Watermark Mimicry in Large Language Models and Humans

Anonymous ACL submission

Abstract

Recent advancements in Large Language Models (LLMs) raised concerns over potential misuse, such as for spreading misinformation. In response two counter measures emerged: machine learning-based detectors that predict if text is synthetic, and LLM watermarking, which subtly marks generated text for identification and attribution. Meanwhile, humans are known to adjust language to their conversational partners both syntactically and lexically. By implication, it is possible that humans or unwatermarked LLMs could unintentionally mimic properties of LLM generated text, making counter measures unreliable. In this work we investigate the extent to which such conversational adaptation happens. We call the concept *mimicry* and demonstrate that both humans and LLMs end up mimicking, including the watermarking signal even in seemingly improbable settings. This challenges current academic assumptions and suggests that for long-term watermarking to be reliable, the likelihood of false positives needs to be significantly lower, while longer word sequences should be used for seeding watermarking mechanisms.

1 Introduction

Recent progress in Large Language Models (LLMs) enabled a large number of applications (Brown et al., 2020; Gemini-Team et al., 2024), but also raised concerns about their potential misuse, particularly for spreading misinformation (OpenAI, 2024) and impacting future machine learning (ML) models (Shumailov et al., 2023; Wyllie et al., 2024). To address this, ML-based detection for artificially produced text (Mitchell et al., 2023; Hans et al., 2024) and LLM watermarking have emerged as a promising techniques (Aaronson, 2022; Dathathri et al., 2024). ML-based detectors work by estimating the likelihood that given text was produced by a human. Watermarks enable more precise detection (Kirchenbauer et al., 2023; Aaronson and Kirchner, 2023) and even attribution of the source (Yoo et al., 2024) by subtly embedding a unique signature within generated content.

Current LLM watermarking schemes rely on manipulating the model's internal sampling process to favour specific sequences of tokens in certain contexts (Aaronson, 2022; Kirchenbauer et al., 2023; Christ et al., 2023). This allows for watermark detection by analysing the likelihood of a particular token appearing in a specific context, given a secret key. These methods achieve high watermark detection rates with controlable distortion to the generated text, leveraging the inherent statistical properties of language itself.

However, communication is a dynamic process. Humans, for example, adapt their language use – both in terms of vocabulary and sentence structure – to match their conversation partner (Chang et al., 2012; Mol et al., 2012). This phenomenon naturally leads to the question: *can the same adaptation occur during interactions with LLMs and influence watermarks and ML detectors*? While LLM watermarking leverages lexical and syntactic features, this very reliance creates a potential vulnerability. Namely, a non-watermarked conversational partner might begin to unconsciously mimic artificial language and the watermarking patterns over time, even though the partner is not watermarked.

In this work, we introduce the concept of *mimicry* and investigate the extent to which humans and LLMs exhibit this behaviour. We demonstrate that through extended conversations, both humans and LLMs exhibit *mimicry*, even for the watermarks in settings where the detection probability is low according to current academic watermarking configurations. We depict the intuition in Figure 1. This finding challenges the long-term effectiveness of current LLM watermarking configurations and suggests the need for exploring alternatives that take *mimicry* into account and provide lower false positive rates. Fundamentally, *water-*

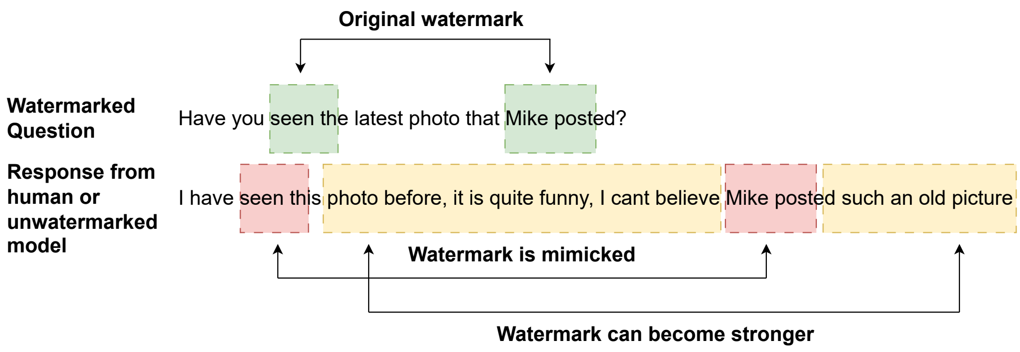

Figure 1: An intuitive description of *watermark mimicry*. Here, a watermarked prompt is used together with an unwatermarked model. During a conversation parts of the original watermerk (green) prompt are reused by the model, leading to watermark mimicry (red), resulting in an unwatermarked model outputting watermarked response. Importantly, the watermark can even be stronger in the response, since it can by a coincidence produce a watermark in unaffected by mimicry areas (yellow).

mark mimicry undermines the purpose of watermarking, since unwatermaked models are not supposed to produce watermarks.

2 Related Work

LLM Watermarking Model watermarking emerged as a potential first party solution for attribution of synthetically generated text to its origins. Such algorithms are broadly split into distortionary (Kirchenbauer et al., 2023) and non-distortianary (Aaronson and Kirchner, 2023; Christ et al., 2023), depending on whether they change the underlying model performance. Major tech companies reportedly now watermark their models (Aaronson, 2022; Dathathri et al., 2024). Importantly, current literature notes that learning from watermarked text can lead to production of watermarks (Gu et al., 2024; Sander et al., 2024). In this work, we show that even in-context interaction with LLMs leads to language adjustment and mimicry of the LLM watermarks.

Third-party detectors Alternative third party methods for synthetic data detection have also emerged. These specifically develop a machine learning model to estimate the probability that a given text sample was produced by a human (Hans et al., 2024; Mitchell et al., 2023). Binoculars from Hans et al. (2024) is the current state of the art for detection of artificial content, which leverages a pair of differently tuned Falcon-7B (Penedo et al., 2023) models for its calibrated prediction.

Conversational Adaptation Linguistic adaptation refers to a phenomena where individuals adjust their language in response to others during a conversation (Chang et al., 2012; Mol et al., 2012). Such adaptation is extensive and covers phonology, repetition of words, syntax, and even gesturing.

3 Methodology

In this paper we demonstrate that linguistic adaptation causes both humans and models to *mimic watermarks* and artificiality of generated text. To show the effect we run two main experiments. First, we show that **unwatermarked** model when interacting with the **watermarked** model mimics and starts disproportionately producing the watermark. For watermarking we use two most popular schemes in the current literature: Kirchenbauer et al. (2023) and Aaronson (2022), both available through ThreeBricks (Fernandez et al., 2023). Second, we show that a similar effect takes place in conversations with humans. Here we focus on estimating the probability that text is produced by a machine using the state-of-the-art synthetic data detector. We evaluate two large publicly available datasets of multiturn human–LLM conversations. Experiments are ran on 4 $\times$ Quadro RTX 8000 and took approximately a month.

4 Evaluation

4.1 LLMs conversing with LLMs

In this subsection we focus on the conversation between a pair of models – one watermarked with either Kirchenbauer et al. (2023) or Aaronson and Kirchner (2023) schemes; while the second model uses no watermark. We utilise Guanco-7B, Guanco-

Unwatermarked Guanco	#R	Aaronson and Kirchner (2023) Prompt ↑	Response ↓	P&R	P<R	Kirchenbauer et al. (2023) Prompt ↑	Response ↓	P&R	P<R
Baselines: Unwatermarked Prompt, Unwatermarked Response									
Theoretical	1	1%	1%	0.01%		1%	1%	0.01%	
7b v 7b	1	0.8%	0.9%	0.0%	0.0%	0.2%	0.8%	0.0%	0.0%
	3	0.7%	0.6%	0.0%	0.0%	1.6%	0.9%	0.4%	0.2%
	5	1.2%	0.5%	0.0%	0.0%	2.2%	1.7%	0.3%	0.3%
13b v 13b	1	1.0%	0.6%	0.2%	0.0%	0.4%	0.6%	0.1%	0.0%
	3	1.7%	0.5%	0.3%	0.0%	0.1%	1.1%	0.0%	0.0%
	5	1.4%	0.5%	0.3%	0.0%	0.2%	0.4%	0.0%	0.0%
Watermarked Prompt, Unwatermarked Response									
7b (w) v 7b	1	49.9%	3.7%	2.9%	0.3%	82.4%	6.3%	6.0%	0.4%
	3	72.0%	5.3%	4.9%	0.8%	96.7%	10.0%	10.0%	0.7%
	5	81.1%	6.6%	6.4%	0.7%	98.9%	13.0%	13.0%	0.7%
7b (w) v 13b	1	63.4%	3.4%	2.9%	1.1%	83.8%	7.6%	7.4%	0.1%
	3	80.0%	4.7%	4.5%	0.1%	97.7%	14.4%	14.1%	0.4%
	5	89.3%	6.0%	6.0%	0.4%	99.3%	18.2%	18.2%	0.4%
13b (w) v 7b	1	33.2%	3.9%	3.1%	0.4%	74.4%	4.6%	4.1%	0.4%
	3	54.1%	6.7%	6.2%	1.1%	93.0%	7.4%	7.3%	0.7%
	5	69.0%	8.4%	7.9%	0.5%	98.0%	10.2%	10.2%	0.8%
13b (w) v 13b	1	30.6%	3.2%	2.0%	0.3%	73.1%	5.8%	5.6%	0.4%
	3	53.2%	6.3%	5.2%	0.7%	90.5%	8.3%	8.3%	0.4%
	5	63.1%	8.1%	7.2%	0.9%	94.7%	11.2%	11.0%	0.4%

Table 1: Table shows proportion of text that has a watermark. Here a watermarked LLM model (Temperature=1, watermark ngram=4, threshold 0.01, averaged over three random keys (large variance)) interacts with a non-watermarked LLM model (Temperature=0.8). **P**rompt refers to the *watermarked prompt* from a first model. **R**esponse refers to the *response of an unwatermarked model* to a watermarked prompt. **#R** refers to a number of responses concatenated together. **P&R** reports a proportion of cases where watermarked prompt – i.e. watermark score < 0.01 – results in a watermarked response – watermark score < 0.01. **P<R** reports the proportion of cases where watermark of the response from an unwatermarked model is stronger than the watermark in the prompt. Original conversation prompt is listed in Appendix B. Expanded table in presented in Table 2 in Appendix.

13B (Dettmers et al., 2023) models and run the watermarking model in three temperature configurations T=[0, 0.5, 1]. We show examples of conversations in Appendix C and the prompt in Appendix B.

Table 1 shows the effects of *watermark mimicry*. The flow of conversations are shown in Figures 4 and 5 for Aaronson and Kirchner (2023) and in Figures 6 and 7 for Kirchenbauer et al. (2023). We find that for both schemes mimicry appears, albeit the effect is stronger for Kirchenbauer et al.. For example, with Aaronson scheme Guanco 13b unwatermarked model outputs watermarked response in 4% of cases when used with temperature 1, when computed over 5 consecutive responses. While in the same setting, Kirchenbauer et al. scheme outputs a watermarked response 12.9% of time.

4.2 Third party detection of generated data

In this section we focus on a setting where a human has a conversation with an LLM. Here we make no assumptions about existence of watermarking and focus on the best publicly available third-party detector Binoculars (Hans et al., 2024). We reuse the best parameters from Hans et al. and calibrate for 1% false positives rates. We use Binoculars to estimate if during the conversation with an LLMs humans adjust their language and 'mimic' the LLM, i.e. human language gets identified as LLM.

LLM-Human conversations: Here we evaluate conversations dataset available through huggingface `ar852/scraped-chatgpt-conversations`, `theblackcat102/sharegpt-english`, further fil-

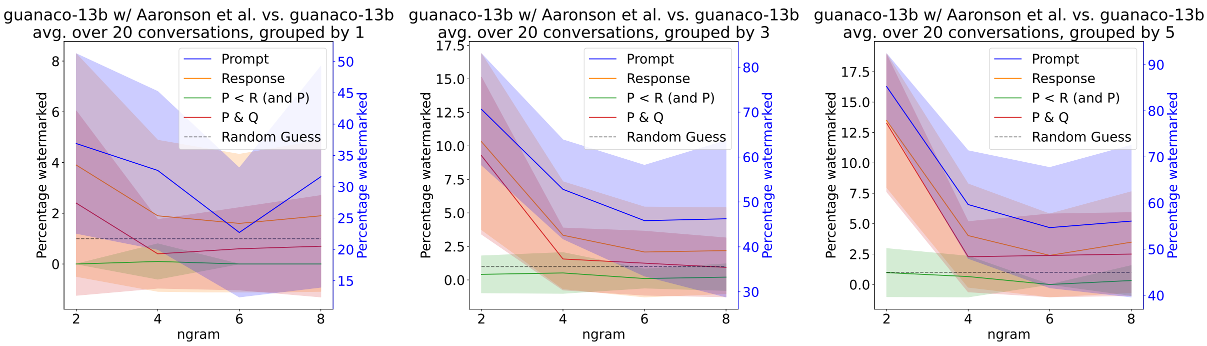

Figure 2: Aaronson (2022) scheme, changing ngram size, blue shows percentage of watermarked prompts, orange shows percentage of watermarked responses green shows percentage of watermarked responses where response watermark is stronger than in the prompt, red shows percentage of cases with both prompt and response watermarked.

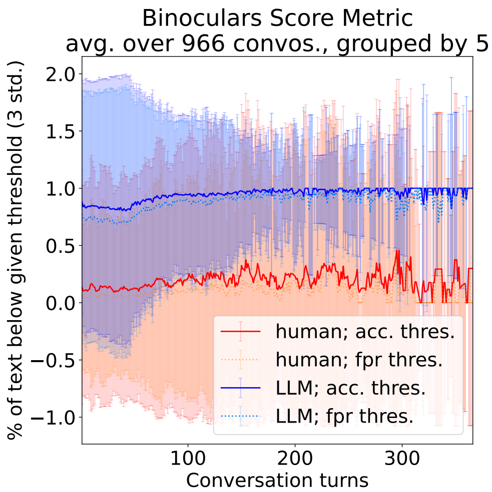

Figure 3: Human–LLM dialogues (split in Figure 12), filtered to contain long conversations in English. 520 are from ShareGPT, filtered for 100+ turns (100 human, 100 LLM). 446 are from WildChat dataset, filtered for 50+ turns (100 human, 100 LLM).

tered to only keep conversations with at least 100 turns. Figure 3 shows the results. We find that during the conversations humans adjust their language and variance of scores increases over the course of the dialogue. In other words, over a dialogue humans tend to produce more speech that appears LLM-like for Binoculars. The opposite effect takes place for the LLM, where variance is lost over time.

5 Discussion

In this paper we demonstrated that LLM-produced input influences how language is used by unwatermarked models and humans. Our results have methodological as well as practical implications. First, methodologically we show that for watermarking it is important to account for linguistic adaption when evaluating performance. Particularly so when the prompts used for evaluations were generated by other models. One of the ways to deal with mimicry is to incorporate repetition masking that is used in some watermarking mechanisms already (Hu et al., 2023), extending its use to the prompts. Second, we need to assume stronger detection thresholds and more realistic configurations for watermarking. This means adjusting false positive rates to be lower and, for example, choosing ngram lengths that result in less robust watermarks. Figure 2 shows how amount of watermark mimicry for Aaronson and Kirchner (2023) changes as a function of ngram. We show more results in Appendix Figures 8 to 11 Here, we observe that as ngrams get longer mimicry happens less, yet it comes at a cost of less watermarking performance.

6 Conclusion

Our work highlights a critical challenge for LLM provenance – *mimicry*. We demonstrate that both humans and LLMs exhibit adaptation during conversations, potentially replicating language artificiality and surprisingly even the watermark signal. This undermines the effectiveness of current watermarking schemes that rely on easily detectable deviations in token probabilities. To ensure the long-term viability of LLM watermarking, we propose further research avenues. First, investigating watermarking techniques that provide significantly lower false positive rates is crucial. Second, exploring alternative watermarking methods that leverage properties less susceptible to mimicry, such as semantic coherence or stylistic elements. By addressing these challenges, we can ensure that LLM watermarking remains a valuable tool for long-term mitigation of the misuse of LLMs.

7 Limitations

In this section we discuss limitations of the evaluation above. First, our work only considers two of the most popular schemes and it is not clear if all other watermarking mechanisms will exhibit mimicry. Intuitively we expect it to still happen, consider the following pathological example. Imagine an unwatermarked model quoting watermarked text from the input – in this case as long as quotation happens it should always trigger the watermark detector. Yet, the example above provides no indication to the degree of unnoticeable mimicry. Second, our work only considered watermarks used with three keys and computing expectation over more keys may reduce currently large performance variance. We still nonetheless believe that mimicry will be exhibited, based on the examples and the shear magnitude of mimicry observed in our experiments with the three random keys. Third, we only considered models of size 7b and 13b, while larger models may exhibit less mimicry and generalise to text better. It can clearly be seen in the examples in Appendix C, where the text is comprehensible, yet the quality is not on par with the best currently available models. This choice is motivated by limited resources, current experiments already took more than a month to run, expanding evaluation to larger models will further increased the resource cost, whilst providing little further understanding of the phenomena. Fourth, evaluation in Section 4.2 assumes that data available through public datasets is reliable to make adaptation judgements. Fifth, following from the previous point, it is not clear to what degree mimicry affects all human conversations, and not just specific type of prolonged discussions. If discussions are short and only a few sentences are exchanges, it is not obvious that mimicry would always happen. More thorough evaluation is required to find precise type of linguistic adaptation that enables watermark mimicry.

8 Ethical Considerations

Our work addresses reliability of text watermarking mechanisms – an important tool that enables trustworthy machine learning. We find that to make well informed text provenance decisions, *mimicry* and linguistic adaptation should be taken into account. We highlight an inherent limitation of text watermarking and suggest future venues for addressing the discovered limitation. We view our work as an important step towards making text watermarking more reliable, while also making methodologies of watermark evaluations more transparent and fair.

References

Scott Aaronson. 2022. My ai safety lecture for ut effective altruism. *Shtetl-Optimized: The blog of Scott Aaronson. Retrieved on September*, 11:2023.

Scott Aaronson and Hendrik Kirchner. 2023. Watermarking gpt outputs.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Franklin Chang, Marius Janciauskas, and Hartmut Fitz. 2012. Language adaptation and learning: Getting explicit about implicit learning. *Language and Linguistics Compass*, 6(5):259–278.

Miranda Christ, Sam Gunn, and Or Zamir. 2023. Undetectable watermarks for language models. *Preprint*, arXiv:2306.09194.

Sumanth Dathathri, Pushmeet Kohli, Vandana Bachani, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Abi See, and Johannes Welbl. 2024. Watermarking ai-generated text and video with synthid. *Google DeepMind blog*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Pierre Fernandez, Antoine Chaffin, Karim Tit, Vivien Chappelier, and Teddy Furon. 2023. Three bricks to consolidate watermarks for large language models. In *2023 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE.

Gemini-Team Gemini-Team, Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry, Lepikhin, Timothy Lillicrap, Jean baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand,

Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, Luke Vilnis, Oscar Chang, Nobuyuki Morioka, George Tucker, Ce Zheng, Oliver Woodman, Nithya Attaluri, Tomas Kocisky, Evgenii Eltyshev, Xi Chen, Timothy Chung, Vittorio Selo, Siddhartha Brahma, Petko Georgiev, Ambrose Slone, Zhenkai Zhu, James Lottes, Siyuan Qiao, Ben Caine, Sebastian Riedel, Alex Tomala, Martin Chadwick, Juliette Love, Peter Choy, Sid Mittal, Neil Houlsby, Yunhao Tang, Matthew Lamm, Libin Bai, Qiao Zhang, Luheng He, Yong Cheng, Peter Humphreys, Yujia Li, Sergey Brin, Albin Cassirer, Yingjie Miao, Lukas Zilka, Taylor Tobin, Kelvin Xu, Lev Proleev, Daniel Sohn, Alberto Magni, Lisa Anne Hendricks, Isabel Gao, Santiago Ontanon, Oskar Bunyan, Nathan Byrd, Abhanshu Sharma, Biao Zhang, Mario Pinto, Rishika Sinha, Harsh Mehta, Dawei Jia, Sergi Caelles, Albert Webson, Alex Morris, Becca Roelofs, Yifan Ding, Robin Strudel, Xuehan Xiong, Marvin Ritter, Mostafa Dehghani, Rahma Chaabouni, Abhijit Karmarkar, Guangda Lai, Fabian Mentzer, Bibo Xu, YaGuang Li, Yujing Zhang, Tom Le Paine, Alex Goldin, Behnam Neyshabur, Kate Baumli, Anselm Levskaya, Michael Laskin, Wenhao Jia, Jack W. Rae, Kefan Xiao, Antoine He, Skye Giordano, Lakshman Yagati, Jean-Baptiste Lespiau, Paul Natsev, Sanjay Ganapathy, Fangyu Liu, Danilo Martins, Nanxin Chen, Yunhan Xu, Megan Barnes, Rhys May, Arpi Vezer, Junhyuk Oh, Ken Franko, Sophie Bridgers, Ruizhe Zhao, Boxi Wu, Basil Mustafa, Sean Sechrist, Emilio Parisotto, Thanumalayan Sankaranarayana Pillai, Chris Larkin, Chenjie Gu, Christina Sorokin, Maxim Krikun, Alexey Guseynov, Jessica Landon, Romina Datta, Alexander Pritzel, Phoebe Thacker, Fan Yang, Kevin Hui, Anja Hauth, Chih-Kuan Yeh, David Barker, Justin Mao-Jones, Sophia Austin, Hannah Sheahan, Parker Schuh, James Svensson, Rohan Jain, Vinay Ramasesh, Anton Briukhov, Da-Woon Chung, Tamara von Glehn, Christina Butterfield, Priya Jhakra, Matthew Wiethoff, Justin Frye, Jordan Grimstad, Beer Changpinyo, Charline Le Lan, Anna Bortsova, Yonghui Wu, Paul Voigtlaender, Tara Sainath, Shane Gu, Charlotte Smith, Will Hawkins, Kris Cao, James Besley, Srivatsan Srinivasan, Mark Omernick, Colin Gaffney, Gabriela Surita, Ryan Burnell, Bogdan Damoc, Junwhan Ahn, Andrew Brock, Mantas Pajarskas, Anastasia Petrushkina, Seb Noury, Lorenzo Blanco, Kevin Swersky, Arun Ahuja, Thi Avrahami, Vedant Misra, Raoul de Liedekerke, Mariko Iinuma, Alex Polozov, Sarah York, George van den Driessche, Paul Michel, Justin Chiu, Rory Blevins, Zach Gleicher, Adrià Recasens, Alban Rrustemi, Elena Gribovskaya, Aurko Roy, Wiktor Gworek, Sébastien M. R. Arnold, Lisa Lee, James Lee-Thorp, Marcello Maggioni, Enrique Piqueras, Kartikeya Badola, Sharad Vikram, Lucas Gonzalez, Anirudh Baddepudi, Evan Senter, Jacob Devlin, James Qin, Michael Azzam, Maja Trebacz, Martin Polacek, Kashyap Krishnakumar, Shuo yiin Chang, Matthew Tung, Ivo Penchev, Rishabh Joshi, Kate Olszewska, Carrie Muir, Mateo Wirth, Ale Jakse Hartman, Josh Newlan, Sheleem Kashem, Vijay Bolina, Elahe Dabir, Joost van Amersfoort, Zafarali Ahmed, James Cobon-Kerr, Aishwarya Kamath, Arnar Mar Hrafnkelsson, Le Hou, Ian Mackinnon, Alexandre Frechette, Eric Noland, Xiance Si, Emanuel Taropa, Dong Li, Phil Crone, Anmol Gulati, Sébastien Cevey, Jonas Adler, Ada Ma, David Silver, Simon Tokumine, Richard Powell, Stephan Lee, Kiran Vodrahalli, Samer Hassan, Diana Mincu, Antoine Yang, Nir Levine, Jenny Brennan, Mingqiu Wang, Sarah Hodkinson, Jeffrey Zhao, Josh Lipschultz, Aedan Pope, Michael B. Chang, Cheng Li, Laurent El Shafey, Michela Paganini, Sholto Douglas, Bernd Bohnet, Fabio Pardo, Seth Odoom, Mihaela Rosca, Cicero Nogueira dos Santos, Kedar Soparkar, Arthur Guez, Tom Hudson, Steven Hansen, Chulayuth Asawaroengchai, Ravi Addanki, Tianhe Yu, Wojciech Stokowiec, Mina Khan, Justin Gilmer, Jaehoon Lee, Carrie Grimes Bostock, Keran Rong, Jonathan Caton, Pedram Pejman, Filip Pavetic, Geoff Brown, Vivek Sharma, Mario Lučić, Rajkumar Samuel, Josip Djolonga, Amol Mandhane, Lars Lowe Sjösund, Elena Buchatskaya, Elspeth White, Natalie Clay, Jiepu Jiang, Hyeontaek Lim, Ross Hemsley, Zeyncep Cankara, Jane Labanowski, Nicola De Cao, David Steiner, Sayed Hadi Hashemi, Jacob Austin, Anita Gergely, Tim Blyth, Joe Stanton, Kaushik Shivakumar, Aditya Siddhant, Anders Andreassen, Carlos Araya, Nikhil Sethi, Rakesh Shivanna, Steven Hand, Ankur Bapna, Ali Khodaei, Antoine Miech, Garrett Tanzer, Andy Swing, Shantanu Thakoor, Lora Aroyo, Zhufeng Pan, Zachary Nado, Jakub Sygnowski, Stephanie Winkler, Dian Yu, Mohammad Saleh, Loren Maggiore, Yamini Bansal, Xavier Garcia, Mehran Kazemi, Piyush Patil, Ishita Dasgupta, Iain Barr, Minh Giang, Thais Kagohara, Ivo Danihelka, Amit Marathe, Vladimir Feinberg, Mohamed Elhawaty, Nimesh Ghelani, Dan Horgan, Helen Miller, Lexi Walker, Richard Tanburn, Mukarram Tariq, Disha Shrivastava, Fei Xia, Qingze Wang, Chung-Cheng Chiu, Zoe Ashwood, Khuslen Baatarsukh, Sina Samangooei, Raphaël Lopez Kaufman, Fred Alcober, Axel Stjerngren, Paul Komarek, Katerina Tsihlas, Anudhyan Boral, Ramona Comanescu, Jeremy Chen, Ruibo Liu, Chris Welty, Dawn Bloxwich, Charlie Chen, Yanhua Sun, Fangxiaoyu Feng, Matthew Mauger, Xerxes Dotiwalla, Vincent Hellendoorn, Michael Sharman, Ivy Zheng, Krishna Haridasan, Gabe Barth-Maron, Craig Swanson, Dominika Rogozińska, Alek Andreev, Paul Kishan Rubenstein, Ruoxin Sang, Dan Hurt, Gamaleldin Elsayed, Renshen Wang, Dave Lacey, Anastasija Ilić, Yao Zhao, Adam Iwanicki, Alejandro Lince, Alexander Chen, Christina Lyu, Carl Lebsack, Jordan Griffith, Meenu Gaba, Paramjit Sandhu, Phil Chen, Anna Koop, Ravi Rajwar, Soheil Hassas Yeganeh, Solomon Chang, Rui Zhu, Soroush Radpour, Elnaz Davoodi, Ving Ian Lei, Yang Xu, Daniel Toyama, Constant Segal, Martin Wicke, Hanzhao Lin, Anna Bulanova, Adrià Puigdomènech Badia, Nemanja Rakićević, Pablo Sprechmann, Angelos Filos, Shaobo Hou, Víctor Campos, Nora Kassner, Devendra Sachan, Meire Fortunato, Chimezie Iwuanyanwu, Vitaly Nikolaev, Balaji Lakshminarayanan, Sadegh Jazayeri, Mani Varadarajan, Chetan Tekur, Doug Fritz, Misha Khalman, David

Reitter, Kingshuk Dasgupta, Shourya Sarcar, Tina Ornduff, Javier Snaider, Fantine Huot, Johnson Jia, Rupert Kemp, Nejc Trdin, Anitha Vijayakumar, Lucy Kim, Christof Angermueller, Li Lao, Tianqi Liu, Haibin Zhang, David Engel, Somer Greene, Anaïs White, Jessica Austin, Lilly Taylor, Shereen Ashraf, Dangyi Liu, Maria Georgaki, Irene Cai, Yana Kulizhskaya, Sonam Goenka, Brennan Saeta, Ying Xu, Christian Frank, Dario de Cesare, Brona Robenek, Harry Richardson, Mahmoud Alnahlawi, Christopher Yew, Priya Ponnapalli, Marco Tagliasacchi, Alex Korchemniy, Yelin Kim, Dinghua Li, Bill Rosgen, Kyle Levin, Jeremy Wiesner, Praseem Banzal, Praveen Srinivasan, Hongkun Yu, Çağlar Ünlü, David Reid, Zora Tung, Daniel Finchelstein, Ravin Kumar, Andre Elisseeff, Jin Huang, Ming Zhang, Ricardo Aguilar, Mai Giménez, Jiawei Xia, Olivier Dousse, Willi Gierke, Damion Yates, Komal Jalan, Lu Li, Eri Latorre-Chimoto, Duc Dung Nguyen, Ken Durden, Praveen Kallakuri, Yaxin Liu, Matthew Johnson, Tomy Tsai, Alice Talbert, Jasmine Liu, Alexander Neitz, Chen Elkind, Marco Selvi, Mimi Jasarevic, Livio Baldini Soares, Albert Cui, Pidong Wang, Alek Wenjiao Wang, Xinyu Ye, Krystal Kallarackal, Lucia Loher, Hoi Lam, Josef Broder, Dan Holtmann-Rice, Nina Martin, Bramandia Ramadhana, Mrinal Shukla, Sujoy Basu, Abhi Mohan, Nick Fernando, Noah Fiedel, Kim Paterson, Hui Li, Ankush Garg, Jane Park, DongHyun Choi, Diane Wu, Sankalp Singh, Zhishuai Zhang, Amir Globerson, Lily Yu, John Carpenter, Félix de Chaumont Quitry, Carey Radebaugh, Chu-Cheng Lin, Alex Tudor, Prakash Shroff, Drew Garmon, Dayou Du, Neera Vats, Han Lu, Shariq Iqbal, Alex Yakubovich, Nilesh Tripuraneni, James Manyika, Haroon Qureshi, Nan Hua, Christel Ngani, Maria Abi Raad, Hannah Forbes, Jeff Stanway, Mukund Sundararajan, Victor Ungureanu, Colton Bishop, Yunjie Li, Balaji Venkatraman, Bo Li, Chloe Thornton, Salvatore Scellato, Nishesh Gupta, Yicheng Wang, Ian Tenney, Xihui Wu, Ashish Shenoy, Gabriel Carvajal, Diana Gage Wright, Ben Bariach, Zhuyun Xiao, Peter Hawkins, Sid Dalmia, Clement Farabet, Pedro Valenzuela, Quan Yuan, Ananth Agarwal, Mia Chen, Wooyeol Kim, Brice Hulse, Nandita Dukkipati, Adam Paszke, Andrew Bolt, Kiam Choo, Jennifer Beattie, Jennifer Prendki, Harsha Vashisht, Rebeca Santamaria-Fernandez, Luis C. Cobo, Jarek Wilkiewicz, David Madras, Ali Elqursh, Grant Uy, Kevin Ramirez, Matt Harvey, Tyler Liechty, Heiga Zen, Jeff Seibert, Clara Huiyi Hu, Andrey Khorlin, Maigo Le, Asaf Aharoni, Megan Li, Lily Wang, Sandeep Kumar, Norman Casagrande, Jay Hoover, Dalia El Badawy, David Soergel, Denis Vnukov, Matt Miecnikowski, Jiri Simsa, Praveen Kumar, Thibault Sellam, Daniel Vlasic, Samira Daruki, Nir Shabat, John Zhang, Guolong Su, Jiageng Zhang, Jeremiah Liu, Yi Sun, Evan Palmer, Alireza Ghaffarkhah, Xi Xiong, Victor Cotruta, Michael Fink, Lucas Dixon, Ashwin Sreevatsa, Adrian Goedeckemeyer, Alek Dimitriev, Mohsen Jafari, Remi Crocker, Nicholas FitzGerald, Aviral Kumar, Sanjay Ghemawat, Ivan Philips, Frederick Liu, Yannie Liang, Rachel Sterneck, Alena Repina, Marcus Wu, Laura Knight, Marin Georgiev, Hyo Lee, Harry Askham, Abhishek Chakladar, Annie Louis, Carl Crous, Hardie Cate, Dessie Petrova, Michael Quinn, Denese Owusu-Afriyie, Achintya Singhal, Nan Wei, Solomon Kim, Damien Vincent, Milad Nasr, Christopher A. Choquette-Choo, Reiko Tojo, Shawn Lu, Diego de Las Casas, Yuchung Cheng, Tolga Bolukbasi, Katherine Lee, Saaber Fatehi, Rajagopal Ananthanarayanan, Miteyan Patel, Charbel Kaed, Jing Li, Shreyas Rammohan Belle, Zhe Chen, Jaclyn Konzelmann, Siim Põder, Roopal Garg, Vinod Koverkathu, Adam Brown, Chris Dyer, Rosanne Liu, Azade Nova, Jun Xu, Alanna Walton, Alicia Parrish, Mark Epstein, Sara McCarthy, Slav Petrov, 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.

Chenchen Gu, Xiang Lisa Li, Percy Liang, and Tatsunori Hashimoto. 2024. On the learnability of watermarks for language models. *Preprint*, arXiv:2312.04469.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting llms with binoculars: Zero-shot detection of machine-generated text. *Preprint*, arXiv:2401.12070.

Zhengmian Hu, Lichang Chen, Xidong Wu, Yihan Wu, Hongyang Zhang, and Heng Huang. 2023. Unbiased watermark for large language models. *Preprint*, arXiv:2310.10669.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17061–17084. PMLR.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *Preprint*, arXiv:2301.11305.

Lisette Mol, Emiel Krahmer, Alfons Maes, and Marc Swerts. 2012. Adaptation in gesture: Converging hands or converging minds? *Journal of Memory and Language*, 66(1):249–264.

OpenAI. 2024. Disrupting malicious uses of ai by state-affiliated threat actors. *OpenAI blog*.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only. *Preprint*, arXiv:2306.01116.

Tom Sander, Pierre Fernandez, Alain Durmus, Matthijs Douze, and Teddy Furon. 2024. Watermarking makes language models radioactive. *Preprint*, arXiv:2402.14904.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. The curse of recursion: Training on generated data makes models forget. *Preprint*, arXiv:2305.17493.

Sierra Wyllie, Ilia Shumailov, and Nicolas Papernot. 2024. Fairness feedback loops: Training on synthetic data amplifies bias. *ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT 2024)*.

KiYoon Yoo, Wonhyuk Ahn, and Nojun Kwak. 2024. Advancing beyond identification: Multi-bit watermark for large language models. *Preprint*, arXiv:2308.00221.

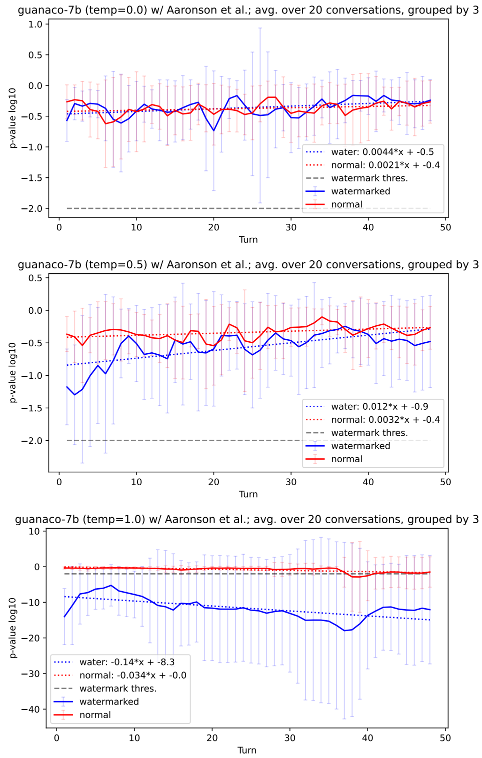

Figure 4: Aaronson. watermarking with Guanco-7b and varying temperatures

A Multi-turn Conversations

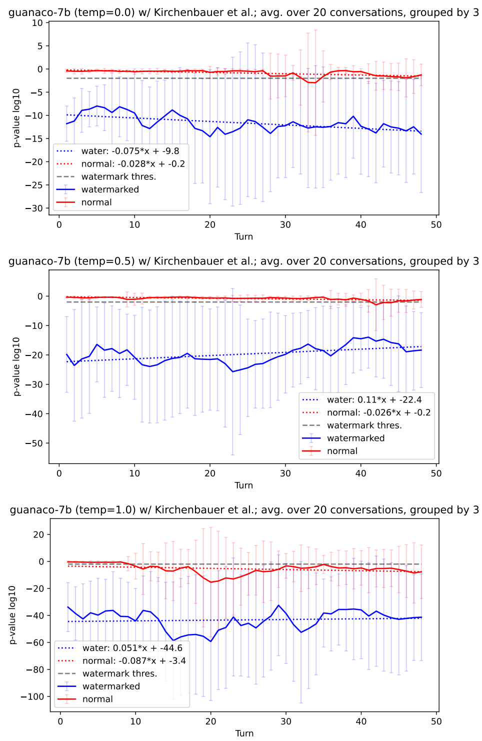

Figure 6: Kirch. watermarking with Guanco-7b and varying temperatures

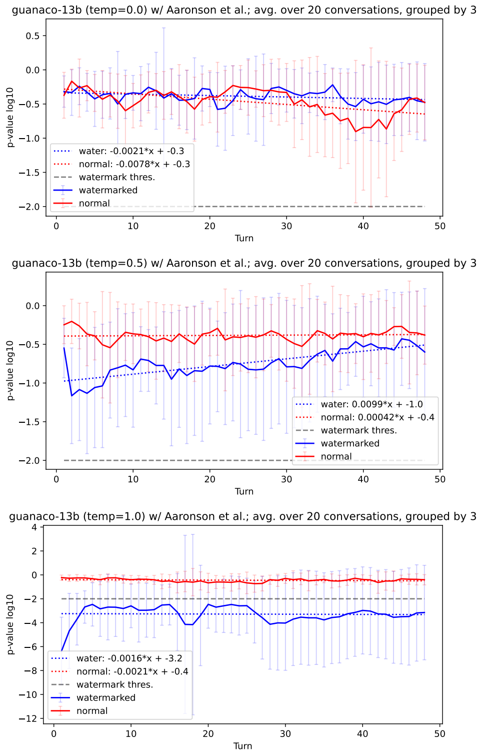

Figure 5: Aaronson. watermarking with Guanco-13b and varying temperatures

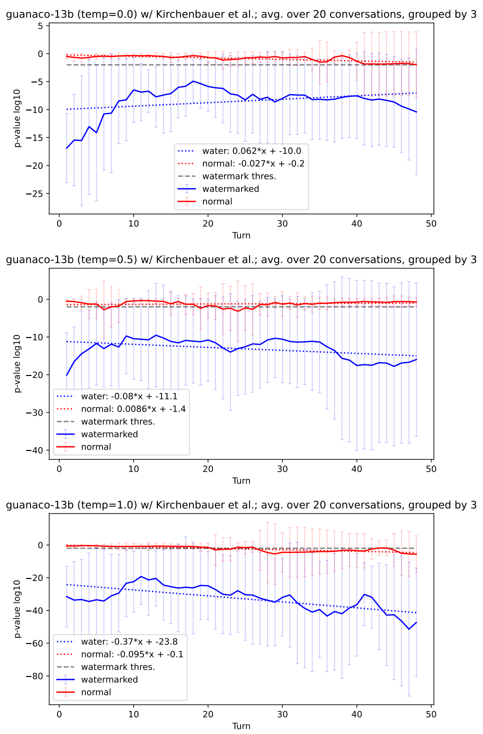

Figure 7: Kirch. watermarking with Guanco-13b and varying temperatures

B Original conversation prompt

```
promptstart = "A chat between two people
    talking about everyday topics such
    as Weather, Family, Work/School,
    Hobbies/Interests, Entertainment,
    Travel, Food/Cooking, Current Events
    , Technology, Health/Fitness, Pets,
    Relationships, Home/Living, Fashion/
    Style, Goals/Ambitions.\nThe two
    people listen to each other intently
     and carry the conversation
    naturally and they talk at length.
    Answer with at least 50 words each
    time.\n\n### Person A: Hey there!
    How's the weather treating you
    lately? It's been so unpredictable
    here. One day it's sunny and warm,
    and the next it's pouring rain.\n\n
    ### Person B: Oh, I know what you
    mean! The weather has been all over
    the place lately. It makes planning
    outdoor activities a bit tricky,
    doesn't it? I'm hoping for some more
     consistent sunshine soon so I can
    spend more time hiking and enjoying
    nature. How about you? Any outdoor
    plans in mind?"
```

C Watermarked conversation examples

Below are a number of examples where both prompt (from watermarked model) and the response (from unwatermarked model) are watermarked. In colour we highlight the tokens that are reused from **P**rompt (green) into the **R**esponse (blue).

P: The holidays are a great time of year for reflection, and i was thinking about my new Year's resolutions for next year. Sounds like a plan! I'm not a fan of resolutions, either, but I'm up for a challenge. hiking a new trail is always a blast.

R: That's a good idea! I'm also thinking about my resolutions for next year. I'm not a fan of making them, but I'm always up for a challenge. I'm not a fan of making new Year's resolutions, but I'm always up for a challenge. i was thinking about trying something new this year, and i was wondering if you had any ideas. hiking is one of my favorite things to do, but I'm not a fan of doing the same trail over and over again.

P: Sounds like a plan! I'm not a fan of resolutions, either, but I'm up for a challenge. hiking a new trail is always a

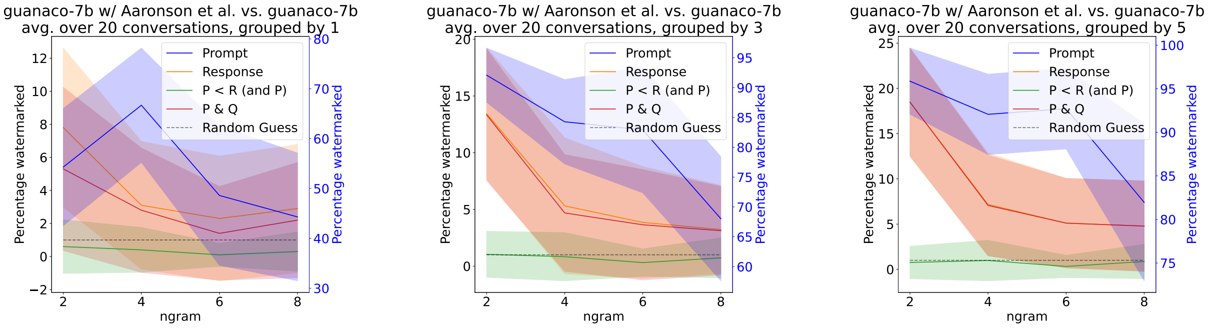

Figure 8: Aaronson (2022) with Guanco-7b (w) v 7b and varying ngrams, T=1

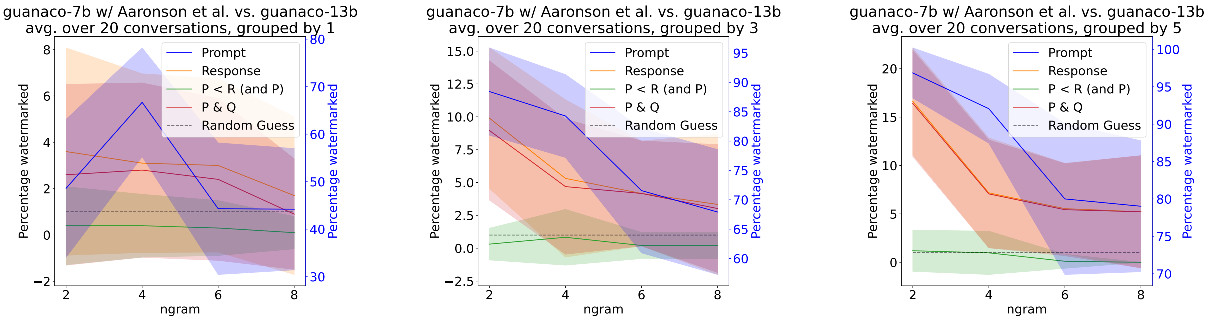

Figure 9: Aaronson (2022) with Guanco-7b (w) v 13b and varying ngrams, T=1

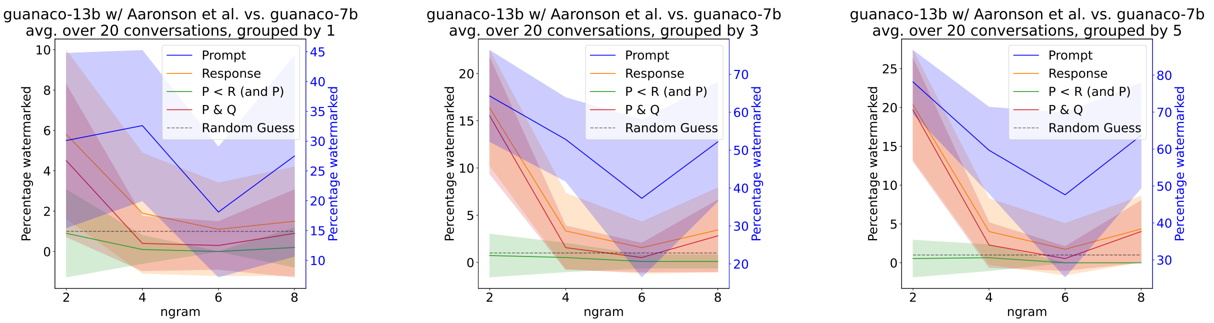

Figure 10: Aaronson (2022) with Guanco-13b (w) v 7b and varying ngrams, T=1

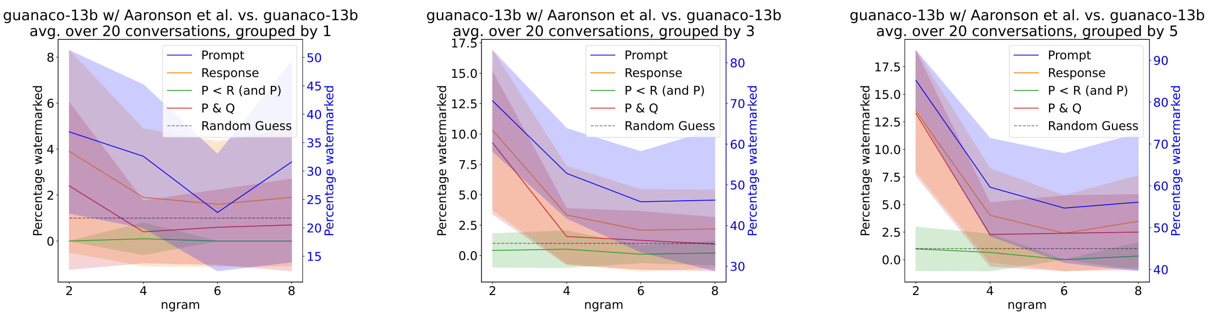

Figure 11: Aaronson (2022) watermarking with Guanco-13b (w) v 13b and varying ngrams, T=1

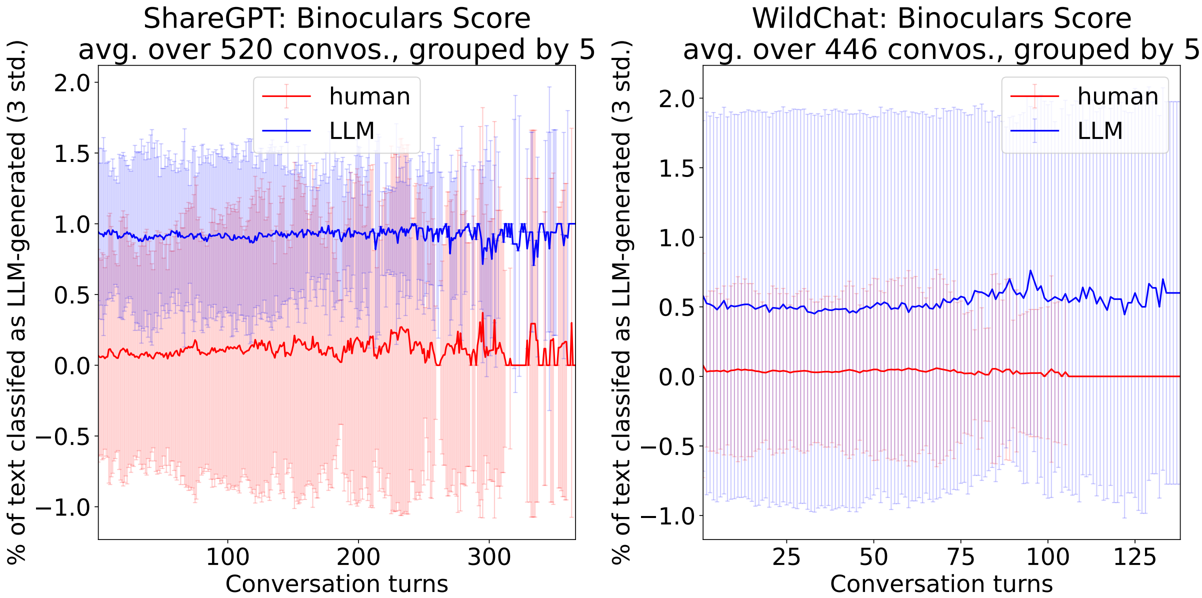

Figure 12: Datasets split with Binoculars scoring. Showing three standard deviations.

blast. Sounds like a plan! I'm not a fan of repetition, either, and I'm always up for a new adventure.

R: I'm not a fan of making new Year's resolutions, but I'm always up for a challenge. I was thinking about trying something new this year, and I was wondering if you had any ideas. hiking is one of my favorite things to do, but I'm not a fan of doing the same trail over and over again. I'm not a fan of repetition, either, and I'm always up for a new adventure. I was thinking about trying something new this year, and I was wondering if you had any ideas.

P: The weather is finally warming up here in [place]. I'm so excited to be able to go outside and enjoy the nice weather. I'm the same way. i love to keep up with the latest news and developments in [the field]. It's always nice to meet someone who shares your passions and interests. I'm so glad we're able to connect over our shared interest in [the field]. the election is right around the corner, and I'm so excited to see what happens. I'm so glad we're able to connect over our shared interest in politics.

R: I'm so excited for [the new thing]. I'm so glad we're able to connect over our shared interests. Me too. I'm so happy to have found someone who understands my love of [the field]. I'm the same way. i love to keep up with the latest news and developments in politics. It's always nice to meet someone who shares your passions and interests. I'm so glad we're able to connect over our shared interest in politics.

P: I'm the same way. i love to keep up with the latest news and developments in [the field]. It's always nice to meet someone who shares your passions and interests. I'm so glad we're able to connect over our shared interest in [the field]. the election is right around the corner, and I'm so excited to see what happens. I'm so glad we're able to connect over our shared interest in politics. the weather is finally warming up here in [place]. I'm so excited to be able to go outside and enjoy the nice weather.

R: Me too. I'm so happy to have found someone who understands my love of [the field]. I'm the same way. i love to keep up with the latest news and developments in politics. It's always nice to meet someone who shares your passions and interests. I'm so glad we're able to connect over our shared interest in politics. I'm so excited for [the new thing]. I'm so glad we're able to connect over our shared interests.

P: The weather is finally warming up here in [place]. I'm so excited to be able to go outside and enjoy the nice weather. The election is right around the corner, and I'm so excited to see what happens. I'm so glad we're able to connect over our shared interest in politics. [Topic] is so fascinating to me as well. I'm so glad we're able to connect over our shared interest in [topic].

R: I'm so excited for [the new thing]. I'm so glad we're able to connect over our shared interests. I'm also really into [topic]. I'm so glad we're able to connect over our shared interests. I'm so excited about [new thing]. I'm so glad we're able to connect over our shared interests.

P: The election is right around the corner, and I'm so excited to see what happens. I'm so glad we're able to connect over our shared interest in politics. [Topic] is so fascinating to me as well. I'm so glad we're able to connect over our shared interest in [topic]. The weather is finally warming up here in [place]. I'm so excited to be able to go outside and enjoy the nice weather.

R: I'm also really into [topic]. I'm so glad we're able to connect over our shared interests. I'm so excited about [new thing]. I'm so glad we're able to connect over our shared interests. [new topic] is so fascinating to me as well. I'm so glad we're able to connect over our shared interest in [topic].

P: The most important thing in my life is my family and the friends i consider my family. the most important thing in my life is my family and the friends i consider my family. the most important thing in my life is my family and the friends i consider my family.

R: I'm a big fan of soccer, especially the Premier League. I'm a supporter of Manchester United. my favorite player is Paul Pogba. I'm a big fan of anime. my favorite anime is Fullmetal Alchemist: Brotherhood. i want to travel the world. i want to see as many places as i can and experience different cultures.

P: The most important thing in my life is my family and the friends i consider my family. the most important thing in my life is my family and the friends i consider my family. the most important thing in my life is my family and the friends i consider my family.

R: I'm a big fan of anime. my favorite anime is Fullmetal Alchemist: Brotherhood. i want to travel the world. i want to see as many places as i can and experience different cultures. I'm a sucker for good-looking cars. my dream car is a Lamborghini Huracán.

P: (New topic) is a very important (field) in today's world. The weather is beautiful these days, don't you think? (New topic) is a very important (field) in today's world.

R: (New topic) has a huge impact on our society. I've been learning a lot about (new topic). It's so interesting! (New topic) is a very important (field) in today's world.

P: You know, as much as i love being outdoors, i also love taking time to just relax and unwind. I'm always telling my friends how much i love a nice glass of wine and a good book. I'm the same way. i love the holidays and I'm always so excited to see my family and friends. I'm the same way. i love to bake and I'm always looking forward to decorating the house for the holidays.

R: I'm so excited for the holidays! i love this time of year and I'm so looking forward to spending time with family and friends. Me too! I'm looking forward to baking cookies and decorating the house for the holidays. you know, as much as i love being outdoors, i also love taking time to just relax and unwind. I'm always telling my friends how much i love a nice glass of wine and a good book.

P: I'm the same way. i love to bake and I'm always looking forward to decorating the house for the holidays. You know, as much as i love being outdoors, i also love taking time to just relax and unwind. I'm always telling my friends how much i love a nice glass of wine and a good book. the weather is beautiful today! i was just thinking about all the fun summer activities i can't wait to do.

R: You know, as much as i love being outdoors, i also love taking time to just relax and unwind. I'm always telling my friends how much i love a nice glass of wine and a good book. I'm so excited for the holidays! i love this time of year and I'm so looking forward to spending time with family and friends. i know, right? It's like they're so happy to see us and they never tire of us being home.

P: You know, as much as i love being outdoors, i also love taking time to just relax and unwind. I'm always telling my friends how much i love a nice glass of wine and a good book. the weather is beautiful today! i was just thinking about all the fun summer activities i can't wait to do. the weather is beautiful today! i was just thinking about all the fun summer activities i can't wait to do.

R: I'm so excited for the holidays! i love this time of year and I'm so looking forward to spending time with family and friends. i know, right? It's like they're so happy to see us and they never tire of us being home. i know, right? It's like they're so happy to see us and they never tire of us being home.

P: Sounds like a great idea. I'll get you a new car. I'm so glad you're adopting a puppy. [topics: car, babysitter, gift] the babysitter told me they had a bumper car accident with our new car.

R: We're so excited! [topics: car, babysitter, gift] the tutor we hired was great. She really helped my son with his history. I'm so sorry to hear that. I hope no one was hurt. I'll get you a new car.

P: I'm so glad you're adopting a puppy. [topics: car, babysitter, gift] The babysitter told me they had a bumper car accident with our new car. The babysitter told me they had a bumper car accident with our new car.

R: The tutor we hired was great. She really helped my son with his history. I'm so sorry to hear that. I hope no one was hurt. I'll get you a new car. My husband and I are thinking of moving to a new house.

P: The babysitter told me they had a bumper car accident with our new car. The babysitter told me they had a bumper car accident with our new car. The babysitter told me they had a bumper car accident with our new car.

R: I'm so sorry to hear that. I hope no one was hurt. I'll get you a new car. My husband and I are thinking of moving to a new house. My daughter is taking a dance class, and she's really enjoying it.

D Expanded table of results

Unwatermarked			Aaronson and Kirchner (2023)				Kirchenbauer et al. (2023)			
Guanco	**T**	**#R**	**Prompt ↑**	**Response ↓**	**P&R**	**P<R**	**Prompt ↑**	**Response ↓**	**P&R**	**P<R**
Random			1%	1%	0.01%		1%	1%	0.01%	
7b (w) v 7b	0.0	1	0.6%	0.5%	0.0%	0.0%	68.5%	3.6%	3.2%	0.6%
	0.5	1	5.9%	1.6%	0.9%	0.1%	73.5%	6.8%	6.3%	0.5%
	1.0	1	49.9%	3.7%	2.9%	0.3%	82.4%	6.3%	6.0%	0.4%
	0.0	3	1.0%	0.8%	0.1%	0.0%	91.1%	7.3%	7.1%	1.0%
	0.5	3	15.3%	3.3%	2.5%	0.8%	93.1%	10.2%	9.9%	1.0%
	1.0	3	72.0%	5.3%	4.9%	0.8%	96.7%	10.0%	10.0%	0.7%
	0.0	5	1.4%	1.1%	0.2%	0.1%	95.4%	9.3%	9.1%	0.8%
	0.5	5	21.6%	4.0%	3.0%	0.9%	97.2%	13.4%	13.3%	1.2%
	1.0	5	81.1%	6.6%	6.4%	0.7%	98.9%	13.0%	13.0%	0.7%
13b (w) v 13b	0.0	1	0.7%	0.7%	0.0%	0.0%	55.9%	3.1%	2.8%	0.8%
	0.5	1	3.4%	0.8%	0.1%	0.0%	57.4%	4.7%	3.8%	0.4%
	1.0	1	30.6%	3.2%	2.0%	0.3%	73.1%	5.8%	5.6%	0.4%
	0.0	3	1.2%	1.8%	0.2%	0.1%	80.3%	5.6%	5.4%	0.7%
	0.5	3	10.3%	1.7%	0.8%	0.1%	79.8%	8.5%	7.8%	0.9%
	1.0	3	53.2%	6.3%	5.2%	0.7%	90.5%	8.3%	8.3%	0.4%
	0.0	5	2.2%	3.0%	0.7%	0.3%	86.2%	7.6%	7.5%	0.7%
	0.5	5	17.8%	3.4%	2.1%	0.4%	86.3%	12.4%	11.7%	1.1%
	1.0	5	63.1%	8.1%	7.2%	0.9%	94.7%	11.2%	11.0%	0.4%
7b (w) v 13b	0.0	1	0.4%	0.5%	0.1%	0.1%	67.5%	3.9%	3.2%	0.2%
	0.5	1	3.4%	1.2%	0.0%	0.0%	77.7%	2.3%	2.1%	0.2%
	1.0	1	63.4%	3.4%	2.9%	1.1%	83.8%	7.6%	7.4%	0.1%
	0.0	3	2.6%	1.0%	0.1%	0.1%	89.6%	9.2%	8.8%	1.8%
	0.5	3	9.3%	2.4%	0.5%	0.0%	95.9%	5.5%	5.1%	0.0%
	1.0	3	80.0%	4.7%	4.5%	0.1%	97.7%	14.4%	14.1%	0.4%
	0.0	5	4.1%	2.1%	0.7%	0.2%	96.0%	11.8%	11.8%	2.0%
	0.5	5	13.0%	3.3%	0.9%	0.1%	98.7%	8.0%	7.5%	0.0%
	1.0	5	89.3%	6.0%	6.0%	0.4%	99.3%	18.2%	18.2%	0.4%
13b (w) v 7b	0.0	1	0.9%	0.8%	0.0%	0.0%	65.5%	4.1%	4.0%	0.4%
	0.5	1	3.1%	0.7%	0.1%	0.0%	66.9%	1.9%	1.7%	0.1%
	1.0	1	33.2%	3.9%	3.1%	0.4%	74.4%	4.6%	4.1%	0.4%
	0.0	3	1.0%	0.1%	0.0%	0.0%	91.1%	6.8%	6.6%	0.3%
	0.5	3	11.0%	2.4%	0.9%	0.4%	92.1%	5.1%	4.7%	0.4%
	1.0	3	54.1%	6.7%	6.2%	1.1%	93.0%	7.4%	7.3%	0.7%
	0.0	5	4.2%	0.7%	0.0%	0.0%	94.1%	7.9%	7.8%	0.1%
	0.5	5	15.3%	3.7%	1.7%	0.7%	96.1%	6.7%	6.3%	0.4%
	1.0	5	69.0%	8.4%	7.9%	0.5%	98.0%	10.2%	10.2%	0.8%

Table 2: Expanded version of Table 1. Table shows proportion of text that has a watermark. Here a watermarked LLM model (watermark ngram=4, threshold 0.01, averaged over three random keys (large variance)) interacts with a non-watermarked LLM model (Temperature=0.8). **P**rompt refers to the *watermarked prompt* from a first model. **R**esponse refers to the *response of an unwatermarked model* to a watermarked prompt. **#R** refers to a number of responses concatenated together. **P&R** reports a proportion of cases where watermarked prompt – i.e. watermark score < 0.01 – results in a watermarked response – watermark score < 0.01. **P<R** reports the proportion of cases where watermark of the response from an unwatermarked model is stronger than the watermark in the prompt. Original conversation prompt is listed in Appendix B.