CSE584 Midterm Project Report: Sentence Generation Classifier

Jihyun Janice Ahn **Bucky Park** **Berk Atil**
The Pennsylvania State University
{jfa5672, hbp5148, bka5352}@psu.edu

Abstract

Advancements in AI and NLP demand improved techniques for text completion and classification, particularly in complex linguistic contexts. This study evaluates seven state-of-the-art large language models using novel two encoder architecture with SBERT and LUAR. The prepared dataset of truncated sentences from public sources enabled extensive evaluation. The LUAR model excelled with over 15% higher accuracy, attributed to its distinct random sampling method enhancing generalizability. Findings highlight the crucial impact of model size and architecture complexity, with larger and multi-layered configurations offering superior performance. This work advances methodology by introducing an encoder setup that strengthens classification and suggests model design optimizations, highlighting potential future research avenues with diverse datasets and broader model exploration.

1 Introduction

With the rapid advancements in artificial intelligence and natural language processing, significant strides have been made in the domains of textual data completion and classification. These enhanced capabilities of generating coherent and contextually appropriate text continuations open up vast potential across various applications, from refining user interactions to propelling automated content creation. Despite these advancements, challenges persist, particularly in developing models that can accurately interpret and replicate complex linguistic styles, thereby necessitating continued exploration into model architectures and training methodologies.

This research embarks on an exploration of various advanced Large Language Models (LLMs) to assess their efficacy in completing incomplete sentences extracted from a diverse array of datasets. The investigation leverages seven state-of-the-art LLMs: GPT-4o, Gemini1.5-Pro, Llama3.1-70B, Llama3.1-8B, Qwen2.5-32B, Qwen2.5-72B, and Claude3.5-Sonnet, each tasked with generating sentence continuations (OpenAI et al., 2024; Dubey et al., 2024; Team et al., 2024; Team, 2024; Anthropic, 2024). Central to this investigation is novel encoder architectures that seeks to leverage distinct embedding spaces for enhanced classification, alongside experiments employing SBERT and LUAR as encoders (Reimers and Gurevych, 2019; Rivera-Soto et al., 2021), further enriching the analysis.

The dataset preparation involved extracting complete sentences from six publicly available datasets, followed by strategic truncation and deduplication processes, ultimately resulting in a robust set of 28,538 incomplete sentences. This structured dataset serves as the foundation for rigorous training, testing, and validation, distributed across various models to ensure balanced evaluations.

The comprehensive experimentation provides pivotal insights into how architectural choices—such as encoder configurations and classification layer complexity—impact performance. Through the meticulous evaluation of these elements, the study not only advances the current understanding of model capabilities but also sets the stage for future innovations in LLMs research and development.

2 Related Work

The increasing use of LLMs has driven the need for effective tools to detect machine-generated text and evaluate performance in sentence completion tasks (Mitchell et al.; Bao et al., 2024; Soto et al., 2024; Guo and Yu, 2023; Abburi et al., 2023; Wang et al., 2023; Akram; Su et al., 2023; Liu et al., 2024; Mireshghallah et al., 2024; Gagiano and Tian). DetectGPT (Mitchell et al.) introduces a novel zero-shot framework for detecting AI-generated content

by analyzing probability curvature, demonstrating that models can differentiate machine-generated text without prior exposure to the content. FAST-DetectGPT (Bao et al., 2024) builds upon this concept by introducing a more efficient detection mechanism that operates in zero-shot scenarios using conditional probability curvature. SeqXGPT (Wang et al., 2023) is an AI-generated text detector focusing on sentences instead of documents. They make use of log probabilities from several models as features. (Gagiano and Tian, 2023) proposes a prompting strategy using predictions and confidence scores of a fine-tuned model. Hence, they aim to make the model think more. In addition, they use a label smoothing strategy to fine-tune a model. (Soto et al., 2024) proposes few-shot learning using style representations, called LUAR (Rivera-Soto et al., 2021) to detect machine generated task. (Guo and Yu, 2023) proposes a model denoising strategy based on the assumption that human-written text should reside outside the distribution of machine-generated text. They first denoise input text with artificially added noise using an LLM. Then, they semantically compare that denoised text with the original text to detect if the text is generated by a machine. Another interesting finding is that smaller and partially trained models are better machine generated text detectors (Mireshghallah et al., 2024). Their approach is based on a curvature test using the likelihood surface of a surrogate detection model.

Another research area is about authorship verification (AV), which decides if a pair of text is written by the same author. Our task is different from AV in the sense that we predict the author. Therefore, our task is more specific because in addition to differentiating the authors, we predict the specific author. The similarity is that writing styles play an important role. (Huang et al., 2024a) reviews the challenges and methodologies of authorship attribution with LLMs. They highlight the importance of generalization across domains and improved explainability in authorship detection methods. The capability of LLMs on authorship attribution and verification has also been analyzed (Huang et al., 2024b). They propose a linguistically informed prompting for zero-shot setting and show that there is no need to fine-tune a model for this task. Similar to (Huang et al., 2024b), (Hung et al., 2023) also proposes different prompting strategies for AV. They add several criteria to the prompt that are relevant to the styles of a text such as tone, mood, idioms, or punctuation. (Rivera-Soto et al., 2021) develops a new model to learn authorship representations using contrastive loss. Their training process tries to make the texts' representations from the same author similar. They use SBERT (Reimers and Gurevych, 2019) as an encoder. (Wegmann et al., 2022) argues that to learn style representations, using an authorship verification task may not be appropriate because, in this way, models might learn content too. Therefore, they propose a content control mechanism using contrastive learning. Their experiments show that learning style representations this way is more effective.

3 Datasets

3.1 Incomplete sentences

For our experiments, we first extracted complete sentences with at least 8 words from six publicly available datasets. After extracting the complete sentences from each dataset, we truncated them by keeping the first half based on the word count. In other words, if a sentence S contains n words, we kept only the first $\lceil \frac{n}{2} \rceil$ words. For example, the sentence with 16 words, *"A man is being pulled on a water ski as he floats in the water casually."*, was truncated to *"A man is being pulled on a water"*, keeping the first 8 words.

After truncation, we identified identical sentences in the dataset and removed these duplicates. With all the processing, we ended up with 28,538 incomplete sentences across the six datasets. Below is a brief description of each dataset we utilized.

3.1.1 HellaSwag

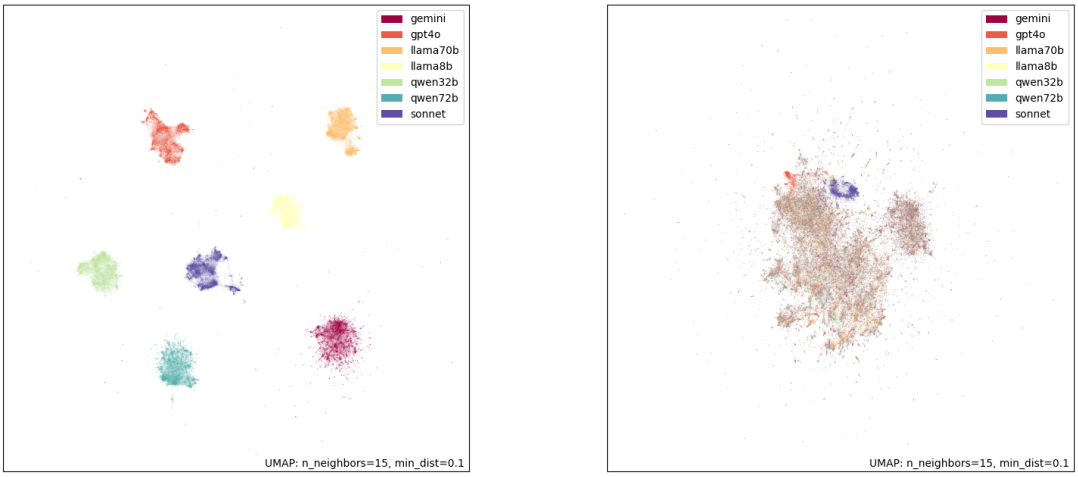
The HellaSwag dataset[1] (Zellers et al., 2019) is designed to test ability of models to predict the most plausible continuation of a given sentence. `ctx_a` column contains complete sentences for each example. We extracted 6,357 sentences from this column.

3.1.2 SNLI Corpus

The Stanford Natural Language Inference(SNLI) Corpus dataset[2] (Bowman et al., 2015) is designed to train and evaluate models on natural language understanding tasks, which was released by the Stanford Natural Language Processing Group.

[1] https://www.kaggle.com/datasets/thedevastator/hellaswag-a-new-commonsense-nli-dataset

[2] https://www.kaggle.com/datasets/stanfordu/stanford-natural-language-inference-corpus/data



(a) Embedded using LUAR

(b) Embedded using SBERT

Figure 1: Two-dimensional visualisations of embeddings of X_js

`sentence1` column contains complete sentences for each example. We extracted 6,928 sentences from this column.

3.1.3 COCO caption

The COCO caption dataset[3] is used for the Microsoft COCO dataset to provide image captions in natural language. `caption` column contains complete sentences for each example. Total number of sentences we can access was 762 and we extracted 559 unique sentences from this column.

3.1.4 MultiNLI

The Multi-Genre Natural Language Inference (MultiNLI) Corpus dataset[4] (Williams et al., 2018) is designed to evaluate models on sentence-level natural language understanding across multiple genres. The dataset was released by the New York University (NYU) Natural Language Processing Group. `sentence1` and `sentence2` columns contain complete sentences for each example. We extracted 4,950 sentences from them.

3.1.5 Brown Corpus

The Brown Corpus dataset [5] is one of the first major collections of text for linguistic research, covering a wide variety of genres in American English. `tokenized_text` column contains complete sentences for each example. We extracted 4,943 sentences from this column.

3.1.6 Bible Corpus

The Bible Corpus dataset[6] contains the full text of the Bible in the American Standard Version. `t` column contains complete sentences for each verse. We extracted 4,801 sentences from this column.

Datasets	# of sentences	# of sentences without duplicates
HellaSwag	7119	6357
SNLI Corpus	7119	6928
COCO caption	762	559
MultiNLI	5000	4950
Brown Corpus	5000	4943
Bible Corpus	5000	4801
Total	30000	28538

Table 1: The number of sentences from each dataset before and after removing duplicates.

3.2 Complete sentences from Large Language Models

To complete the given incomplete sentences, we used 7 models: GPT-4o, Gemini1.5-Pro, Llama3.1-70B, Llama3.1-8B, Qwen2.5-32b, Qwen2.5-72B, and Claude3.5-Sonnet (OpenAI et al., 2024; Dubey

[3] https://github.com/tylin/coco-caption (Chen et al., 2015)

[4] https://cims.nyu.edu/~sbowman/multinli/

[5] https://www.kaggle.com/datasets/nltkdata/brown-corpus/data (Francis and Kucera, 1964)

[6] https://www.kaggle.com/datasets/oswinrh/bible?select=t_asv.csv

et al., 2024; Team et al., 2024; Team, 2024; Anthropic, 2024). For Gemini and Claude, we generated the following sentences using the prompt *"Complete the given sentence. Do not give any explanation."*, while for the others, we used the prompt *"Complete the given sentence"*. We then stored all the generated sentences from each model in separate files and combined them into a single file.[7] While all the models except Gemini generated following sentences for every given incomplete sentence, Gemini did not generate following sentences for seven of the incomplete sentences due to sensitivity reasons, and we excluded these seven sentences from the Gemini dataset. We preprocessed datasets and as shown on the Tab. 1, we have about 28K unique sentences and each dataset sources has similar distributions except COCO caption as mentioned in section 3.1.3.

4 Classification Model

We propose a simple but effective neural architecture. Our dual encoder (Bromley et al., 1993) learns different embedding spaces for the X_is and X_js. We concatenate the embeddings of X_is and X_js and feed them into a simple neural classifier consisting of one or two fully connected layers. As encoders, we experiment with SBERT (Reimers and Gurevych, 2019), which was designed to learn representations for semantic similarity, and LUAR (Rivera-Soto et al., 2021), which was intended to model authorship styles. We use cross-entropy loss, Adam optimizer (Kingma and Ba, 2014), batch size of 8, and learning rate of 0.0001. Because of time constraints, we could not do a lot of hyperparameter optimizations. Our hypothesis was that LUAR should perform way better based on the findings in (Rivera-Soto et al., 2021) and Fig 1. LUAR is more capable of distinguishing different styles of texts.

The experiment was initiated with methods aimed at enhancing the diversity and robustness of the data within the evaluation framework. The augmented datasets were labeled from 0 to 6, each number indicating the specific LLM responsible for generating a given sentence. This consolidated dataset served as the primary input. of rigorous training and evaluation, 70% of the data was allocated for training, 10% for testing, and the remaining 20% for validation. This stratified approach ensured balanced distribution across all dataset splits, which is essential for a comprehensive assessment of model performance.

To evaluate the influence of different architectural setups, the performances of single-encoder and dual-encoder configurations were compared. This assessment provided insights on how the architectural type impacts the models' effectiveness. Moreover, differential impacts of **classification layer** complexity were investigated by conducting experiments with one-layer and two-layer setups. The one-layer setup utilized a linear classification layer, whereas the two-layer scenario incorporated an additional linear layer with a ReLU activation function. This exploration highlighted the role of network depth and non-linearity in classification accuracy.

Environment setups for each model were meticulously configured due to their unique requirements. The LUAR model necessitated a specialized setup, prompting differentiated environmental configurations for each model. For instance, the SBERT configuration employed the latest module versions from the transformers library to ensure compatibility and optimal performance. Detailed module version specifications used for the LUAR experiment are available in the GitHub repository[8], providing transparency and supporting reproducibility for future research.

Lastly, the **ablation study** concentrated on evaluating the contribution of individual components within the classification model. Models were trained using only X_j, while other configuration parameters remained constant, except for the classification layers. By systematically omitting specific inputs, the study discerned the impact of each component on overall model performance, thereby emphasizing the significance of each feature within the augmented setup. This analysis was integral in understanding model nuances and guiding future improvements.

5 Results

5.1 Main Experiment

For each encoder, we experimented with six architectural setups, including the ablation case. First, we applied both single- and dual-encoder configurations to the datasets. Additionally, we created two setups by varying the number of classification

[7] https://github.com/berkatil/cse584-midterm-project/blob/main/dataset/combined_data.csv

[8] https://github.com/berkatil/cse584-midterm-project/blob/main/environment.yml

layers, 1 and 2, for each encoding method. Each experiment results were represented with confusion matrix and stored in our github repository[9]. Fig 2 is one of the confusion matrix.

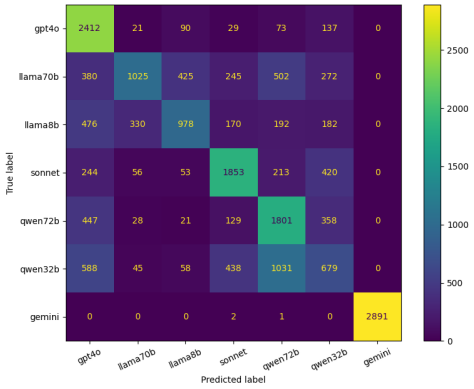


Figure 2: LUAR result with 3 epoch, dual encoder, and 2 classification layers

Since our task is a classification problem, we evaluate their performances using the following metrics: Precision, Recall, F1 Score, Accuracy. The results in these metrics are presented in Table 2.

As expected, LUAR significantly outperformed, achieving around 0.6 across all metrics. In contrast, the performances of SBERT varied roughly between 0.3 to 0.4, depending on the setups, even in the ablation case. We will further elaborate on the results in the next section.

5.2 Ablation Study

For the ablation study, we applied both SBERT and LUAR encoder as a single encoder configuration. We provided the generated sentences X_js only without the given incomplete sentences for this study and results are shown in Table 3.

6 Discussion

The results of the experiment underscore the superior performance of the LUAR model, which outperformed its counterparts by achieving more than 15% higher accuracy, alongside other evaluative scores as depicted in Table 2. This outcome not only aligns with the initial hypothesis but also validates the efficacy of the LUAR model's design, particularly its employment of random sampling during training. This additional step has been referenced in literature to enhance model generalizability (Rivera-Soto et al., 2021; Boenninghoff et al., 2020).

Analyzing the impact of epoch quantity, it was observed that increasing the number of epochs consistently enhanced model performance. Furthermore, the architecture of classification layers was found to significantly affect model outcomes. A single-layer classifier tends to oversimplify logical deductions, potentially leading to underfitting, whereas models utilizing two classification layers exhibited markedly improved performance metrics.

Model	# of unique tokens	average # of words
GPT4o	33174	19.03
Llama70b	39115	12.50
Llama8b	34721	16.68
Sonnet	28450	12.37
Qwen72b	22913	6.52
Qwen32b	27646	8.43
Gemini	20684	4.10

Table 4: Model response statistics

When evaluating different classifiers, it emerged that sentence classification efficiency is higher for models such as GPT4o, Sonnet, and Gemini. Conversely, both classifiers encountered difficulties in classifying sentences generated by the qwen32b model.

The effect of model size was evident in comparative analyses within the study, particularly among the Llama and Qwen models. While Llama demonstrated consistent performance regardless of size, variations in Qwen model sizes significantly influenced the LUAR classifier's performance. It is usually easier to find larger models, suggesting a correlation between model size and classification efficacy.

An additional experiment was conducted to explore the correlation between dataset characteristics and model accuracy, focusing on factors such as the number of unique tokens and average word length (see Table 4). This analysis highlighted that generated sentences by LLMs such as GPT4o, Sonnet, and Gemini were more effective, while outputs from the Qwen32b model posed classification challenges. Although certain correlations between accuracy, unique tokens, and word count were identified, they were not statistically significant, likely attributable to the limited sample size. Nonetheless, the study indicates that model size potentially

[9] https://github.com/berkatil/cse584-midterm-project

Encoder	Setup	Precision	Recall	F1 Score	Accuracy
SBERT	Dual, 1 Layer	0.370121	0.375202	0.356964	0.377922
	Dual, 2 Layers	**0.436489**	**0.430565**	**0.423831**	**0.432599**
	Single, 1 Layer	0.302669	0.311536	0.294239	0.313294
	Single, 2 Layers	0.367233	0.375006	0.362895	0.376522
LUAR	Dual, 1 Layer	**0.628634**	0.607513	0.584749	0.609173
	Dual, 2 Layers	0.616155	0.606646	0.593206	**0.611661**
	Single, 1 Layer	0.610068	**0.607937**	**0.596391**	0.610884
	Single, 2 Layers	0.606333	0.593647	0.581397	0.599741

Table 2: Performance of each setup with different metrics

Encoder	Setup	Precision	Recall	F1 Score	Accuracy
SBERT	1 Layer	0.366762	0.388330	0.363484	0.392433
	2 Layers	**0.436988**	**0.448674**	**0.434210**	**0.451568**
LUAR	1 Layer	**0.621281**	0.613296	0.590984	0.615807
	2 Layers	0.612787	**0.620237**	**0.594096**	**0.620368**

Table 3: Performance of the ablation case with different metrics

plays a crucial role in influencing classification efficiency across diverse large language models. This finding underscores the importance of considering model size in the development and deployment of classification systems.

Upon analyzing the ablation study results, a comparison between the best outcomes from Table 2 and Table 3 reveals that the ablation study consistently produced higher scores across all metrics. This suggests that the X_is included in the main experiment have a minimal impact on performance and may act as noise in the dataset. One possible explanation is that providing identical X_is simply completes the input sentences without contributing any meaningful information for the classification task.

7 Conclusion

In conclusion, this study comprehensively evaluated the performance of various large language models (LLMs) and neural architectures in tasks related to sentence completion and classification. The LUAR model demonstrated marked superiority, corroborating the initial hypothesis and effectively distinguishing between textual styles. It was observed that enhancing the number of epochs and classification layers significantly bolstered model performance, thus mitigating tendencies of underfitting observed in simpler classifiers.

Significant insights emerged regarding model size, particularly in the Llama and Qwen families. Our classifier exhibited stable performance regardless of size for Llama models, contrasting with the Qwen models, where larger configurations appreciably enhanced classification accuracy. This finding underscores the necessity of factoring in model size for optimal model deployment. Additionally, while models like GPT4o, Sonnet, and Gemini effectively managed complex sentence structures, they faced challenges with some outputs, reflecting the critical influence of underlying architecture on model competency.

Furthermore, an analysis of dataset characteristics, such as unique token count and word length, revealed some degree of correlation with accuracy. However, these correlations were less significant compared to the dominant impact of model size and structure. Therefore, future research could explore larger datasets to better understand these dynamics and consider additional architectural innovations to further advance model efficacy.

8 Limitation and Future Works

Limitation: Although the dataset was selected with the intention of maintaining fairness, the limited sample size from the COCO caption dataset (559 X_js) necessitated supplementation from another dataset, which could introduce slight biases in the findings. Moreover, in the context of the LUAR model experiments, an earlier version

of the transformer module (4.38.1) had to be utilized instead of the latest version employed for the SBERT experiments, potentially affecting the results.

Future Work: Future research could explore the integration of additional classifiers beyond the two initially selected, to provide a more comprehensive analysis. Furthermore, an in-depth examination of performance variations across different datasets would represent a promising direction for subsequent experimental efforts.

References

Harika Abburi, Kalyani Roy, Michael Suesserman, Nirmala Pudota, Balaji Veeramani, Edward Bowen, and Sanmitra Bhattacharya. 2023. A Simple yet Efficient Ensemble Approach for AI-generated Text Detection. *arXiv preprint*. ArXiv:2311.03084 [cs].

Arslan Akram. An Empirical Study of AI Generated Text Detection Tools.

Anthropic. 2024. Claude 3.5 sonnet.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature. *arXiv preprint*. ArXiv:2310.05130 [cs].

Benedikt Boenninghoff, Julian Rupp, Robert M. Nickel, and Dorothea Kolossa. 2020. Deep bayes factor scoring for authorship verification. *Preprint*, arXiv:2008.10105.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1993. Signature verification using a "siamese" time delay neural network. In *Proceedings of the 6th International Conference on Neural Information Processing Systems*, NIPS'93, page 737–744, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue

Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

W. Nelson Francis and Henry Kucera. 1964. *Manual of Information to Accompany A Standard Corpus of Present-Day Edited American English, for Use with Digital Computers*. Department of Linguistics, Brown University.

Rinaldo Gagiano and Lin Tian. A Prompt in the Right Direction: Prompt Based Classification of Machine-Generated Text Detection.

Rinaldo Gagiano and Lin Tian. 2023. A prompt in the right direction: Prompt based classification of machine-generated text detection. In *Proceedings of the 21st Annual Workshop of the Australasian Language Technology Association*, pages 153–158.

Zhen Guo and Shangdi Yu. 2023. AuthentiGPT: Detecting Machine-Generated Text via Black-Box Language Models Denoising. *arXiv preprint*. ArXiv:2311.07700 [cs].

Baixiang Huang, Canyu Chen, and Kai Shu. 2024a. Authorship Attribution in the Era of LLMs: Problems, Methodologies, and Challenges. *arXiv preprint*. ArXiv:2408.08946 [cs].

Baixiang Huang, Canyu Chen, and Kai Shu. 2024b. Can Large Language Models Identify Authorship? *arXiv preprint*. ArXiv:2403.08213 [cs].

Chia-Yu Hung, Zhiqiang Hu, Yujia Hu, and Roy Ka-Wei Lee. 2023. Who Wrote it and Why? Prompting Large-Language Models for Authorship Verification. *arXiv preprint*. ArXiv:2310.08123 [cs].

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Zeyan Liu, Zijun Yao, Fengjun Li, and Bo Luo. 2024. On the Detectability of ChatGPT Content: Benchmarking, Methodology, and Evaluation through the Lens of Academic Writing. *arXiv preprint*. ArXiv:2306.05524 [cs].

Niloofar Mireshghallah, Justus Mattern, Sicun Gao, Reza Shokri, and Taylor Berg-Kirkpatrick. 2024. Smaller Language Models are Better Black-box Machine-Generated Text Detectors. *arXiv preprint*. ArXiv:2305.09859 [cs].

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Rafael A. Rivera-Soto, Olivia Elizabeth Miano, Juanita Ordonez, Barry Y. Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. 2021. Learning Universal Authorship Representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 913–919, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rafael Rivera Soto, Kailin Koch, Aleem Khan, Barry Chen, Marcus Bishop, and Nicholas Andrews. 2024. Few-Shot Detection of Machine-Generated Text using Style Representations. *arXiv preprint*. ArXiv:2401.06712 [cs].

Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. 2023. DetectLLM: Leveraging Log Rank Information for Zero-Shot Detection of Machine-Generated Text. *arXiv preprint*. ArXiv:2306.05540 [cs].

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya, Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud, Linda Friso, Abhimanyu Goyal, Ben Caine, Kris Cao, Da-Woon Chung, Matthew Lamm, Gabe Barth-Maron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Orgil Batsaikhan, Mateo Wirth, James Qin, Ivo Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurumurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya Attaluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine, Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillicrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette, Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang, Thi Avrahami, Ye Zhang, Emanuel Taropa, Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano, HyunJeong Choe, Alex Tomala, Chalence Safranek-Shrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire Fortunato, Christina Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakićević, Dominik Grewe, Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Sid Dalmia, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong, Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny Huang, Vedant Misra, Jeremy Chen, RJ Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkelsson, Marcello Maggioni, Daniel Zheng, Yury Sulsky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer, Lars Lowe Sjos, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh Newlan, Zeyncep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlas, Ada Ma, Juraj Gottweis, Jinwei Xing, Chenjie Gu, Jin Miao, Christian Frank, Zeynep Cankara, Sanjay Ganapathy, Ishita Dasgupta, Steph Hughes-

Fitt, Heng Chen, David Reid, Keran Rong, Hongmin Fan, Joost van Amersfoort, Vincent Zhuang, Aaron Cohen, Shixiang Shane Gu, Anhad Mohananey, Anastasija Ilic, Taylor Tobin, John Wieting, Anna Bortsova, Phoebe Thacker, Emma Wang, Emily Caveness, Justin Chiu, Eren Sezener, Alex Kaskasoli, Steven Baker, Katie Millican, Mohamed Elhawaty, Kostas Aisopos, Carl Lebsack, Nathan Byrd, Hanjun Dai, Wenhao Jia, Matthew Wiethoff, Elnaz Davoodi, Albert Weston, Lakshman Yagati, Arun Ahuja, Isabel Gao, Golan Pundak, Susan Zhang, Michael Azzam, Khe Chai Sim, Sergi Caelles, James Keeling, Abhanshu Sharma, Andy Swing, YaGuang Li, Chenxi Liu, Carrie Grimes Bostock, Yamini Bansal, Zachary Nado, Ankesh Anand, Josh Lipschultz, Abhijit Karmarkar, Lev Proleev, Abe Ittycheriah, Soheil Hassas Yeganeh, George Polovets, Aleksandra Faust, Jiao Sun, Alban Rrustemi, Pen Li, Rakesh Shivanna, Jeremiah Liu, Chris Welty, Federico Lebron, Anirudh Baddepudi, Sebastian Krause, Emilio Parisotto, Radu Soricut, Zheng Xu, Dawn Bloxwich, Melvin Johnson, Behnam Neyshabur, Justin Mao-Jones, Renshen Wang, Vinay Ramasesh, Zaheer Abbas, Arthur Guez, Constant Segal, Duc Dung Nguyen, James Svensson, Le Hou, Sarah York, Kieran Milan, Sophie Bridgers, Wiktor Gworek, Marco Tagliasacchi, James Lee-Thorp, Michael Chang, Alexey Guseynov, Ale Jakse Hartman, Michael Kwong, Ruizhe Zhao, Sheleem Kashem, Elizabeth Cole, Antoine Miech, Richard Tanburn, Mary Phuong, Filip Pavetic, Sebastien Cevey, Ramona Comanescu, Richard Ives, Sherry Yang, Cosmo Du, Bo Li, Zizhao Zhang, Mariko Iinuma, Clara Huiyi Hu, Aurko Roy, Shaan Bijwadia, Zhenkai Zhu, Danilo Martins, Rachel Saputro, Anita Gergely, Steven Zheng, Dawei Jia, Ioannis Antonoglou, Adam Sadovsky, Shane Gu, Yingying Bi, Alek Andreev, Sina Samangooei, Mina Khan, Tomas Kocisky, Angelos Filos, Chintu Kumar, Colton Bishop, Adams Yu, Sarah Hodkinson, Sid Mittal, Premal Shah, Alexandre Moufarek, Yong Cheng, Adam Bloniarz, Jaehoon Lee, Pedram Pejman, Paul Michel, Stephen Spencer, Vladimir Feinberg, Xuehan Xiong, Nikolay Savinov, Charlotte Smith, Siamak Shakeri, Dustin Tran, Mary Chesus, Bernd Bohnet, George Tucker, Tamara von Glehn, Carrie Muir, Yiran Mao, Hideto Kazawa, Ambrose Slone, Kedar Soparkar, Disha Shrivastava, James Cobon-Kerr, Michael Sharman, Jay Pavagadhi, Carlos Araya, Karolis Misiunas, Nimesh Ghelani, Michael Laskin, David Barker, Qiujia Li, Anton Briukhov, Neil Houlsby, Mia Glaese, Balaji Lakshminarayanan, Nathan Schucher, Yunhao Tang, Eli Collins, Hyeontaek Lim, Fangxiaoyu Feng, Adria Recasens, Guangda Lai, Alberto Magni, Nicola De Cao, Aditya Siddhant, Zoe Ashwood, Jordi Orbay, Mostafa Dehghani, Jenny Brennan, Yifan He, Kelvin Xu, Yang Gao, Carl Saroufim, James Molloy, Xinyi Wu, Seb Arnold, Solomon Chang, Julian Schrittwieser, Elena Buchatskaya, Soroush Radpour, Martin Polacek, Skye Giordano, Ankur Bapna, Simon Tokumine, Vincent Hellendoorn, Thibault Sottiaux, Sarah Cogan, Aliaksei Severyn, Mohammad Saleh, Shantanu Thakoor, Laurent Shefey, Siyuan Qiao, Meenu Gaba, Shuo yiin Chang, Craig Swanson, Biao Zhang, Benjamin Lee, Paul Kishan Rubenstein, Gan Song, Tom Kwiatkowski, Anna Koop, Ajay Kannan, David Kao, Parker Schuh, Axel Stjerngren, Golnaz Ghiasi, Gena Gibson, Luke Vilnis, Ye Yuan, Felipe Tiengo Ferreira, Aishwarya Kamath, Ted Klimenko, Ken Franko, Kefan Xiao, Indro Bhattacharya, Miteyan Patel, Rui Wang, Alex Morris, Robin Strudel, Vivek Sharma, Peter Choy, Sayed Hadi Hashemi, Jessica Landon, Mara Finkelstein, Priya Jhakra, Justin Frye, Megan Barnes, Matthew Mauger, Dennis Daun, Khuslen Baatarsukh, Matthew Tung, Wael Farhan, Henryk Michalewski, Fabio Viola, Felix de Chaumont Quitry, Charline Le Lan, Tom Hudson, Qingze Wang, Felix Fischer, Ivy Zheng, Elspeth White, Anca Dragan, Jean baptiste Alayrac, Eric Ni, Alexander Pritzel, Adam Iwanicki, Michael Isard, Anna Bulanova, Lukas Zilka, Ethan Dyer, Devendra Sachan, Srivatsan Srinivasan, Hannah Muckenhirn, Honglong Cai, Amol Mandhane, Mukarram Tariq, Jack W. Rae, Gary Wang, Kareem Ayoub, Nicholas FitzGerald, Yao Zhao, Woohyun Han, Chris Alberti, Dan Garrette, Kashyap Krishnakumar, Mai Gimenez, Anselm Levskaya, Daniel Sohn, Josip Matak, Inaki Iturrate, Michael B. Chang, Jackie Xiang, Yuan Cao, Nishant Ranka, Geoff Brown, Adrian Hutter, Vahab Mirrokni, Nanxin Chen, Kaisheng Yao, Zoltan Egyed, Francois Galilee, Tyler Liechty, Praveen Kallakuri, Evan Palmer, Sanjay Ghemawat, Jasmine Liu, David Tao, Chloe Thornton, Tim Green, Mimi Jasarevic, Sharon Lin, Victor Cotruta, Yi-Xuan Tan, Noah Fiedel, Hongkun Yu, Ed Chi, Alexander Neitz, Jens Heitkaemper, Anu Sinha, Denny Zhou, Yi Sun, Charbel Kaed, Brice Hulse, Swaroop Mishra, Maria Georgaki, Sneha Kudugunta, Clement Farabet, Izhak Shafran, Daniel Vlasic, Anton Tsitsulin, Rajagopal Ananthanarayanan, Alen Carin, Guolong Su, Pei Sun, Shashank V, Gabriel Carvajal, Josef Broder, Iulia Comsa, Alena Repina, William Wong, Warren Weilun Chen, Peter Hawkins, Egor Filonov, Lucia Loher, Christoph Hirnschall, Weiyi Wang, Jingchen Ye, Andrea Burns, Hardie Cate, Diana Gage Wright, Federico Piccinini, Lei Zhang, Chu-Cheng Lin, Ionel Gog, Yana Kulizhskaya, Ashwin Sreevatsa, Shuang Song, Luis C. Cobo, Anand Iyer, Chetan Tekur, Guillermo Garrido, Zhuyun Xiao, Rupert Kemp, Huaixiu Steven Zheng, Hui Li, Ananth Agarwal, Christel Ngani, Kati Goshvadi, Rebeca Santamaria-Fernandez, Wojciech Fica, Xinyun Chen, Chris Gorgolewski, Sean Sun, Roopal Garg, Xinyu Ye, S. M. Ali Eslami, Nan Hua, Jon Simon, Pratik Joshi, Yelin Kim, Ian Tenney, Sahitya Potluri, Lam Nguyen Thiet, Quan Yuan, Florian Luisier, Alexandra Chronopoulou, Salvatore Scellato, Praveen Srinivasan, Minmin Chen, Vinod Koverkathu, Valentin Dalibard, Yaming Xu, Brennan Saeta, Keith Anderson, Thibault Sellam, Nick Fernando, Fantine Huot, Junehyuk Jung, Mani Varadarajan, Michael Quinn, Amit Raul, Maigo Le, Ruslan Habalov, Jon Clark, Komal Jalan, Kalesha Bullard, Achintya Singhal, Thang Luong, Boyu Wang, Sujeevan Rajayogam, Julian Eisenschlos, Johnson Jia, Daniel Finchelstein, Alex Yakubovich, Daniel Balle, Michael Fink, Sameer Agarwal, Jing Li, Dj Dvijotham, Shalini Pal, Kai Kang, Jaclyn

Konzelmann, Jennifer Beattie, Olivier Dousse, Diane Wu, Remi Crocker, Chen Elkind, Siddhartha Reddy Jonnalagadda, Jong Lee, Dan Holtmann-Rice, Krystal Kallarackal, Rosanne Liu, Denis Vnukov, Neera Vats, Luca Invernizzi, Mohsen Jafari, Huanjie Zhou, Lilly Taylor, Jennifer Prendki, Marcus Wu, Tom Eccles, Tianqi Liu, Kavya Kopparapu, Francoise Beaufays, Christof Angermueller, Andreea Marzoca, Shourya Sarcar, Hilal Dib, Jeff Stanway, Frank Perbet, Nejc Trdin, Rachel Sterneck, Andrey Khorlin, Dinghua Li, Xihui Wu, Sonam Goenka, David Madras, Sasha Goldshtein, Willi Gierke, Tong Zhou, Yaxin Liu, Yannie Liang, Anais White, Yunjie Li, Shreya Singh, Sanaz Bahargam, Mark Epstein, Sujoy Basu, Li Lao, Adnan Ozturel, Carl Crous, Alex Zhai, Han Lu, Zora Tung, Neeraj Gaur, Alanna Walton, Lucas Dixon, Ming Zhang, Amir Globerson, Grant Uy, Andrew Bolt, Olivia Wiles, Milad Nasr, Ilia Shumailov, Marco Selvi, Francesco Piccinno, Ricardo Aguilar, Sara McCarthy, Misha Khalman, Mrinal Shukla, Vlado Galic, John Carpenter, Kevin Villela, Haibin Zhang, Harry Richardson, James Martens, Matko Bosnjak, Shreyas Rammohan Belle, Jeff Seibert, Mahmoud Alnahlawi, Brian McWilliams, Sankalp Singh, Annie Louis, Wen Ding, Dan Popovici, Lenin Simicich, Laura Knight, Pulkit Mehta, Nishesh Gupta, Chongyang Shi, Saaber Fatehi, Jovana Mitrovic, Alex Grills, Joseph Pagadora, Dessie Petrova, Danielle Eisenbud, Zhishuai Zhang, Damion Yates, Bhavishya Mittal, Nilesh Tripuraneni, Yannis Assael, Thomas Brovelli, Prateek Jain, Mihajlo Velimirovic, Canfer Akbulut, Jiaqi Mu, Wolfgang Macherey, Ravin Kumar, Jun Xu, Haroon Qureshi, Gheorghe Comanici, Jeremy Wiesner, Zhitao Gong, Anton Ruddock, Matthias Bauer, Nick Felt, Anirudh GP, Anurag Arnab, Dustin Zelle, Jonas Rothfuss, Bill Rosgen, Ashish Shenoy, Bryan Seybold, Xinjian Li, Jayaram Mudigonda, Goker Erdogan, Jiawei Xia, Jiri Simsa, Andrea Michi, Yi Yao, Christopher Yew, Steven Kan, Isaac Caswell, Carey Radebaugh, Andre Elisseeff, Pedro Valenzuela, Kay McKinney, Kim Paterson, Albert Cui, Eri Latorre-Chimoto, Solomon Kim, William Zeng, Ken Durden, Priya Ponnapalli, Tiberiu Sosea, Christopher A. Choquette-Choo, James Manyika, Brona Robenek, Harsha Vashisht, Sebastien Pereira, Hoi Lam, Marko Velic, Denese Owusu-Afriyie, Katherine Lee, Tolga Bolukbasi, Alicia Parrish, Shawn Lu, Jane Park, Balaji Venkatraman, Alice Talbert, Lambert Rosique, Yuchung Cheng, Andrei Sozanschi, Adam Paszke, Praveen Kumar, Jessica Austin, Lu Li, Khalid Salama, Wooyeol Kim, Nandita Dukkipati, Anthony Baryshnikov, Christos Kaplanis, XiangHai Sheng, Yuri Chervonyi, Caglar Unlu, Diego de Las Casas, Harry Askham, Kathryn Tunyasuvunakool, Felix Gimeno, Siim Poder, Chester Kwak, Matt Miecnikowski, Vahab Mirrokni, Alek Dimitriev, Aaron Parisi, Dangyi Liu, Tomy Tsai, Toby Shevlane, Christina Kouridi, Drew Garmon, Adrian Goedeckemeyer, Adam R. Brown, Anitha Vijayakumar, Ali Elqursh, Sadegh Jazayeri, Jin Huang, Sara Mc Carthy, Jay Hoover, Lucy Kim, Sandeep Kumar, Wei Chen, Courtney Biles, Garrett Bingham, Evan Rosen, Lisa Wang, Qijun Tan, David Engel, Francesco Pongetti, Dario de Cesare, Dongseong Hwang, Lily Yu, Jennifer Pullman, Srini Narayanan, Kyle Levin, Siddharth Gopal, Megan Li, Asaf Aharoni, Trieu Trinh, Jessica Lo, Norman Casagrande, Roopali Vij, Loic Matthey, Bramandia Ramadhana, Austin Matthews, CJ Carey, Matthew Johnson, Kremena Goranova, Rohin Shah, Shereen Ashraf, Kingshuk Dasgupta, Rasmus Larsen, Yicheng Wang, Manish Reddy Vuyyuru, Chong Jiang, Joana Ijazi, Kazuki Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemniy, Kiam Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadsy, Prakash Shroff, Inderjit Dhillon, Tejasi Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jiageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush Dubey, Abhi Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang Nguyen, Ji Liu, Madhavi Sewak, Bryce Petrini, DongHyun Choi, Ivan Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chakladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya Samangouei, Riham Mansour, Tomasz Kępa, François-Xavier Aubet, Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Merey, Martin Baeuml, Trevor Strohman, Junwen Bai, Slav Petrov, Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Pengyu Wang, Linyang Li, Ke Ren, Botian Jiang, Dong Zhang, and Xipeng Qiu. 2023. SeqXGPT: Sentence-Level AI-Generated Text Detection. *arXiv preprint*. ArXiv:2310.08903 [cs].

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. Same Author or Just Same Topic? Towards Content-Independent Style Representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 249–268, Dublin, Ireland. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceed-*

ings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1112–1122. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800.