demonstrating the k skills in the context of the provided topic. Once the (Student) language model generates some text, it then must be **graded** by a (possibly different) grading language model (i.e., Grader). A simplified version of the prompts used in the generation and grading stages is depicted in Figure 1.

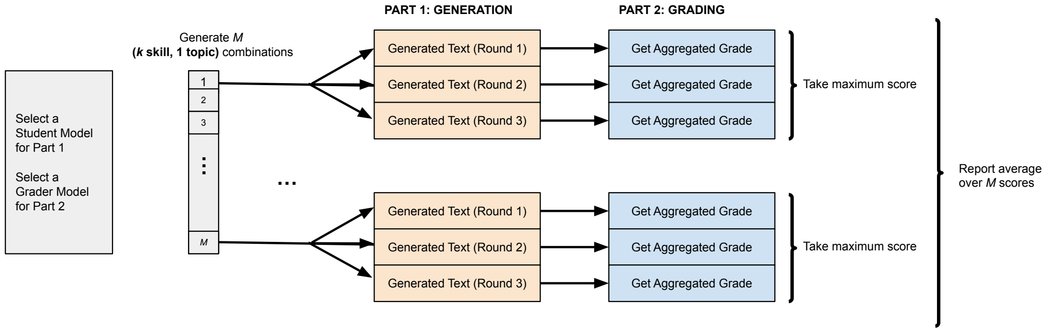

Figure 2: **Illustration of** SKILL-MIX (k) **pipeline.** In our experiments, we use $M = 30$. For a more detailed illustration of grading a single piece of generated text, see Figure 3.

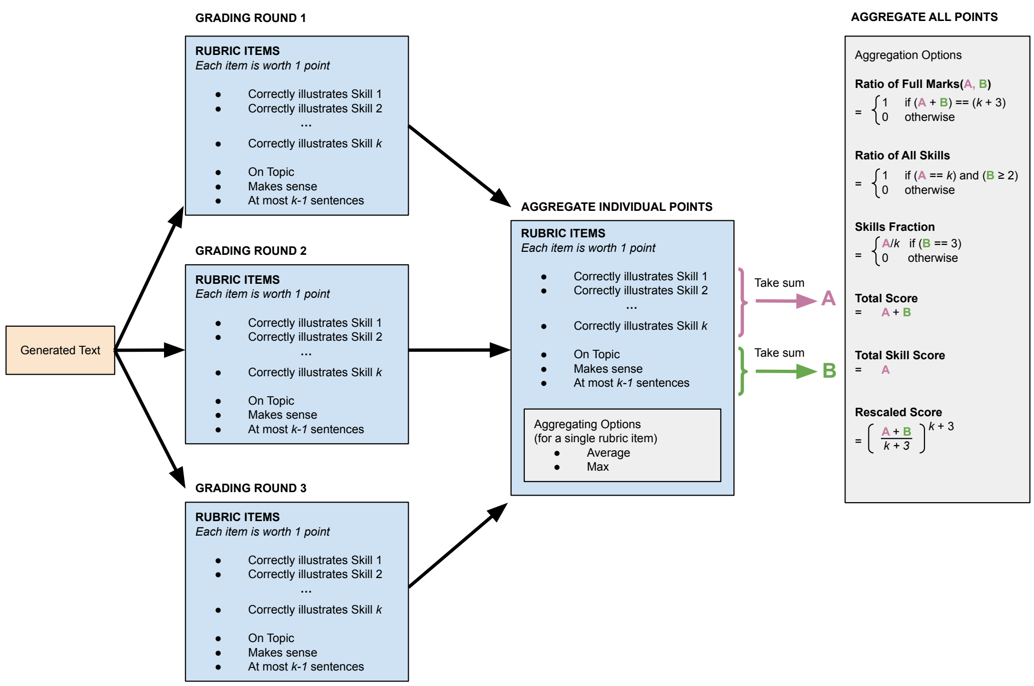

Figure 3: **Illustration of obtaining aggregated grade** This illustration depicts the process used to grade a single generated piece of text.

Models used for generation Our dataset is designed to test general skills, and hence many language models may be used in the generation step. However, since the language model must be able to respond to prompt instructions ("generate k skills"), we only pick models that have been instruction-tuned. These models include LLaMA-2-7B-Chat, LLaMA-2-13B-Chat, LLaMA-2-70B-Chat (Touvron et al., 2023), GPT-3.5-turbo, GPT-4 (OpenAI, 2023), Falcon-180B-Chat (Almazrouei et al., 2023), Xwin-LM-70B-V0.1 (Xwin-LM Team, 2023), Mistral-7B-Instruct-v0.1 (Mistral AI Team, 2023), Qwen-14B-Chat (Bai et al., 2023). Note Xwin-LM-70B-V0.1, Mistral-7B-Instruct-v0.1, and Qwen-14B-Chat were released only a few days before the deadline, and all of them claim to perform better on benchmarks compared to state-of-the-art models of even larger size.

Models used for grading We find that some language models are more suitable for grading than others. Some models have difficulty recognizing the presence of skills, even when they are demonstrated correctly. We use LLaMA-2-70B-Chat and GPT-4 after manually spot-checking grading samples and ensuring they aligned with human grading.

Details of model configurations We do not use quantization on any of the models. For the GPT family, we use OpenAI API with default generation configuration and the minimal system prompt "`You are a helpful assistant.`" For the LLaMA-2 family, we use 2 A100 GPU and run with no system prompt, 0.7 temperature, 1.0 repetition penalty, and 512 max new tokens. For Falcon-180B-Chat, we use the prompt format mentioned in the official Huggingface blog of Falcon-180B-Chat (Schmid et al., 2023) and the same parameters as LLaMA-2 family. For Xwin-LM-70B-V0.1, we use the official prompt format (Xwin-LM Team, 2023) and the same hyperparameters as those used for the LLaMA-2 family. For Mistral-7B-Instruct-v0.1, we access the prompt format with `tokenizer.apply_chat_template` function and again the same parameters as the LLaMA-2 family. For Qwen-14B-Chat, we directly use the `model.chat` function as mentioned in their official Github repository.

SKILL-MIX (k) consists of picking a random topic and a random subset of k skills (with their respective definitions and illustrative examples) from the list described above, and asking the (Student) model to produce a short piece of text which illustrates all the k skills in the context of the topic of interest. A simplified example prompt appears in Figure 1, but our final prompt includes an additional question that asks the (Student) model to look over and possibly improve its first answer. We find the second answer can be much better than the first one with proper prompt engineering. In addition, we ask the model to separate its answer and explanation with "Answer" and "Explanation." Otherwise, the model may not separate its generated answer from its explanation, hampering the parsing of the answer. We provide more details about the generation prompts in Appendix F.1.1

The authors took the test (6 questions each with $k = 4$) to assess the feasibility and difficulty level. The average time taken to understand the prompt and type the answer was more than 7 minutes. This is not an easy test for humans!

D Auto-grading Method

For each prompt sampled from SKILL-MIX (k), the (Student) model's corresponding answer is graded according to the following criteria: (1) the k skills are present and used properly in the output; (2) the output is on the given topic; (3) the number of sentences in the output is within the provided limit, which we set to be $k - 1$; and (4) the output is a piece of sensible text. For any of the above subtasks, partial credit can be assigned if the answer partially satisfies the requirement.

The generations by the models were graded using GPT-4 as well as LLaMA-2-70B-Chat, and these grades were then spot-checked by the paper authors. In the trial run, we focused on tweaking the method for best results and consistency using a small set of around 20 generations. As usual, the assessments generated by the grading models (especially LLaMA-2) were somewhat sensitive to the prompt. We tweaked the grading prompt by including a summarized version of the generation prompt, providing definitions and illustrative examples of the individual k skills, and requesting for the graded output to follow a particular format.

While both models are creditable graders, they, like all current LLMs, were unreliable at simple arithmetic, which is important for calculating a total score. We changed the prompt to require the grader to output separate grades for individual components (proper use of skills, good fit to the topic, and producing sensible text), which were subsequently aggregated (see Figure 3) using a separate but simple Python script[4].

To require separate grades for individual components, we asked GPT-4 to provide a rubric-table style grade, whereas for for LLaMA-2, we simply asked the model to include "Point earned: 1" if the requirement is met, and "Point earned: 0" otherwise, for each rubric item in the evaluation. More details, as well as the full grading prompt, appear in Appendix F.1.2.

Human Grading: Better? With a small test we conducted with five NLP researchers, we found that human grading is noisy, and human graders might need significant training so they can agree on a grading rubric. The standard deviation between the human grading is high, and even higher than the

[4]We adjusted the point awarded for meeting the number of sentences requirement based on the ground truth. While adjustments were rare, they were more common for LLaMA-2 than GPT-4.

difference between the average human grading and GPT-4 grading.[5] This also indicates that GPT-4 and LLaMA-2 graders are reasonable graders compared to humans. More details of the test can be found in Appendix E.

E Human Grading Test

We will now describe a test we conducted to measure the grading quality of human graders. Our five volunteers were Ph.D. students and Postdocs working in the field of natural language processing and large language models. They were given 5 outputs generated by GPT-4 on SKILL-MIX (4). The same grading prompt for machine grading was given to human graders, asking them to give each individual point for the criteria. In this case ($k = 4$), each output could receive at most 7 points.

For each point, we computed the mean and standard deviation among the human graders and then averaged over all the points (35 in total). The average standard deviation is 0.261. On the other hand, we compare the average of human grading with GPT-4 and LLaMA-2 grading. For each point, we computed the absolute difference between the mean among the humans and machine grading. Then, we took the average over 35 points. We found that the average difference is 0.257 for GPT-4 grading and 0.268 for LLaMA-2 grading. This means if we assume the average of human grading is the ground truth, then human graders and machine graders have similar errors.

We also observe that human graders in general give lower scores for text making sense, probably because the output is usually a shortened version of the model's first attempt. In contrast, the machine graders usually award the point for "making sense" for more than 90% of the generated answers.

F Prompt Design

Our evaluation is roughly broken down into two parts. In the first part, we conduct **generation**, where a language model is given a set of k skills and a topic and asked to generate some text demonstrating the k skills. Once some text has been generated, it then must be **graded** by a (possibly different) language model.

F.1 Prompt design

We experiment with different prompts for both generation and grading. We find that prompts can perform quite differently. We list some examples of prompts we considered below.

F.1.1 Generation prompts

We try several different prompts when asking the models to generate tuples of k skills, and find that prompt selection does influence the quality of the generation. In general, our prompts contain two questions, giving the model the chance to look over its first answer and improve it. Without giving specific instructions about the format, we found it hard to parse the model output for grading because the model may not separate the generated answer from its explanation. Hence, we asked the model to separate its answer and explanation with "Answer:" and "Explanation:". Here is our first prompt with the formatting instructions:

```
Greetings!  I am interested in natural language processing and I was wondering if you could help me
generate an example of text that illustrates multiple skills in semantics or syntax.  The example
should be a single piece of text with {num_sentences_str} in the context of {topic} that illustrates
all of the following skills:  {skills_str}.

{skills_defs_and_examples}

Please keep the text short so it can fit in {num_sentences_str}, and please make sure the concepts
can be found fully from the text.  Please start the text with 'Answer:'  and start the explanation
with 'Explanation:'.  Thanks very much!
```

[5] The difference between the average human grading and LLaMA-2 grading is slightly higher than the standard deviation between humans.

```
Thanks very much. Now could you look it over and shorten your example but make sure it still
illustrates all the skills?
```

Using this prompt, we observe that the models do not always follow the instructions, especially for the second answer. We also observed that the second answer is sometimes worse than the first generation, partially because some of the skills are also removed when shortening the answer.

To overcome the shortages, we present our 8th attempt at the generation prompt below.

```
Greetings! I am interested in natural language processing and I was wondering if you could help me
generate an example of text that illustrates multiple skills in semantics or syntax. The example
should be a minimal natural piece of text with up to a few lines in the context of {topic} that
illustrates all of the following skills: {skills_str}. Please keep the text as short as possible,
and make sure the concepts can be found fully from the text.

For reference, here are the definitions and examples for the concepts:

{skills_defs_and_examples_simple}

Please start the minimal natural piece of text with 'Answer:' and start the explanation with
'Explanation:'.

Thanks very much!
```

```
Thanks very much. Could you please look over the minimal natural piece of text and possibly improve
and shorten it (up to {num_sentences_str})? If you make changes, please make sure that the text
still illustrates all skills and remains on topic.

Again, please start the improved minimal natural piece of text with 'Answer:' and start the
explanation with 'Explanation:'.

Thanks very much!
```

Although there still exists rare failure cases for some models (demonstrated in Appendix F.2.1), the instructions are mostly followed by all models with this version of the prompt. The quality of the second answer is also significantly better when we remove the length constraints on the first answer, and give the model the option of shortening the text (rather than the requirement) for the second answer.

In the later tuning of the prompt, we further emphasize the improvement of the second answer, and find the performance of GPT-4 continues to improve. The same prompt also improves Xwin-LM-70B-V0.1, but makes LLaMA-2 family perform slightly worse on $k > 2$ (in those cases, we use the previous prompt). Our final prompt is recorded below.

```
Greetings! I am interested in natural language processing and I was wondering if you could help me
generate an example of text that illustrates multiple skills in semantics or syntax. The example
should be a minimal natural piece of text with up to a few lines in the context of {topic} that
illustrates all of the following skills: {skills_str}. Please keep the text as short as possible,
and make sure the concepts can be found fully from the text.

For reference, here are the definitions and examples for the concepts:

{skills_defs_and_examples_simple}

Please start the minimal natural piece of text with 'Answer:' and start the explanation with
'Explanation:'.

Thanks very much!
```

```
Thanks very much. Could you please look over your answer and improve it? Please make sure that the
new answer better illustrates all skills and remains on topic. On top of that, please also make sure
the new answer only contains up to {num_sentences_str}.

Again, please start the improved minimal natural piece of text with 'Answer:' and start the
explanation with 'Explanation:'.

Thanks very much!
```

Example of LLaMA-2-70B-Chat generation response

Sure, here's an example of a minimal natural piece of text that illustrates both emotional self regulation and folk physics in the context of sewing:

Answer: "I'm so frustrated with this sewing project. I've been working on it for hours and it's still not turning out right. I just want to throw my sewing machine out the window!"

Explanation: This text illustrates emotional self regulation because it shows the speaker's ability to recognize and manage their emotions. The speaker is feeling frustrated and angry, but instead of acting on those emotions and doing something impulsive (like throwing the sewing machine out the window), they pause and express their feelings in a more controlled way. This shows that they are regulating their emotions in order to avoid doing something they might regret.

The text also illustrates folk physics because it implies a basic understanding of cause-and-effect and the behavior of objects in the physical world. The speaker's statement about throwing the sewing machine out the window implies that they understand that objects will fall to the ground if dropped, and that breaking the machine would be a consequence of their actions.

This demonstrates an untrained human perception of basic physical phenomena, which is a key aspect of folk physics.

Sure, here's an improved example that better illustrates both emotional self-regulation and folk physics in the context of sewing, and is only one sentence long:

Answer: "I'm taking a deep breath and pausing before ripping out this seam for the fifth time, because I know I'll just get frustrated and make more mistakes if I don't calm down."

Explanation: This text illustrates emotional self-regulation because it shows the speaker actively regulating their emotions by taking a deep breath and pausing before proceeding with a potentially frustrating task. It also illustrates folk physics because it implies an understanding of the cause-and-effect relationship between the speaker's emotions and their ability to perform the task accurately. The speaker recognizes that if they don't calm down, they are more likely to make mistakes, demonstrating an understanding of the impact of their emotions on their physical actions.

F.1.2 Grading prompts

We also try several different prompts when asking the model to grade a response. Unsurprisingly, we find that prompt selection once again does influence grading. In general, our prompts ask the (Grader) model to score the student's response according to the rubric criteria.

Below is the final prompt we provided to GPT-4 for grading.

Greetings! I was wondering if you could help me grade the following answer given by a student.

I'll first describe the question that was given to the student, and then give you the student's answer, and the grading rubric.

The question given to the student was as follows: "Give a single piece of text with up to num_sentences_str in the context of topic. This single piece of text should illustrate all of the following skills: skills_str."

The student's answer was: "student_answer"

For reference, here are the definitions for the skills: skills_defs_and_examples_simple

Using a rubric table format, please grade the student's answer with positive scoring. Each criterion is worth 1 point. The criteria are: rubric_items. The table should only have the following columns: 'Criteria', 'Points Earned'. In the 'Points Earned' column, please provide only numerical values with no additional formatting. Please introduce the table with 'Here's the grading table:' and please include a row for 'Total Points Earned' at the end of the table. Finally, please start your grading explanation with 'Explanation':

Below is the final prompt we provided to LLaMA-2-70B-Chat for grading.

Greetings! I was wondering if you could help me grade the following answer given by a student.

```
I’ll first describe the question that was given to the student, and then give you the student’s
answer, and the grading rubric.

The question given to the student was as follows: ‘‘Give a single piece of text with up to
num_sentences_str in the context of topic. This single piece of text should illustrate all of
the following skills: skills_str.’’

The student’s answer was: ‘‘student_answer’’

For reference, here are the definitions for the skills: skills_defs

Please grade the student’s answer using the following criteria: rubric_items. For each item in the
rubric, please include in your evaluation ‘Point earned: 1’ if the requirement is met and ‘Point
earned: 0’ otherwise. Please finish your grading with ‘Grade:’ followed by a number indicating the
total points earned by the student. Thanks so much.
```

The following failure cases guided up towards our final grading prompt and manual-point-extraction scheme.

- The model would assign points for individual rubric criteria, but then total them up incorrectly
- We say we are grading a student’s answer. There is some amount of LLM sycophancy if we say it is a professor’s answer.
- Without definition in prompt, the model mistakes certain skills for the common English meanings.

We will now provide examples of these failure cases based on real outputs.

F.1.3 Unable to output correct total score

The following example is based on a real grading response by LLaMA-2-70B-Chat. Observe that the grader model assigns points for each of the rubric criteria, but does not correctly total them up (the correct total is 4, not 3). Because scenarios like these were common, we used a simple Python script to extract the individual points associated with each rubric item to calculate the final score.

```
Sure, I’d be happy to help you grade the student’s answer. Here’s my evaluation using the given
rubric:

1. Contains red herring: The student’s answer demonstrates red herring in the sentence, where Mrs.
Thompson diverts attention from the main issue at hand, so Point earned: 1. 2. Contains modus
ponens: The student’s answer does not contain any explicit illustration or use of modus ponens.
Therefore, Point earned: 0. 3. Pertains to Gardening: The student’s answer clearly relates to
Gardens, as it mentions the garden’s sprinkling system. Point earned: 1. 4. Text makes sense:
The student’s answer is a coherent sentence that makes sense in the context of gardening. Point
earned: 1. 5. At most one sentence: The student’s answer is a single sentence, which meets the
requirement. Point earned: 1.

Grade: 3 out of 5.

The student’s answer demonstrates the use of red herring and pertains to gardening, but it does not
contain modus ponens. The text makes sense and is limited to a single sentence, which meets the
given criteria.
```

F.1.4 Mistakes with skill meaning (use of common English meaning)

The following example is based on a real grading response by LLaMA-2-70B-Chat. Observe that the grader model interprets “complex question” by its meaning in colloquial English, rather than the meaning associated with logical fallacy.

```
Sure, I’d be happy to help you grade the student’s answer. Here’s my evaluation using the given
```

```
rubric:

1. Contains red herring: The student's answer demonstrates red herring in the sentence, where Mrs.
Thompson diverts attention from the main issue at hand, so Point earned: 1. 2. Contains complex
question: The student's answer contains the complex question asked by Mrs. Thompson, ''How do you
reconcile your current career with what you dreamed of doing as a child?'' Therefore, Point earned:
1. 3. Pertains to Gardening: The student's answer clearly relates to Gardens, as it mentions the
garden's sprinkling system. Point earned: 1. 4. Text makes sense: The student's answer is a
coherent sentence that makes sense in the context of gardening. Point earned: 1. 5. At most one
sentence: The student's answer is a single sentence, which meets the requirement. Point earned: 1.

Grade: 5 out of 5.

The student's answer demonstrates the use of red herring and complex question. The text makes sense
and is limited to a single sentence, which meets the given criteria.
```

F.2 Failure cases with final prompt

With our final prompt, most of the generation and grading work well. However, there are still failure cases, especially for relatively weaker models and larger k. Descriptions of some failure cases are below:

- Model is unable to follow the instructions, e.g., unable to output "Answer:" before the text for generation, and unable to output the table for grading.
- Model would directly use the word (name of skill) in the generation instead of demonstrating the skill
- Some of the k skills are simply left out of the generation
- Partially correct skills, e.g. logic skills where a premise was not explicitly stated, or the implication was partially wrong.

We will now provide examples of these failure cases. Note that these examples are not real; instead, they are modeled off of real outputs from particular models and values of k. We do this to respect the release of 10 skills and 10 topics randomly sampled from out master list of skills and topics, to avoid potential "cramming" for SKILL-MIX .

F.2.1 Unable to follow the instructions

Falcon-180B-Chat sometimes does not follow the instructions for generation. This example below is inspired by a real response provided by Falcon-180B-Chat for $k = 3$:

```
**Answer:**

Explanation: The **red herring** is shown in the first and second sentence, where Mrs. Thompson
diverts attention from the main issue at hand. The **spatial reasoning** is present in the second
sentence with the discussion of "how better plumbing routes could save water". The **self-serving
bias** is displayed in the second sentence where Mrs. Thompson "tried to deflect blame from personal
high water usage".

At the town meeting about the water shortage, Mrs. Thompson, who had recently installed a lavish
traditional gardening system in her yard sprinkler system to water her garden, blamed the town's poor
plumbing for the crisis. While many were drawn into her spatial reasoning about how better plumbing
routes could save water, some recognized that her self-serving bias was evident, as she tried to
deflect blame from personal high water usage and onto the town's infrastructure.
```

Here is another failure case inspired by a real response by LLaMA-2-7B-Chat for $k = 4$:

```
Answer: Sure, I'd be happy to help you with that! Here's an example of a minimal natural piece of
```

> text that illustrates multiple skills in semantics or syntax:
>
> Explanation:
>
> "At the town meeting about the water shortage, Mrs. Thompson, who had recently implemented a lavish traditional gardening system in her yard, asked, "Given that our town's plumbing is outdated and we all want green spaces, shouldn't we first consider overhauling our water distribution system?" While many were drawn into her spatial reasoning about how better plumbing routes could save water, some recognized that her self-serving bias was evident, as she tried to deflect blame from personal high water usage and onto the town's infrastructure."
>
> This text demonstrates the following skills in semantics or syntax:
>
> 1. Red Herring: In the second sentence, during a discussion about a water shortage, Mrs. Thompson introduces an irrelevant point to detract attention from her new gardening system.
>
> 2. Spatial Reasoning: In the second sentence, Mrs. Thompson uses spatial reasoning to discuss how better plumbing routes could save water.
>
> 3. Self Serving Bias: In the second sentence, some towns members recognize that Mrs. Thompson discusses plumbing routes in the first place stems from her need to maintain her self-esteem.
>
> 4. Complex Question: In the first sentence, Mrs. Thompson asks a question that indicates the implicit assumption that everyone in the town wants green space.
>
> Overall, this text shows how these skills in semantics or syntax can be applied in a natural and coherent way in a short piece of text.

In these situations, models for grading may receive "**" or an empty string as the student's answer. So they sometimes fail to output the table of points. The following is a typical failure case output by GPT-4:

> As the student's answer was missing, I will unfortunately be unable to grade it or provide the necessary feedback. Could you please provide me with the student's answer?

F.2.2 Using the name of the skill directly in the text

Here is an example inspired by real text generated by GPT-4. The requested skills are red herring, spatial reasoning, self serving bias, and complex question.

> Answer: "At the town meeting about the water shortage, Mrs. Thompson, who had recently implemented a lavish traditional gardening system in her yard, asked, "Given that our town's plumbing is outdated and we all want green spaces, shouldn't we first consider overhauling our water distribution system?" While many were drawn into her spatial reasoning about how better plumbing routes could save water, a few discerned it as a red herring and recognized that her self-serving bias was evident, as she tried to deflect blame from personal high water usage and onto the town's infrastructure."

F.2.3 Some of the k skills left out of the generation

The example below is inspired by real text generated by LLaMA-2-70B-Chat for $k = 3$. The requested skills are metaphor, statistical syllogism, and red herring. The topic of interest is "Dueling." Out of all the requested skills, only metaphor is present.

> Answer: "I'm not sure if I'll duel tomorrow. My opponent's six-shooter is a wild card, but my queasy stomach and off-target aim may be liabilities.

F.2.4 Partially correct skills

Here is an example inspired by real text generated by GPT-4. The requested skills are modus ponens, red herring, and metaphor.

> Answer: "If needles were the keys to crafting melodies, then every perfect stitch would be a note in a harmonious symphony; but speaking of symphonies, have you ever noticed how the early bird's song sounds just like Mozart?"

F.2.5 Unclear if sentence grammatical

Here is an example inspired by real text generated by GPT-4 with requested skill metaphor.

```
Answer: "Gardening, the mind's soil yields a bouquet of confusion."
```

G Additional experimental results

In this section, we test various instruction-tuned models (including the LLaMA-2 family, GPT family, Falcon-180B-Chat, Xwin-LM-70B-V0.1, Mistral-7B-Instruct-v0.1, and Qwen-14B-Chat) on their performance on SKILL-MIX for various k. (For details regarding our procedure, please see Appendix C and D.) For convenience, we use *saturation point* to denote the value of k at which a model's score in SKILL-MIX drops off.

We answer the following questions

- What differences arise between grading by GPT-4 vs. LLaMA-2?
- What is the effect of increasing k on SKILL-MIX performance? What is the saturation point for the instruction-tuned models?
- What is the relationship between model scale and saturation point? We are particularly interested in answering this question for the family of LLaMA-2 models, since they share the same training set and methodology.

Setup We evaluate various instruction-tuned models on SKILL-MIX (k) with $k = 2, 3, 4$. We continue to evaluate a model with $k = 5$ (potential also $k = 6$) if it does not saturate at $k = 4$. We use both GPT-4 and LLaMA-2-70B-Chat as Grader.

For each SKILL-MIX (k), we evaluate all models on 30 (k skills, 1 topic) combinations. We provide each specific combination of k skills to the (Student) model on three instances (see Figure 2). Each of the three generated texts are also graded three times (to reduce the randomness caused by the Grader), in total creating nine grading results for each (k skills, 1 topic) (see Figure 3).

Metrics Each generated text can receive up to $k + 3$ points: 1 point for each correctly illustrated skill, 1 point for sticking to the topic, 1 point for coherence / making sense, and 1 point for having at most $k - 1$ sentence. Recall that we grade each generated text three times. In each round of grading, we parse each of the criteria individually from the Grader model's output. For each criterion, we then collect the majority vote among the three grading rounds. The voted points are then converted into various metrics of interest. We define the following metrics of interest as follows:

- *Ratio*[6] *of Full Marks*: 1 if all $k + 3$ points are earned, and 0 otherwise
- *Ratio of All Skills*: 1 if k points are awarded for the k skills and at least 2 points are awarded for the remaining criteria (which allows "cheating" by exceeding sentence limit, not using topic, or not making sense), and 0 otherwise
- *Skill Fraction*: the fraction of points awarded for the k skills if all 3 points are awarded for the remaining criteria, and 0 otherwise.
- *Total Score*: sum of the individual points awarded
- *Total Skill Score*: sum of the points awarded for the k skills
- *Rescaled Score*: $\left(\frac{c}{k+3}\right)^{k+3}$ where c is the total score

For each metric, the maximum value among the 3 generations is computed, and then averaged across the 30 combinations. We then take the maximum value of the metrics among the 3 generations for a given (k skill, 1 topic) combination, and average the maximum value across the 30 combinations.

Differences in Grader Scores From Tables 1 and 5, we clearly observe that LLaMA-2 is a more generous grader than GPT-4. We also observe that when LLaMA-2 is used as the grader, it prefers generations outputted by the LLaMA-2 family. For example, across different metrics, GPT-4 generally gives a higher score to Mistral-7B-Instruct-v0.1 than to LLaMA-2-7B-Chat, but LLaMA-2 grader gives a much higher score to LLaMA-2-7B-Chat than Mistral-7B-Instruct-v0.1 for $k \geq 3$. Overall, we found via spot-checking the output that GPT-4 is a more accurate and reliable grader than LLaMA-2.

[6] This is called "ratio" because the metric is later averaged over the 30 combinations, even though this metric is 0 and 1 for a single generation.

Table 4: **(Additional metrics) Performance of various instruction-tuned student (generating) models on SKILL-MIX (k) graded by GPT-4.** Total Score/Total Skill Score/Rescaled Score are reported for each student model at $k = 2, 3, 4$.

Student (generator)	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$
LLaMA-2-7B-Chat	3.63/.87/.28	3.77/.93/.10	4.10/1.27/.06	-/-/-	-/-/-
LLaMA-2-13B-Chat	3.87/1.33/.39	4.10/1.50/.17	4.87/1.97/.15	-/-/-	-/-/-
LLaMA-2-70B-Chat	4.27/1.30/.52	4.37/1.43/.20	4.83/2.00/.11	-/-/-	-/-/-
GPT-3.5-turbo	4.57/1.63/.72	4.70/1.80/.34	5.33/2.37/.23	-/-/-	-/-/-
GPT-4	4.90/1.93/.93	5.73/2.73/.82	6.43/3.43/.63	7.20/4.23/.52	7.90/4.90/.40
Mistral-7B-Instruct-v0.1	3.80/1.23/.32	4.20/1.30/.15	4.37/1.73/.07	-/-/-	-/-/-
Qwen-14B-Chat	4.03/1.13/.45	4.27/1.50/.19	4.63/1.73/.10	-/-/-	-/-/-
Xwin-LM-70B-V0.1	4.40/1.50/.61	5.17/2.37/.47	5.70/2.97/.37	6.10/3.47/.19	-/-/-
Falcon-180B-Chat	4.10/1.27/.47	4.37/1.47/.18	4.57/1.73/.10	-/-/-	-/-/-

Increasing k degrades SKILL-MIX performance We observe that the ratio of full marks and the ratio of all skills can decrease dramatically when k increases. With the exception of GPT-4, GPT-3.5-turbo and Xwin-LM-70B-V0.1, all models saturate on or before $k = 3$ with GPT-4 grading. Amongst the small models, LLaMA-2-7B-Chat and Mistral-7B-Instruct-v0.1 saturate at $k = 2$. Since LLaMA-2 is more generous, the saturation point is usually delayed by 1 for LLaMA-2 grading.

Relationship between model scale and saturation point We find that as capacity increases on LLaMA-2, so does the saturation point. Observe that for LLaMA-2-7B-Chat, LLaMA-2-13B-Chat, LLaMA-2-70B-Chat, the saturation points (of GPT-4 grading) are $k = 2$, 3, and 3, respectively. Additionally, for any fixed k and metric type, higher model capacity corresponds to a better score amongst the LLaMA-2 model family. However, these observations do not necessarily hold true for models from different families. For example, Falcon-180B-Chat has more model parameters than Xwin-LM-70B-V0.1, yet the saturation point of Xwin-LM-70B-V0.1 is higher than that of Falcon-180B-Chat, and Xwin-LM-70B-V0.1 also outperforms Falcon-180B-Chat across all metrics for $k = 2, 3, 4$.

A deviation from model rankings on popular LLM leaderboards Recent models (i.e., Falcon-180B-Chat, Xwin-LM-70B-V0.1, Qwen-14B-Chat, Mistral-7B-Instruct-v0.1) are often introduced with their performance evaluated on AlpacaEval or Hugging Face's Open LLM Leaderboard (which contains ARC (Clark et al., 2018), HellaSwag (Zellers et al., 2019b), MMLU (Hendrycks et al., 2020), and TruthfulQA (Lin et al., 2021)), along with a comparison to the LLaMA-2 and GPT families. We find that their superior performance on those evaluations may not extend to SKILL-MIX :

- Falcon-180B-Chat ranks higher than LLaMA-2-70B-Chat on Open LLM Leaderboard, but performs worse on SKILL-MIX for both GPT-4 and LLaMA-2 grading.
- Xwin-LM-70B-V0.1 takes on first place on AlpacaEval, beating GPT-4. However, Xwin-LM-70B-V0.1 is clearly worse than GPT-4 on SKILL-MIX .
- Qwen-14B-Chat outperforms LLaMA-2-70B-Chat on MMLU, HumanEval (Chen et al., 2021) and GSM8K (Cobbe et al., 2021), but performs worse than LLaMA-2-70B-Chat for $k = 2, 3, 4$ with both GPT-4 and LLaMA-2 grading.
- Mistral-7B-v0.1 outperforms LLaMA-2 13B on all benchmarks that the Mistral AI team tested. Mistral-7B-Instruct-v0.1 (the model after instruction tuning) outperforms LLaMA-2-13B-Chat on MT-Bench (Zheng et al., 2023). Yet, the situation is reversed on SKILL-MIX .

Differences between model ranking on popular LLM leaderboards vs. SKILL-MIX provide evidence of "cramming for the leaderboard", further validating that SKILL-MIX is a good evaluation benchmark.

H Best practices for SKILL-MIX ecosystem

SKILL-MIX differs from most existing evaluations in two ways: (1) there is no dataset per se; instead, by using N skills and T topics, the tasks (prompts) are generated randomly on the fly from $\binom{N}{k}T$ possible combinations; and (2) for moderate k, the task is not easy for humans to solve, or even to grade; at a minimum, high-quality human labor is needed.

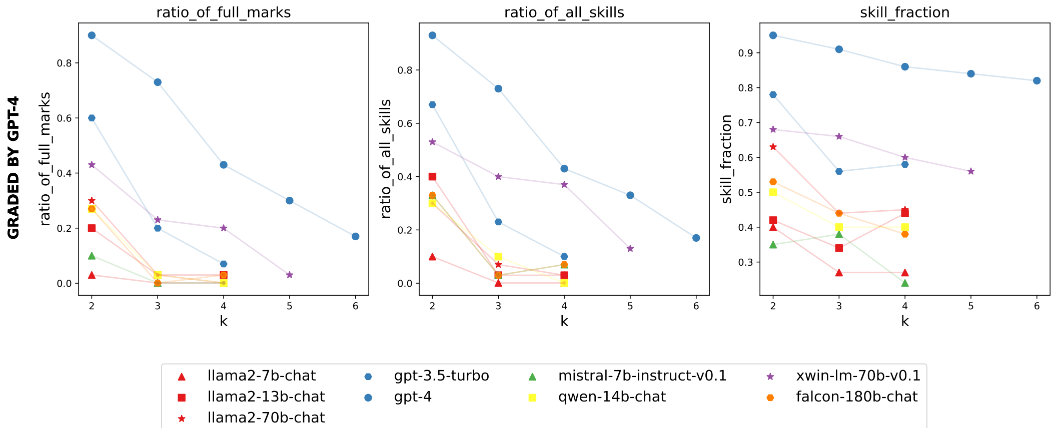

Figure 4: **Performance of various instruction-tuned student (generating) models on** SKILL-MIX (k) **graded by GPT-4.** For the accompanying table, see Table 1.

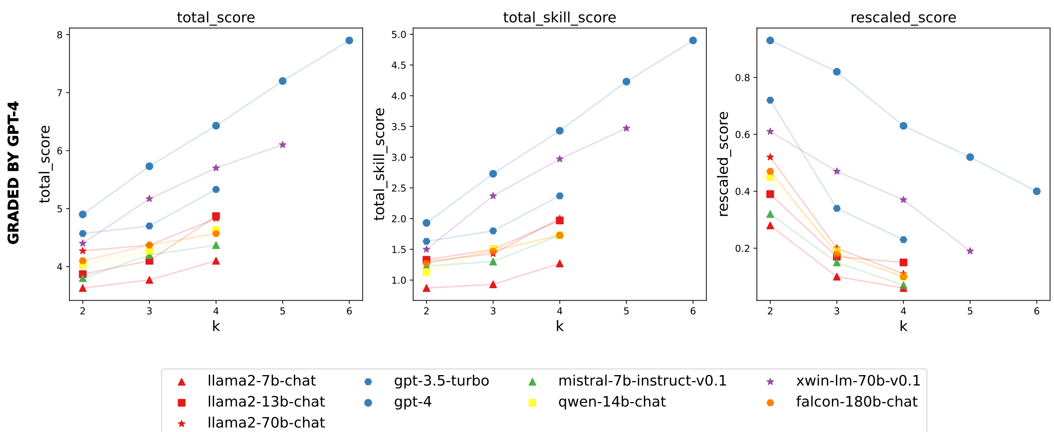

Figure 5: **(Additional metrics) Performance of various instruction-tuned student (generating) models on** SKILL-MIX (k) **graded by GPT-4.** For the accompanying table, see Table 4.

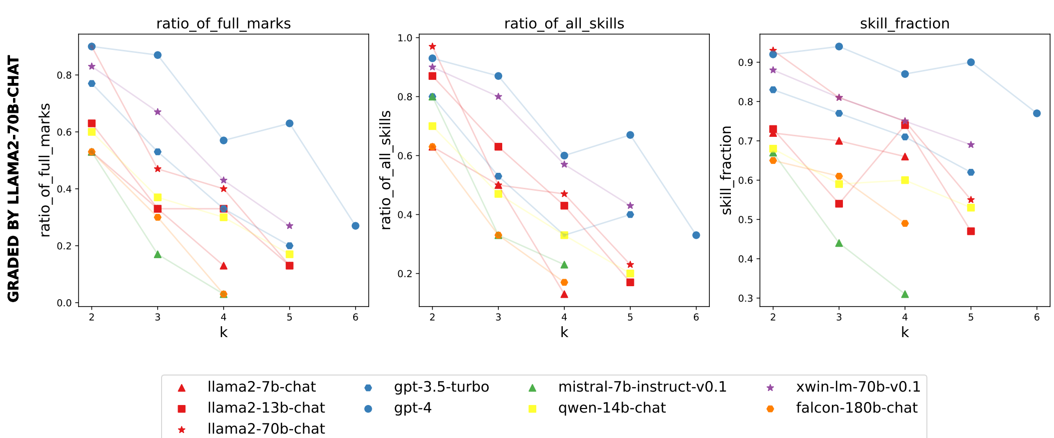

Figure 6: **Performance of various instruction-tuned student (generating) models on** SKILL-MIX (k) **graded by LLaMA-2-70B-Chat.** For the accompanying table, see Table 5.

Table 5: **Performance of various instruction-tuned student (generating) models on SKILL-MIX (k) graded by LLaMA-2-70B-Chat.** Ratio of Full Marks/Ratio of All Skills/Skill Fraction are reported for each student model at $k = 2, 3, 4$. Evaluations on $k = 5, 6$ are skipped if the Ratio of Full Marks drops below 0.3 with smaller k. Details on prompts can be found in Appendix C. See Table 6 for additional metrics.

Student (generator)	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$
LLaMA-2-7B-Chat	.53/.63/.72	.33/.50/.70	.13/.13/.66	-/-/-	-/-/-
LLaMA-2-13B-Chat	.63/.87/.73	.33/.63/.54	.33/.43/.74	.13/.17/.47	-/-/-
LLaMA-2-70B-Chat	.90/.97/.93	.47/.50/.81	.40/.47/.75	.13/.23/.55	-/-/-
GPT-3.5-turbo	.77/.80/.83	.53/.53/.77	.33/.33/.71	.20/.40/.62	-/-/-
GPT-4	.90/.93/.92	.87/.87/.94	.57/.60/.87	.63/.67/.90	.27/.33/.77
Mistral-7B-Instruct-v0.1	.53/.80/.67	.17/.33/.44	.03/.23/.31	-/-/-	-/-/-
Qwen-14B-Chat	.60/.70/.68	.37/.47/.59	.30/.33/.60	.17/.20/.53	-/-/-
Xwin-LM-70B-V0.1	.83/.90/.88	.67/.80/.81	.43/.57/.75	.27/.43/.69	-/-/-
Falcon-180B-Chat	.53/.63/.65	.30/.33/.61	.03/.17/.49	-/-/-	-/-/-

Table 6: **(Additional metrics) Performance of various instruction-tuned student (generating) models on SKILL-MIX (k) graded by LLaMA-2-70B-Chat.** Total Score/Total Skill Score/Rescaled Score are reported for each student model at $k = 2, 3, 4$.

Student (generator)	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$
LLaMA-2-7B-Chat	4.33/1.57/.66	5.10/2.40/.51	5.67/2.77/.32	-/-/-	-/-/-
LLaMA-2-13B-Chat	4.57/1.90/.74	5.03/2.50/.51	6.17/3.30/.52	5.67/3.10/.21	-/-/-
LLaMA-2-70B-Chat	4.90/1.97/.93	5.43/2.47/.64	6.03/3.20/.53	5.97/3.37/.26	-/-/-
GPT-3.5-turbo	4.63/1.73/.82	5.27/2.37/.64	5.77/2.87/.46	6.73/4.03/.38	-/-/-
GPT-4	4.80/1.90/.92	5.83/2.83/.90	6.50/3.53/.70	7.50/4.53/.73	7.73/5.00/.42
Mistral-7B-Instruct-v0.1	4.47/1.80/.67	4.30/1.83/.32	4.87/2.47/.18	-/-/-	-/-/-
Qwen-14B-Chat	4.33/1.60/.69	4.60/2.10/.48	5.40/2.70/.41	5.77/3.00/.25	-/-/-
Xwin-LM-70B-V0.1	4.83/1.90/.89	5.63/2.80/.77	6.07/3.33/.57	6.83/4.10/.44	-/-/-
Falcon-180B-Chat	4.07/1.47/.63	4.87/2.00/.44	4.90/2.23/.20	-/-/-	-/-/-

However, note that a dataset of $O(T \log T + N \log N)$ random prompts and corresponding productions already expose (with high probability) the full set of skills and topics, as well as many interesting ways to combine them. Preliminary results (the final paper will have a more definitive study) suggest that fine-tuning on such a dataset of synthetic productions can improve scores on SKILL-MIX . Since the goal of our evaluation is to test general-purpose text generation capability rather than ability on

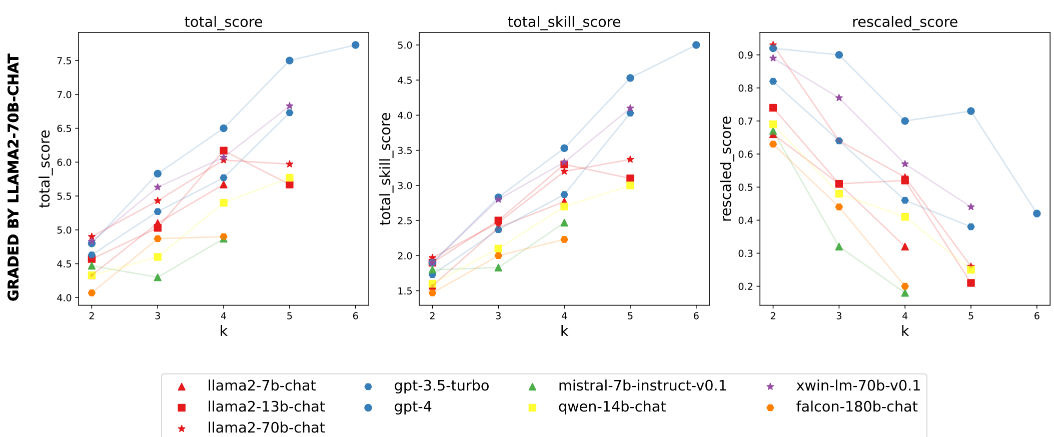

Figure 7: **(Additional metrics) Performance of various instruction-tuned student (generating) models on SKILL-MIX (k) graded by LLaMA-2-70B-Chat.** For the accompanying table, see Table 6.

the particular skills and topics, we propose to release only a random subset of 10% of skills and topics.

But, ultimately, one needs an evaluation ecosystem. This might consist of independent research groups developing versions of SKILL-MIX (in other words, tests that are chosen randomly from a high number of potential questions) with very different sets of skills and topics, including basic skills related to science, economics, law, etc. Releasing a random subset of, say, 10% of the skills and topics of a new evaluation gives the rest of the world an idea of what it tests –and how it is reasonably distinct from other existing evaluations– while retaining its difficulty. If the research groups are seen as trustworthy, this ecosystem's continuous assessment of AI capabilities might become important inputs into policy discussions in the future.

Difficulty of grading An open question in the above picture is how to grade harder versions of SKILL-MIX in the future. The obvious idea is to use the current champion model for the evaluation (provided the model retains no memory of its past interactions). But, a natural question arises: whether to trust the champion's grade for itself. This relates to an interesting debate in pedagogy (Bloom et al., 1956; Anderson & Krathwohl, 2001; Forehand, 2005): *Which is harder: to ace the exam or to grade it well?* While the answer may seem obvious in quantitative or scientific fields (i.e., acing is harder), this wasn't obvious in other fields. Today it is more broadly accepted that grading is indeed easier[7], which suggests that the champion can probably grade itself (assuming the model retains no memory of past interactions), Human spot-checking seems advisable. In our experience, GPT-4's grading capabilities for a particular k seem to be better than its SKILL-MIX score on the same k; we plan to further investigate this relationship in future work. On a related note, we found that human spot-checking becomes much easier and more accurate if the grader includes reasoning for grading decisions.

[7]But the variance we saw among human graders on our SKILL-MIX reminds us of the difficulty of grading exams where there is no obvious best answer.