The Extractive-Abstractive Axis:
Measuring Content “Borrowing” in Generative Language Models

Nedelina Teneva [1]

Abstract

Generative language models produce highly *abstractive* outputs by design, in contrast to *extractive* responses in search engines. Given this characteristic of LLMs and the resulting implications for content Licensing & Attribution, we propose the the so-called *Extractive-Abstractive axis* for benchmarking generative models and highlight the need for developing corresponding metrics, datasets and annotation guidelines. We limit our discussion to the text modality.

1. Introduction

The widespread adoption of Large Language Models (LLMs) has created many practical data governance challenges, among which Licensing & Attribution has emerged as a key one (Jernite et al., 2022). The interplay between generative language models and copyright law, the fair use doctrine and licensing requirements is of broad research and practical interest to legal practitioners, and increasingly, developers and users of LLMs. This topic is not new: content owners’ rights have been of interest to the legal community since the inception of the web and the subsequent wide spread use of search engines (Travis, 2008). Traditionally, search engines have been powered by information retrieval techniques, which take as input a user query and output a query answer by parsing out relevant paragraphs, sentences or phrases from a web-scale corpus of documents to produce an *attributable extractive answer* to the query.

The advent of LLMs – which Liu et al. (2023a) call “generative search engines” – is leading to a paradigm shift from *attributable extractive* question answering and summarization methodologies to increasingly *abstractive* ones. To produce these abstractive responses, generative models (Lewis et al., 2019; Raffel et al., 2020) synthesize information from multiple sources/text documents using sequence-to-sequence LLMs such that the generated answers may be highly abstractive or otherwise not readily attributable – as they are in search engines – to a specific content source such as a document on the web with a unique URI identifier [1] . Given this shift, we propose the *Extractive–Abstractive axis* for quantifying the propensity of LLMs for content borrowing. We highlight the need for relevant metrics, benchmarks and annotations and list some practical challenges in Section 4.

2. The Extractive–Abstractive Axis

Being able to quantify a generative language model’s extractiveness/abstractiveness level – in other words where the model lies on what we call the *Extractive-Abstractive axis* – with respect to one or several sources (e.g., a text snippet, a web page or social media post), is necessary for evaluating whether (and how much) a generative AI application is using content from copyrighted or licensed sources. Intuitively, LLM answers with high levels of content borrowing in the absence of proper attribution constitute a higher risk for copyright infringement. By way of a practical example: a news publisher would like to determine if their article was used for training a LLM without their permission. If the publisher had access to the LLM pre-training and fine tuning corpus they can examine each training document and compare it to their article. If the LLM is only commercially accessible through APIs (e.g., ChatGPT (Bubeck et al., 2023)), the publisher may query it in an attempt to examine if its responses contain snippets from their article. Depending on the abstractiveness level of the responses, the publisher may be facing a rather complex prompting task while at the same time providing additional training information to an already potentially copyright-infringing LLM operator.

Quantification along the Extractive-Abstractive axis is of practical use to content owners, the developers of generative language applications and third parties for several reasons: **(1)** Being able to quantify the level of language/content borrowing will allows content owners or third parties such as algorithmic auditors (Raji & Buolamwini, 2019) to quantify how prone a trained LLM is to content borrowing. **(2)** Such metrics will enable the designers of generative

[1] Megagon Labs, Mountain View, CA, USA. Correspondence to: Nedelina Teneva <nedteneva@gmail.com>.

Accepted to the 1^{st} *Workshop on Generative AI and Law*, co-located with the *International Conference on Machine Learning*, Honolulu, Hawaii, USA. 2023. Copyright 2023 by the author(s).

[1] https://datatracker.ietf.org/doc/html/rfc3986

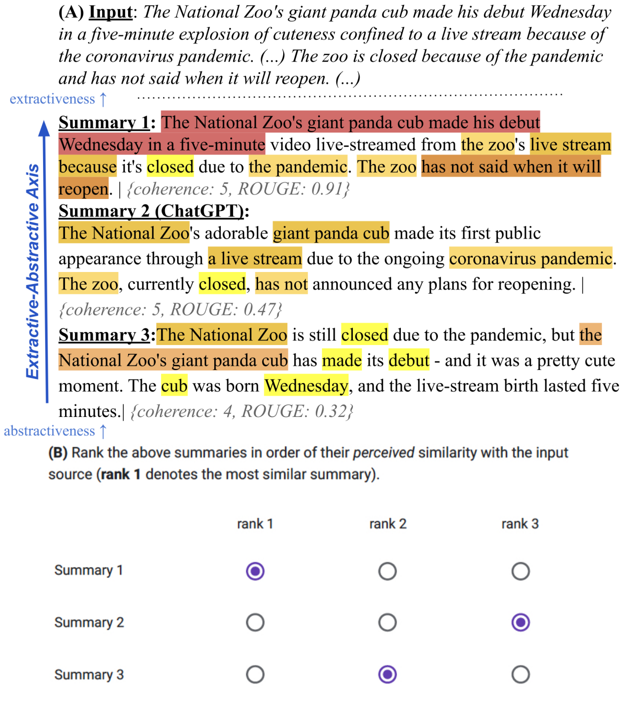
language-based applications to minimize their legal risks (e.g. copyright infringing) by identifying highly extractive responses at inference/run time. **(3)** More generally, such tools will also help organizations with the assessment of LLM related Licensing & Attribution risks pre- and post-deployment or liability assessment of off-the-shelf tools such as black box LLMs. **(4)** In courts these metrics can potentially be used for quantifying if a LLM-generated text is substantially similar to copyrighted content (or derivative of such). We may even imagine cases in which the empirical propensity (measured on benchmarking datasets) of generative models for borrowing large amounts of content may also play a role in Licensing & Attribution matters.

3. Metrics, Datasets and Annotation Tasks

Metrics. Existing Natural Language Processing (NLP) tasks such as question answering, machine translation, extractive and abstractive summarization (see Appx. A.2 for definitions) use various automatic metrics to measure the similarity between generated answers and the true "gold" answers. Some of these include (see Liu et al. (2023b); Fabbri et al. (2021) for additional ones): **(1)** token overlap metrics (e.g., ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002)) compare the similarity between two texts based on the n-grams (contiguous sequence of tokens) overlap between them; **(2)** vector-based metrics such as BERTScore (Zhang et al., 2019) and BARTScore (Yuan et al., 2021) measure text sequence similarity based on text representations learnt by neural models; **(3)** metrics relying on the assumption that if two texts are similar, they should be able to address the same set of questions – one example is QAEval (Deutsch et al., 2020); **(4)** additional metrics such as n-grams ratio (Narayan et al., 2018) or coverage (Grusky et al., 2018) are used for measuring summarization quality.

In principle, while some of the above mentioned NLP metrics can be repurposed to measure LLMs along the Extractive-Abstractive axis in matters related to Licensing & Attribution, no empirical studies exist on this topic and there are no evaluation benchmarks to guide such analysis. Since these automatic metrics have not been previously applied in the context of Licensing& Attribution, there are no empirical studies of whether they correlate with content owners' perception of Licensing & Attribution.

Datasets and Human Annotations. Like all NLP models, LLMs are evaluated with respect to the downstream *user perception* of the answer quality – see Rogers et al. (2023) for a review and taxonomy of the vast number of NLP datasets. They are, however, not evaluated with respect to the *content owners' perception* of how well their content is used for answering users' questions. Given LLM's propensity for content borrowing, it is critical that the experience and rights of content owners are balanced with those of application users.



(A) **Input**: *The National Zoo's giant panda cub made his debut Wednesday in a five-minute explosion of cuteness confined to a live stream because of the coronavirus pandemic. (...) The zoo is closed because of the pandemic and has not said when it will reopen. (...)*

Summary 1: The National Zoo's giant panda cub made his debut Wednesday in a five-minute video live-streamed from the zoo's live stream because it's closed due to the pandemic. The zoo has not said when it will reopen. | *{coherence: 5, ROUGE: 0.91}*

Summary 2 (ChatGPT): The National Zoo's adorable giant panda cub made its first public appearance through a live stream due to the ongoing coronavirus pandemic. The zoo, currently closed, has not announced any plans for reopening. | *{coherence: 5, ROUGE: 0.47}*

Summary 3: The National Zoo is still closed due to the pandemic, but the National Zoo's giant panda cub has made its debut - and it was a pretty cute moment. The cub was born Wednesday, and the live-stream birth lasted five minutes.| *{coherence: 4, ROUGE: 0.32}*

(B) Rank the above summaries in order of their *perceived* similarity with the input source (**rank 1** denotes the most similar summary).

	rank 1	rank 2	rank 3
Summary 1	●	○	○
Summary 2	○	○	●
Summary 3	○	●	○

Figure 1. **(A)** Illustration of 3 summaries (of the same input) along the Extractive-Abstractive Axis (similarity measured by ROUGE) and their example coherence scores. Text by Dreyer et al. (2023), with permission. **(B)** Example guideline for annotating the perceived similarity between the summaries. Details in Appx. A.1.

A simple approach for benchmarking content owners' perception of Licensing & Attribution quality is by repurposing existing NLP datasets. For example, summarization tasks, which already rely heavily on human annotation, may be particularly well suited for benchmarking generative models along the Extractive-Abstractive axis. Currently, summarization tasks are benchmarked using human annotators who rate the generated summaries along dimensions such as summary relevance, fluency, coherence (Fabbri et al., 2020) or the level of factual alignment between the summary and the underlying source text being summarized (e.g., answer consistency (Fabbri et al., 2020), faithfulness (Ladhak et al., 2021) and factuality (Dreyer et al., 2023)). As an example, Figure 1A shows summaries with their example coherence scores. These summarization benchmarks can be augmented with (legal) expert annotation of Licensing & Attribution quality assessing whether, e.g. (1) the similarity between the input text and summary is acceptable, (2) there are copyright concerns, (3) any extractive snippets are properly attributed to the source. Some of these dimensions may be more easy to assess in a comparative manner, rather than individually as long as proper annotator agreement is established. Figure 1B illustrates this point with an example annotation question which ranks the 3 summaries in order of decreasing perceived similarity.

4. Practical Challenges and Limitations

There are several practical challenges associated with measuring generative models along the Extractive-Abstractive axis which we categorize below.

Evaluation Challenges: Human evaluation, especially of longer answers, is a hard and actively studied research problem (Rogers et al., 2023). Content Licensing & Attribution nuances and expertise required can pose challenges to the human evaluation of the generated responses described in Section 3 and Figure 1. Additionally, while this study focuses mainly on English language and we note that content borrowing may be different in other languages.

Usability Challenges and Conflicting Interests: (1) Correlation between faithfulness/factuality and extractiveness (Dreyer et al., 2023; Ladhak et al., 2021) observed in summarization tasks implies that a certain level of extractiveness may be needed in the generated answers in order to balance mis/disinformation concerns. This observation may heighten the necessity of measuring content borrowing in LLMs along the Extractive-Abstractive axis. (2) Interactions with LLMs can be used as additional training signals for the underlying LLM system so a practical challenge is how generative models can be audited for Licensing & Attribution purposes without further aiding their development.

Ethical Challenges: Content borrowing poses numerous ethical challenges in addition to legal ones. In order to mitigate such challenges, incentives and broader policies may be needed in order to alleviate the concerns of both content owners and LLM end users. Adversarial scenarios also needs to be considered: LLMs can be tuned for specific abstraction levels which means that copyrighted content used for pre-training and fine tuning can be intentionally obfuscated. In such cases, developing methodologies for identifying copyright infringement in black box LLMs becomes even more critical.

Acknowledgements

We would like to thank John Santerre for their input in the early stages of the project; Estevam Hruschka, Sajjadur Rahman and the anonymous reviewers for their draft feedback and suggestions.

References

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Deutsch, D., Bedrax-Weiss, T., and Roth, D. Towards question-answering as an automatic metric for evaluating the content quality of a summary. arxiv. *arXiv preprint arXiv:2010.00490*, 2020.

Dreyer, M., Liu, M., Nan, F., Atluri, S., and Ravi, S. Evaluating the tradeoff between abstractiveness and factuality in abstractive summarization. In *Findings of the European Association for Computational Linguistics: EACL 2023*, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. URL https://arxiv.org/abs/2108.02859.

Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., and Radev, D. Summeval: Re-evaluating summarization evaluation. *arXiv preprint arXiv:2007.12626*, 2020.

Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., and Radev, D. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 2021.

Grusky, M., Naaman, M., and Artzi, Y. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 708–719, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1065. URL https://aclanthology.org/N18-1065.

Jernite, Y., Nguyen, H., Biderman, S., Rogers, A., Masoud, M., Danchev, V., Tan, S., Luccioni, A. S., Subramani, N., Johnson, I., et al. Data governance in the age of large-scale data-driven language technology. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2206–2222, 2022.

Ladhak, F., Durmus, E., He, H., Cardie, C., and McKeown, K. Faithful or extractive? on mitigating the faithfulness-abstractiveness trade-off in abstractive summarization. *arXiv preprint arXiv:2108.13684*, 2021.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Liu, N. F., Zhang, T., and Liang, P. Evaluating verifiability in generative search engines. *arXiv preprint arXiv:2304.09848*, 2023a.

Liu, Y., Fabbri, A. R., Zhao, Y., Liu, P., Joty, S., Wu, C.-S., Xiong, C., and Radev, D. Towards interpretable and efficient automatic reference-based summarization evaluation. *arXiv preprint arXiv:2303.03608*, 2023b.

Narayan, S., Cohen, S. B., and Lapata, M. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*, 2018.

Nenkova, A. and McKeown, K. A survey of text summarization techniques. *Mining text data*, pp. 43–76, 2012.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

Raji, I. D. and Buolamwini, J. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 429–435, 2019.

Rogers, A., Gardner, M., and Augenstein, I. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *ACM Computing Surveys*, 55(10):1–45, 2023.

Soares, M. A. C. and Parreiras, F. S. A literature review on question answering techniques, paradigms and systems. *Journal of King Saud University-Computer and Information Sciences*, 32(6):635–646, 2020.

Travis, H. Opting out of the internet in the united states and the european union: Copyright, safe harbors, and international law. *Notre Dame L. Rev.*, 84:331, 2008.

Yuan, W., Neubig, G., and Liu, P. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277, 2021.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

A. Appendix

A.1. Figure Details

The figure shows three summaries of the input text snippet shown at the top of the figure; each summary is of different degree of extractives/abstractiveness (measured by ROUGE score (Lin, 2004)). Summaries 1 and 3 are reproduced with permission from Dreyer et al. (2023)'s study. Summary 2 was obtained by prompting ChatGPT (using the free plan) in May 2023 as follows: "Summarize this snippet: 'The National Zoo's giant panda cub made his debut Wednesday in a five-minute explosion of cuteness confined to a live stream because of the coronavirus pandemic. The zoo is closed because of the pandemic and has not said when it will reopen.'"

For each summary in (A), we show 1) an example scoring of the coherence; and 2) automatically computed ROUGE-L score (recall) between the summary and the input using the `py-rouge` library (https://pypi.org/project/py-rouge/). Following Dreyer et al. (2023), fragments extracted from the input are marked from red (longer fragments) to yellow (shorter fragments).

A.2. NLP Task Definitions

Question Answering refers to the task of answering asked by humans in natural language using either a pre-structured database or a collection of text documents (see (Soares & Parreiras, 2020) for a review). There are various subtypes of question answering such as factoid question answering or multiple choice question answering.

Summarization tasks aim produce summaries of single or multiple documents to answer questions that require longer responses. The goal is to convey the key information in the input text. In *extractive* summarization, the summarizers identify the most important sentences in the input, which can be either a single document or a cluster of related documents, and string them together to form a summary (Nenkova & McKeown, 2012). In *abstractive* summarization, the summary contains synthesized text which may not be explicitly present in the input text.