Exploring Precision and Recall to assess the quality and diversity of LLMs

Florian Le Bronnec, Alexandre Vérine, Benjamin Negrevergne, Yann Chevaleyre, Alexandre Allauzen

Exploring Precision and Recall to assess the quality and diversity of LLMs

Florian Le Bronnec[*,1,2] **Alexandre Verine**[*,1]
Benjamin Negrevergne[1] **Yann Chevaleyre**[1] **Alexandre Allauzen**[1]
[1]Miles Team, LAMSADE, Université Paris-Dauphine, Université PSL, CNRS, 75016 Paris, France
[2]Sorbonne Université, CNRS, ISIR, F-75005 Paris, France

Abstract

The implementation of CAME is publicly available This paper introduces a novel evaluation framework for Large Language Models (LLMs) such as LLAMA-2 and MISTRAL, focusing on the adaptation of Precision and Recall metrics from image generation to text generation. This approach allows for a nuanced assessment of the quality and diversity of generated text without the need for aligned corpora. By conducting a comprehensive evaluation of state-of-the-art language models, the study reveals significant insights into their performance on open-ended generation tasks, which are not adequately captured by traditional benchmarks. The findings highlight a trade-off between the quality and diversity of generated samples, particularly when models are fine-tuned with human feedback. This work extends the toolkit for distribution-based NLP evaluation, offering insights into the practical capabilities and challenges that current LLMs face in generating diverse and high-quality text. We release our code and data[1]

1 Introduction

In recent years and months, there has been a rapid democratization of Large Language Models (LLMs), exemplified by platforms such as ChatGPT and HuggingChat. These models are now widely accessible for a diverse range of tasks, from composing emails and application letters to performing multi-document summarization, generating medical prescriptions, and even crafting poetry and novels. The expanding spectrum of applications underscores the ubiquity of LLMs in our daily lives, necessitating the development of novel evaluation frameworks to accommodate their growing relevance.

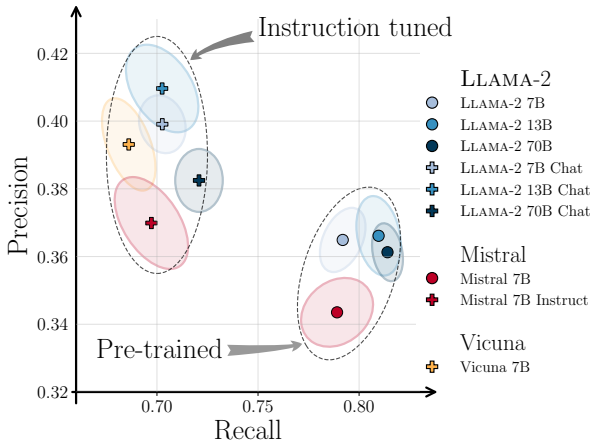


Figure 1: Precision and Recall of various models on generating the WebText dataset, with the 2 standard deviation error ellipsis. Chat and pre-trained models different behaviors are clearly captured by our metrics.

Until recently, benchmarks were designed to target specific tasks such as machine translation, summarization, and question answering, among others. However, LLMs now encompass a wide range of tasks, prompting the community to reconsider methods for comparing and assessing these models. One proposed solution is to gather a diverse set of tasks to better evaluate the versatility of modern generative models. For example, the Open LLM Leaderboard (Beeching et al., 2023) presents a unified framework consisting of closed questions that focus on the model's ability to provide concise and accurate answers. However, it is important to note that these evaluations are sample-based, relying on generated samples from the model compared against aligned references, written by humans.

In contrast, the recent line of work on *distribution-based* metrics greatly departs from the *sampled based* evaluation. By considering LLMs and datasets as empirical distributions, the new metrics attempt to quantify how they differ and how they overlap. This shift drastically changes the scope of the evaluation. Beyond performance

[*]Authors contributed equally to this work. Corresponding authors: florian.le-bronnec@dauphine.psl.eu, alexandre.verine@dauphine.psl.eu.

[1]https://github.com/AlexVerine/pr-4-llm.

measures based on human references, the goal is to estimate the disparity between some data distribution (P) exhibited by human-written texts and the distribution learned by a LLM (Q), eliminating the need for aligned corpora. As exemplified by the development of MAUVE (Pillutla et al., 2021), this kind of approach opens new perspectives to really compare LLMs in terms of their generative abilities rather than for some peculiar tasks.

However, many factors take part in assessing open-ended text generation and focusing on a single measure may restrict the significance of the evaluation. This is well known in the field of Information Retrieval, for example, where Precision and Recall are at the core of evaluation. While they have been previously tailored for image generation by Sajjadi et al. (2018) and Kynkäänniemi et al. (2019), we propose to adapt them to LLMs. With an extensive set of experiments involving prominent LLMs such as LLAMA-2 and Mistral, we show that these two new metrics significantly improve the evaluation with better understanding and characterization of the flaws in text generation. Precision and Recall allow us to clearly distinguish between samples quality or adequacy (for Precision) and a lack of diversity in the model outputs (for Recall). Empirical results show that these two measures are necessary for in-depth comparison of LLMs. For instance, we are able to quantify that fine-tuning models with human feedback significantly improves sample quality, albeit at the expense of sample diversity, as evidenced in Figure 1 by the trade-off between Precision and Recall.

As a summary, our contributions are threefold:

- We adapt the concepts of Precision and Recall, traditionally used in image generation, to evaluate the performance of large language models. Our method offers a novel lens to assess the quality and diversity of text generation without the need for aligned corpora.
- We carry out a thorough evaluation of state-of-the-art language models, such as Llama and Mistral, using our proposed framework. This analysis provides a detailed comparison of the performance of these models in terms of quality and diversity on challenging open-ended generation tasks that are out of the scope of traditional evaluation methods.
- Our investigation sheds light on the impact of fine-tuning LLMs with human feedback. We present empirical evidence that, while such fine-tuning can improve the quality of generated samples, it also reduce their diversity, highlighting a crucial trade-off between Precision and Recall.

In summary, our contributions advance the field of distribution-based NLP evaluation by introducing novel metrics tailored to the specific tasks of open-ended generation. Through empirical analysis and insights, we provide a deeper understanding of LLM generation capabilities.

2 Related works

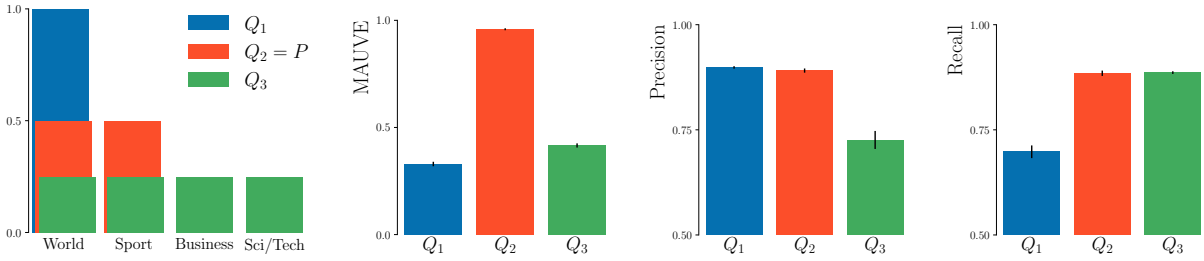
Historically, assessing text generation tasks has been dependent on human resources, such as annotated texts or aligned corpora. This is particularly true in areas like machine translation and automatic summarization, where N-gram based metrics have been developed, notably BLEU (Papineni et al., 2001) and ROUGE (Lin, 2004). These metrics perform a rough comparison between the generated text and human-produced and aligned reference texts. More recent developments have introduced model-based metrics, such as BERTScore[2] (Zhang* et al., 2020), which aim to capture semantic similarities more effectively. However, these evaluations still largely depend on "aligned references," meaning human-provided examples that constrain the expected range of generated outputs. The challenge becomes significantly greater with open-ended generation tasks, where the range of acceptable responses is vast and cannot be encapsulated by a single reference.

2.1 Standard evaluation of generative models

Perplexity is the historical measure for language modeling (Bahl et al., 1977). Easy to use and cheap to compute, practical motivations clearly explain its persistency for evaluation, even if it has notable limitations. As demonstrated in the work of Pillutla et al. (2021), perplexity works at the token level, only considering the surface forms, hence missing semantic information. Moreover, the metric is calculated on a per-sample basis, thereby failing to capture the diversity of generated texts.

Nowadays, the prevalent method for evaluating recent LLMs involves benchmarks with many tasks with easily verifiable correct answers (Touvron et al., 2023; Google, 2023; Beeching et al., 2023).

[2]BERTScore uses precision and recall scores that differ significantly from ours, as they evaluate the similarity between a single generated text and its aligned reference



(a) Distributions of labels. (b) MAUVE of Q w.r.t. P. (c) Precision of Q w.r.t. P. (d) Recall of Q w.r.t. P.

Figure 2: Illustration of what can bring Precision and Recall compared to MAUVE. We consider a reference dataset P composed of articles from 2 labels, `World` and `Sport`. Q_2 is made of articles from the same distribution. We compare it with two other datasets: Q_1 composed only of `World` articles and Q_3 composed of even numbers of `World, Sport, Business` and `Sci/Tech` articles. Relatively to Q_2, the MAUVE scores of Q_1 and Q_3 are almost identical, while Precision and Recall help differentiating how the distributions actually differ from the reference P.

Typically, this entails the generation of responses to short questions prompted in the instructions. The responses are then compared to a reference using metrics such as Exact-Match. While effective in assessing the comprehension and reasoning capabilities of LLMs, this kind of approach does not assess the generation skills of the models.

To overcome the limitations of automatic metrics, another trend considers human evaluation, especially for complex tasks. Beyond its cost, the guidelines must be carefully designed, since various criteria could be necessary to rate texts and compare model outputs. At the end human evaluation is prone to high variance and is difficult to reproduce. Another issue, is that diversity is not evaluated is standard human evaluation protocols, since texts are rated individually or by pairs.

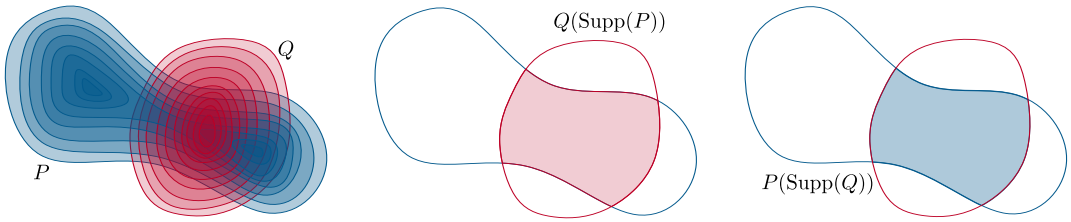
2.2 Evaluation of open-ended generation

While the prevailing evaluations primarily focus on the quality of generated texts, some studies aim to characterize the diversity of these outputs. One approach involves computing the proportion of distinct N-grams generated, as demonstrated in (Li et al., 2016). Another notable proposition by (Zhu et al., 2018) introduces Self-BLEU which computes the BLEU score of each generated text against all other generated texts as a reference. However, N-gram matching cannot capture semantic features. For instance, a model generating random words achieves a perfect Self-BLEU score.

In the current landscape, most of the existing methods either prioritize quality or rely on simplistic characterization of diversity. Nevertheless, the recent introduction of MAUVE Pillutla et al. (2021) makes a significant shift toward diversity. By drawing inspiration from the domain of image generation, this new metric compares generated texts to a reference distribution without requiring aligned texts. MAUVE consists of a divergence-based metric that captures certain properties of text and exhibits correlation with human judgment. Subsequent research by Pimentel et al. (2023) confirms the validity of this approach by leveraging common divergence measures. MAUVE and subsequent works propose a simple metric that allows summarizing the evaluation of both quality and diversity in a single measure. Our work builds upon these efforts; we aim to demonstrate that using two distinct measures enables to distinguish between lack of quality and lack of diversity. This distinction becomes necessary to enable a deeper and more precise understanding of generative models, as illustrated in a simple example in Figure 2.

(a) Q_1. P ≈ 0.95, R ≈ 0. (b) Q_2. P ≈ 0.4, R ≈ 0.4.

Figure 3: Example of distribution of images. P is the reference distribution of images of the CelebA dataset (Liu et al., 2015), Q_1 and Q_2 are two different distributions of images. Q_1 has high quality, but low diversity, while Q_2 has high diversity and a low quality. Numbers and images are from Kynkäänniemi et al. (2019).



(a) Distributions P and Q (b) Precision to assess quality (c) Recall to assess diversity

Figure 4: Precision and Recall for distribution-based metrics. (a) Distributions P and Q. (b) Precision is the proportion of the support of Q that generates P. (c) Recall is the proportion of the support P generated by Q.

3 Background on Precision and Recall

Nowadays generative models for images are able to output ultra-realistic samples, making classical metrics such as Heusel et al. (2017) too coarse to assess the full distributional properties of the models. This jump of performance has motivated the definition of precision and diversity for generative models. This is exemplified on Figure 3. Q_1 represents a model that produces realistic images but lacks diversity in the generated people's appearance. Conversely, Q_2 depicts a model generating a broad range of people with lower quality. Capturing this significant tradeoff is essential for comprehending generative model behavior. Inspired by the precision and recall metrics of binary classifications, the pioneering work of Sajjadi et al. (2018) proposed to redefine these two terms to assess the quality and diversity of generative models. These metrics are based on a discretized estimation of the reference and model's distribution. Kynkäänniemi et al. (2019) pointed out some limitations of those metrics and proposed a new way to compute the Precision and Recall, leveraging more robust support estimation. This method is now widely adopted in the literature and is typically used along with the Fréchet Inception Distance (Heusel et al., 2017). Still, these metrics have some limitations, like the sensibility to outliers (Naeem et al., 2020). This motivates further ongoing work on this topic, including theoretical contributions (Simon et al., 2019; Naeem et al., 2020; Alaa et al., 2022; Cheema and Urner, 2023; Verine et al., 2023) or practical methods (Kim et al., 2023).

Precision and Recall definition. Given its importance in the field of image generation, we adopt the definition of Precision and Recall introduced by Kynkäänniemi et al. (2019). Alternative definitions and approaches are discussed in Appendix A.

Definition 1. *Let P and Q be two distributions over a space $\mathcal{X} \in \mathbb{R}^d$. We denote* $\mathrm{Supp}(P)$ *and* $\mathrm{Supp}(Q)$ *their support. The Precision and the Recall of Q with respect to P are defined as:*

$$\text{Precision} = Q(\mathrm{Supp}(P)) \tag{1a}$$

$$\text{Recall} = P(\mathrm{Supp}(Q)). \tag{1b}$$

Precision and Recall in practice. Since Precision and Recall involve an estimation of the support, standard pipelines involve a projection of the samples into an latent space, using pre-trained models such as Inception-v3 (Szegedy et al., 2015) or VGG (Simonyan and Zisserman, 2015). Support of the distributions is then estimated using a k-nearest neighbors algorithm. We present this algorithm in more detail in Section 4.

Quality and diversity in NLP. Although the practice of reporting Precision and Recall is well-established in image generation (Dhariwal and Nichol, 2021; Sauer et al., 2022; Song et al., 2023), it is a very recent practice in the field of text generation. Pillutla et al. (2021) led the way in this field by introducing MAUVE, a pioneering metric that builds on the theoretical concepts of quality and diversity, as initially proposed by Djolonga et al. (2020), and detailed in Appendix A.

The authors showcased the intriguing characteristics of their metrics across various GPT2 models and decoding algorithms. This metric is highly effective in evaluating how well the generated distribution aligns with the reference. The methodology behind it was subsequently validated and simplified by Pimentel et al. (2023) and further evaluated with alternative frameworks by Pillutla et al. (2023).

Despite the progress made, these approaches present a challenge: interpreting them as a measure of quality and diversity is not straightforward as they condense these two concepts into a single divergence measure. Characterizing Precision and Recall independently has only been mention by

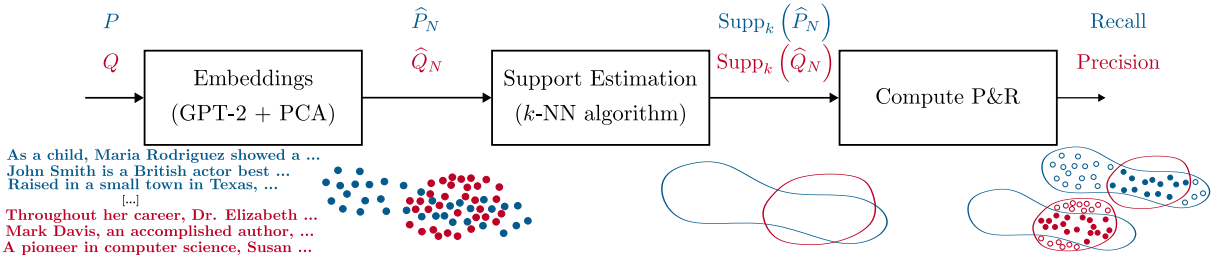


Figure 5: Our pipeline to compute the Precision and Recall metrics. Texts are projected into a latent space of a pre-trained model, where a k-NN estimation is performed to estimate the relative overlaps of P and Q.

Pillutla et al. (2023) who conducted some toys experiments on small models or decoding algorithms.

Therefore, we believe that Precision and Recall for text generation is in its early stages. We argue that a deeper understanding of these metrics and their practical implications is needed.

4 Precision and Recall to assess text generation

In this paper, we introduce a framework for calculating Precision and Recall between distributions on texts. These distributions characterize a dataset, or a model and its generative capability. The extended version of Precision and Recall builds upon the notion of support of distributions, which is in our case challenging to define and estimate.

4.1 Pipeline to compute Precision and Recall

We propose the following pipeline to compute the Precision and Recall metrics, as shown in Figure 5:

1. Select N sequences randomly in the dataset to build the set of references: $\mathcal{X}^{\text{ref}} = \left\{\boldsymbol{x}_1^{\text{ref}}, \dots, \boldsymbol{x}_N^{\text{ref}}\right\}$. Sample N sequences from the distribution Q to assess and build the set of outputs: $\mathcal{X}^{\text{out}} = \left\{\boldsymbol{x}_1^{\text{out}}, \dots, \boldsymbol{x}_N^{\text{out}}\right\}$.
2. Samples pre-processing:
   (a) Compute the latent representations of each set using the embedding function ϕ:
   $$\mathcal{X}_\phi^{\text{ref}} = \left\{\phi(\boldsymbol{x}_1^{\text{ref}}), \dots, \phi(\boldsymbol{x}_N^{\text{ref}})\right\}$$
   $$\mathcal{X}_\phi^{\text{out}} = \left\{\phi(\boldsymbol{x}_1^{\text{out}}), \dots, \phi(\boldsymbol{x}_N^{\text{out}})\right\}$$
   (b) Perform Principal Component Analysis (PCA) on the union of sets $\mathcal{X}_\phi^{\text{ref}}$ and $\mathcal{X}_\phi^{\text{out}}$ to reduce the dimensionality of the data by retaining 90% of the variance. This process yields the sets $\mathcal{Z}_\phi^{\text{ref}}$ and $\mathcal{Z}_\phi^{\text{out}}$, which consist of points, respectively, distributed on $\widehat{P}_N$ and $\widehat{Q}_N$, the empirical distributions in the latent space.
3. Support estimation:
   (a) For each set $\mathcal{Z}_\phi \in \left\{\mathcal{Z}_\phi^{\text{ref}}, \mathcal{Z}_\phi^{\text{out}}\right\}$, for each point $\boldsymbol{z}_i \in \mathcal{Z}_\phi$ computes the pairwise distances. Define $B_k(\boldsymbol{z}_i, \mathcal{Z}_\phi)$ as the ball centered on $\boldsymbol{z}_i$ with radius being the distance to the k-th nearest neighbor in $\mathcal{Z}_\phi$.
   (b) Each support is defined as the union of balls $B_k(\boldsymbol{z}_i, \mathcal{Z}_\phi)$ for all $\boldsymbol{z}_i$ in $\mathcal{Z}_\phi$:
   $$\mathrm{Supp}_k(\widehat{P}_N) = \bigcup\nolimits_{\boldsymbol{z}_i \in \mathcal{Z}_\phi^{\text{ref}}} B_k(\boldsymbol{z}_i, \mathcal{Z}_\phi^{\text{ref}})$$
   $$\mathrm{Supp}_k(\widehat{Q}_N) = \bigcup\nolimits_{\boldsymbol{z}_i \in \mathcal{Z}_\phi^{\text{out}}} B_k(\boldsymbol{z}_i, \mathcal{Z}_\phi^{\text{out}})$$
4. Precision and Recall computation:
$$\text{Precision} = \frac{1}{N} \sum_{\boldsymbol{z}_i \in \mathcal{Z}_\phi^{\text{out}}} \mathbb{1}_{\boldsymbol{z}_i \in \mathrm{Supp}_k(\widehat{P}_N)} \tag{2a}$$
$$\text{Recall} = \frac{1}{N} \sum_{\boldsymbol{z}_i \in \mathcal{Z}_\phi^{\text{ref}}} \mathbb{1}_{\boldsymbol{z}_i \in \mathrm{Supp}_k(\widehat{Q}_N)} \tag{2b}$$

The parameters are detailed in the next section.

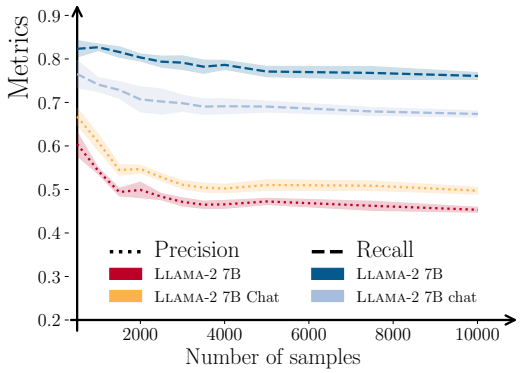


Figure 6: Evolution of the metrics Precision and Recall computed with $k = 4$ as the number of samples N increases. The shaded area represents the standard deviation computed over five random seeds for the set outputs. The means reach a plateau around $N = 3000$.

4.2 Precision and Recall for text generation.

The embeddings used for the MAUVE metric have proven effective at capturing both word level (Pimentel et al., 2023) and content level (Pillutla et al., 2021) properties of texts. Building on this success, we adopt the same embedding function, ϕ, which leverages the output from the last layer of GPT-2_{LARGE}, coupled with a PCA. The PCA serves a dual purpose: it reduces data dimensionality and also filters out noise, which is particularly beneficial for k-NN based manifold estimation.

To choose the other parameters – the number of samples N and the neighbor parameter k – we investigate how Precision and Recall are affected by their variations. We utilize the WebText generation task with LLAMA-2 7B Chat (comprehensively described in Section 5.3) as a test case, prompting the model to generate between 100 to 10,000 samples. The findings are depicted in Figure 6. Additional details can be found in Appendix B.1.

Our observations are: (1) Precision and Recall values stabilize with an increasing N, plateauing at approximately $N = 3000$. (2) As k increases, Precision and Recall values gradually approach 1, aligning with the expectation that a larger k leads to a wider estimated support. In the rest of the paper, we set N to 4000 samples and k to 4.

5 Practical use cases for LLMs

Equiped with our freshly introduced Precision and Recall we now present an analysis of the quality of generation of a broad range of LLMs. We focus on three different tasks to illustrate the versatility of our metrics and the new insights they can provide.

5.1 Experimental settings

For the models, we stress the distinction between *pre-trained* models and *instruction-tuned* (or *chat*) models. The former are models that have been trained on a large corpus of text, and the latter have been fine-tuned with instructions and optionally aligned with human feedback.

Models. We considered the following recent LLMs:

- LLAMA-2 {7, 13, 70}B counterparts fine-tuned and aligned with human feedback (Touvron et al., 2023).
- VICUNA (Zheng et al., 2023), derived from LLAMA-2 7B model trained on a large open-source instruction dataset.

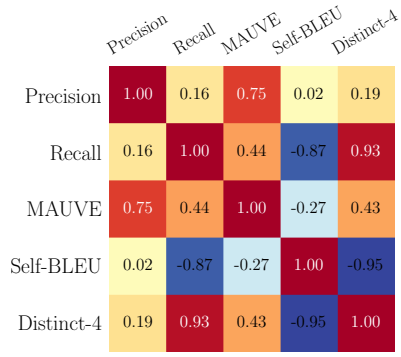


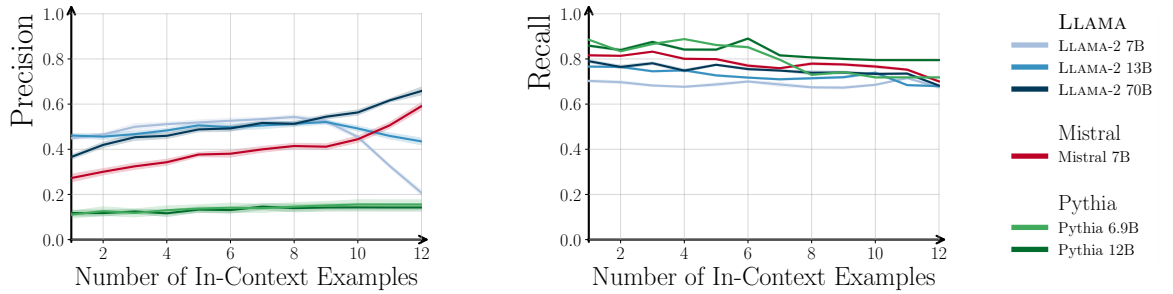
Figure 7: Correlation between Precision and Recall for Wikipedia's biographies generation.

- MISTRAL 7B and its insctruction-tuned version MISTRAL 7B Instruct (Jiang et al., 2023).
- PYTHIA {6.9,12}B (Biderman et al., 2023)

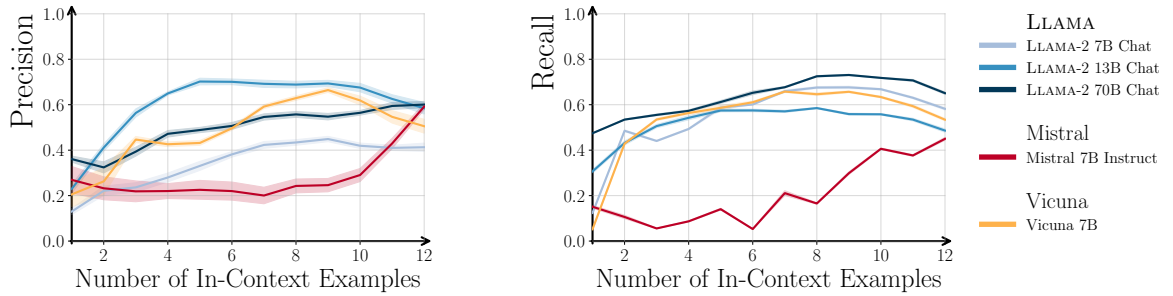
Text generation. For all tasks, we use the model to generate 4000 samples. For WebText and Creative writings, we additionally use five different seeds for the random generator to account for the stochasticity of the generation process. We then compute the average Precision and Recall over the the different seeds.

WebText generation. We run generations of a wide suits of LLMs, of sizes ranging from 7B parameters to 70B parameters, with both pre-trained and instruction-tuned models. In this setting, we make models generate texts close to their pre-training data distribution. Models are prompted with a minimal instruction of 10 words extracted from the WebText train set, and we evaluate their ability to continue the prompt. To account for the different tokenizations, input and generation lengths are constrained in terms of words, rather than in terms of tokens.

Biographies generation. We evaluate the abilities of LLMs to generate Wikipedia-like biographies, given some in-context examples of biographies. The benchmark for comparison consists of summary sections from Wikipedia pages of individuals who have been distinguished with either "Good Article" or "Featured Article" accolades. The models receive a prompt that includes an instruction followed by a varying number of in-context biography examples. Our evaluation focuses on analyzing how the quality and diversity of the generated biographies evolves as we increase the number of in-context examples provided.



(a) Pre-trained models.



(b) Instruction-tuned models.

Figure 8: Evolution of Precision and Recall for Wikipedia's biographies generation based on the number of in-context examples. Standard deviations are depicted with transparency. Pre-trained models consistently exhibit higher Recall compared to Chat models. Precision and Recall of instruction-tuned models progressively improve until reaching a plateau after a certain number of in-context examples.

Creative texts generation. We task the models with generating creative texts based on a set of 50 manually-crafted creative instructions such as `Write about a dream you had` or `Write a script for a short film`. We leverage the stochastic decoding procedure of models to generate a large number of output based on these 50 prompts. The specificity of this task is that no reference dataset is available.

Appendix B provides full details on experiments.

5.2 Preliminary results

We first present some preliminary results to illustrate the potential of our metrics.

AG news topics modeling. We consider a reference distribution P made up of journal articles on two topics (`World, Sport`), extracted from the AG news dataset (Zhang et al., 2015). We then simulate three models. Q_1 corresponds to a model that generates texts on a single topic. Q_2 generates texts from the two reference topics. Q_3 generates text on divergent topics compared to P. When considering P as a reference, Q_1 therefore suffers from a lack of diversity and Q_3 from a lack of quality. The experiment is illustrated in Figure 2. Precision and Recall metrics allow us to clearly distinguish between these cases, while,for instance, MAUVE would give the same score for both, demonstrating that our metrics can capture the trade-off between quality and diversity on a content-based level.

Correlation with other metrics. Figure 7 shows the correlation between Precision and Recall and other distribution-based metrics on the Wikipedia Biography dataset in our experiments. We observe that 1) Recall expectedly reflects word-level diversity as indicated by its correlation with Self-BLEU and Distinct-N 2) Precision shows no correlation with these metrics but does correlate with MAUVE. This suggests that Precision captures the adequacy of distributions, but not diversity, essentially indicating the quality of the generated texts. A complementary analysis of MAUVE on the WebText dataset is available in Appendix B.

5.3 Evaluation of open-ended generation

Instruction-tuned models are more precise and less diverse than pre-trained models. Figure 1 shows the Precision / Recall graph of various models. Instruction tuned models have a higher Precision and a lower Recall than their pre-trained counterparts. This is consistent with the expected effect of human preference alignment and instruction tuning: models are encouraged to generate more "human-like" texts, which do not encompass

	LLAMA-2 70B Chat	LLAMA-2 13B Chatct	LLAMA-2 7B Chat	Mistral 7B Instruct	Vicuna 7B
Precision	$\mathbf{1.00 \pm 0.00}$	0.95 ± 0.01	0.94 ± 0.01	0.99 ± 0.01	0.93 ± 0.01
Recall	$\mathbf{1.00 \pm 0.00}$	0.96 ± 0.01	0.92 ± 0.01	0.83 ± 0.08	0.96 ± 0.02

(a) Estimation of overlaps of the distribution of LLAMA-2 70B Chat.

	LLAMA-2 70B Chat	LLAMA-2 13B Chat	LLAMA-2 7B Chat	**Mistral 7B Instruct**	Vicuna 7B
Precision	0.82 ± 0.07	0.77 ± 0.08	0.78 ± 0.09	$\mathbf{1.00 \pm 0.00}$	0.67 ± 0.01
Recall	0.98 ± 0.01	0.93 ± 0.01	0.91 ± 0.01	$\mathbf{1.00 \pm 0.00}$	0.95 ± 0.01

(b) Estimation of overlaps of the distribution of MISTRAL 7B Instruct.

Table 1: Relative comparison of models based on Precision and Recall on the Creative text generation task. Texts generated by the first models are used as a reference against which are compared the others. The values are averaged over 5 generation seeds and the standard deviation is given. The model used as a reference distribution is in **bold**.

any notion of diversity.

Larger models are more diverse. The Recall is consistently better for larger models. Larger models have a better expressive power, and this diversity is reflected by their higher Recall.

5.4 Generating Biographies.

The results are presented in Figures 8a and 8b. We plot the Precision and the Recall as a function of the number of in-context examples.

The number of in-context examples makes the model more precise. For both pre-trained and instruction-tuned models, the Precision increases with the number of in-context examples. This is consistent with the intuition that several in-context examples help the model generate texts that are closer to the expected distribution.

Increasing the number of in-context examples make Chat models more diverse. Figure 8b confirms our previous findings that instruction-tuned models are less diverse than their pre-trained counterparts. However, by increasing the number of in-context examples, we observe a strong regain in diversity: chat models leverage the diversity in the prompt to generate more diverse texts.

Chat models plateau. After a few numbers of in-context examples, we observe that both the Precision and the Recall plateau. This suggests that the model has captured the distribution of the prompt after few examples and that adding more does not bring additional information.

5.5 Creative text generation

For this task, no reference dataset is available. Still, our novel Precision and Recall metrics offer a means to compare models, where conventional methods fall short.

Estimation of overlaps of the distribution of models. Since our metrics Precision and Recall corresponds to different support overlap, they can be used to analyze models comparatively.

Models of the same family share close distributions but size induces some differences. Table 1a shows that compared to the reference LLAMA-2 70B Chat, Precision and Recall of LLAMA-2 7B Chat and LLAMA-2 13B Chat increases with their size.

Intepreting distributions overlap. Compared to LLAMA-2 70B Chat, MISTRAL 7B Instruct has an almost perfect Precision but suffers from a lower Recall. Geometrically, this means that its distribution fits almost perfectly in the support of LLAMA-2 70B Chat, but does not cover it totally. Table 1b illustrates even further this phenomena: compared to MISTRAL 7B Instruct, every model has a quite low precision but a high recall. Geometrically, the LLAMA-2-based covers a wide portion of the support of MISTRAL 7B Instruct, but also assigns mass to points outside this distribution.

6 Conclusion

We introduced Precision and Recall as automated metrics for comparing generative model outputs against a reference distribution. These metrics serve as independent indicators of the quality and diversity of generated text. In challenging scenarios where evaluation protocols are scarce, such as tasks with high generation complexity, we demonstrated the efficacy of our metrics in assessing the quality and diversity of model outputs. Furthermore, our metrics revealed nuanced model behaviors that lie

beyond the scope of traditional evaluation benchmarks. We believe our work marks a significant step towards advancing the evaluation of generative models, offering insights to better understand their capabilities and facilitating the development of more sophisticated evaluation protocols within the community.

7 Limitations

Our study proposes to adapt the framework of Kynkäänniemi et al. (2019) from image to text generation. However, limitations of the orinal image-based frame work have been highlighted (Naeem et al., 2020; Kim et al., 2023) and may also apply to our adaptation in NLP. Notably, the sensitivity to outliers and the generalizability of these metrics across different domains warrant further investigation.

The metrics proposed in our study rely on an auxiliary model, specifically GPT-2 embeddings, to rate the quality of text generations. While we follow previous work demonstrating successful use of such embeddings, the choice of embedding may significantly impact the performance and applicability of these metrics. Future research could explore the effects of alternative embedding models on the Precision and Recall measurements in text generation tasks.

We present these metrics as valuable new instruments for evaluating a model's open-ended generation capabilities. Nonetheless, to comprehensively assess LLMs across specific dimensions and tasks, these metrics may be supplemented by other approaches.

We restricted our analysis to English language on most popular LLMs.

8 Ethical considerations

Precision and Recall are distributional-based metrics, and therefore depend heavily on the data they are tested against. If these data have biases, our metrics might unintentionally favor models that repeat these biases. This issue is critical because it can lead to unfair outcomes, especially for underrepresented groups. This highlights the need for a carefully crafted reference dataset.

To tackle this problem, we suggest a practical solution: calculate Precision and Recall for specific groups within the data, such as different genders or ethnic minorities. This helps to check whether the model is as accurate and fair for one group as it is for another. Our goal is to make sure that the models we are evaluating do not overlook or misrepresent any group. Further investigations should be conducted on this consideration.

9 Acknowledgements

This work has been partly funded through project ACDC ANR-21-CE23-0007. This project was provided with computing AI and storage resources by GENCI at IDRIS thanks to the grants 20XX-AD011014022R1, 20XX-AD011014053 and 20XX-A0151014627 on the supercomputer Jean Zay's V100/A100 partition.

References

Ahmed M. Alaa, Boris van Breugel, Evgeny Saveliev, and Mihaela van der Schaar. 2022. How Faithful is your Synthetic Data? Sample-level Metrics for Evaluating and Auditing Generative Models. ArXiv:2102.08921 [cs, stat].

Lalit Bahl, Jim Baker, Frederick Jelinek, and Robert Mercer. 1977. Perplexity : a measure of the difficulty of speech recognition task. In *Program of the 94th Meeting of the Acoustical Society of America*, volume 62:S63, Miami Beach, Florida.

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

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling.

Fasil Cheema and Ruth Urner. 2023. Precision Recall Cover: A Method For Assessing Generative Models. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, pages 6571–6594. PMLR. ISSN: 2640-3498.

Prafulla Dhariwal and Alex Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. ArXiv:2105.05233 [cs, stat].

Josip Djolonga, Mario Lucic, Marco Cuturi, Olivier Bachem, Olivier Bousquet, and Sylvain Gelly. 2020. Precision-Recall Curves Using Information Divergence Frontiers. ArXiv:1905.10768 [cs, stat].

Gemini Team Google. 2023. Gemini: A family of highly capable multimodal models.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Pum Jun Kim, Yoojin Jang, Jisu Kim, and Jaejun Yoo. 2023. TopP&R: Robust Support Estimation Approach for Evaluating Fidelity and Diversity in Generative Models. ArXiv:2306.08013 [cs].

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2019. Improved Precision and Recall Metric for Assessing Generative Models. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.* ArXiv: 1904.06991.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Generating text from structured data with application to the biography domain. *CoRR*, abs/1603.07771.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. ArXiv:1411.7766 [cs].

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. 2020. Reliable Fidelity and Diversity Metrics for Generative Models. ArXiv:2002.09797 [cs, stat].

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, page 311, Philadelphia, Pennsylvania. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. ArXiv:1912.01703 [cs, stat].

Krishna Pillutla, Lang Liu, John Thickstun, Sean Welleck, Swabha Swayamdipta, Rowan Zellers, Sewoong Oh, Yejin Choi, and Zaid Harchaoui. 2023. MAUVE Scores for Generative Models: Theory and Practice. ArXiv:2212.14578 [cs].

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. MAUVE: Measuring the Gap Between Neural Text and Human Text using Divergence Frontiers. In *Advances in Neural Information Processing Systems*, volume 34, pages 4816–4828. Curran Associates, Inc.

Tiago Pimentel, Clara Meister, and Ryan Cotterell. 2023. On the usefulness of embeddings, clusters and strings for text generator evaluation.

Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. 2018. Assessing Generative Models via Precision and Recall. In *32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada*. ArXiv: 1806.00035.

Axel Sauer, Katja Schwarz, and Andreas Geiger. 2022. StyleGAN-XL: Scaling StyleGAN to Large Diverse Datasets. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings*, pages 1–10, Vancouver BC Canada. ACM.

Loic Simon, Ryan Webster, and Julien Rabin. 2019. Revisiting precision recall definition for generative modeling. In *Proceedings of the 36th International Conference on Machine Learning*, pages 5799–5808. PMLR. ISSN: 2640-3498.

Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. ArXiv:1409.1556 [cs] version: 6.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. 2023. Consistency Models. ArXiv:2303.01469 [cs, stat].

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the Inception Architecture for Computer Vision. ArXiv:1512.00567 [cs] version: 3.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura,

Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Alexandre Verine, Benjamin Negrevergne, Muni Sreenivas Pydi, and Yann Chevaleyre. 2023. Precision-Recall Divergence Optimization for Generative Modeling with GANs and Normalizing Flows. ArXiv:2305.18910 [cs].

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. ArXiv:1910.03771 [cs].

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level Convolutional Networks for Text Classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

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

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. *SIGIR*.

A Definition of Precision-Recall Curves

In this section, we introduce another object called the PR-Curve, which is closely connected to the MAUVE Score. Then, we demonstrate using a straightforward example that the MAUVE score struggles to distinguish between a lack of diversity and a lack of quality. This difficulty can actually be explained by the shortcomings of the underlying curve used by the MAUVE score.

A.1 Definition

Witnessing the need to distinguish between quality and diversity, and making the parallel with the notion of precision and recall from classification, Sajjadi et al. (2018) introduced a new definition of Precision and Recall for discrete distributions.

Definition 2 (Precision-Recall trade-off (Sajjadi et al., 2018))**.** *A distribution Q has a precision α and a recall β with respect to a distribution P if there exists the distributions μ, ν_P and ν_Q such that:*

$$P = \alpha\mu + (1-\alpha)\nu_P \tag{3a}$$
$$Q = \beta\mu + (1-\beta)\nu_Q. \tag{3b}$$

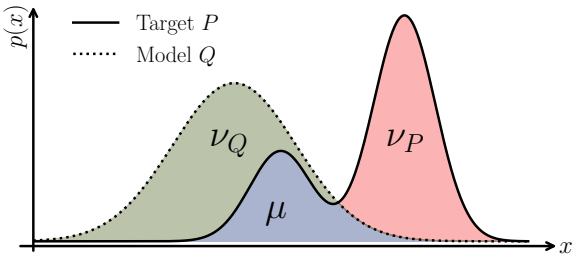


Figure A.1: Precision-Recall trade-off illustration

This definition is illustrated in Figure A.1. The distribution ν_P represents the part of P that cannot be generated by μ and thus Q, and ν_Q represents the part of P that should not be generated by μ. For every distribution μ Figure A.1 defines a couple, *i.e.* a trade-off, (α, β), and we can define the Precision Recall Curve as the set of all Pareto-optimal trade-offs. More recently, Simon et al. (2019) has extended the definition for any continuous distribution P and Q:

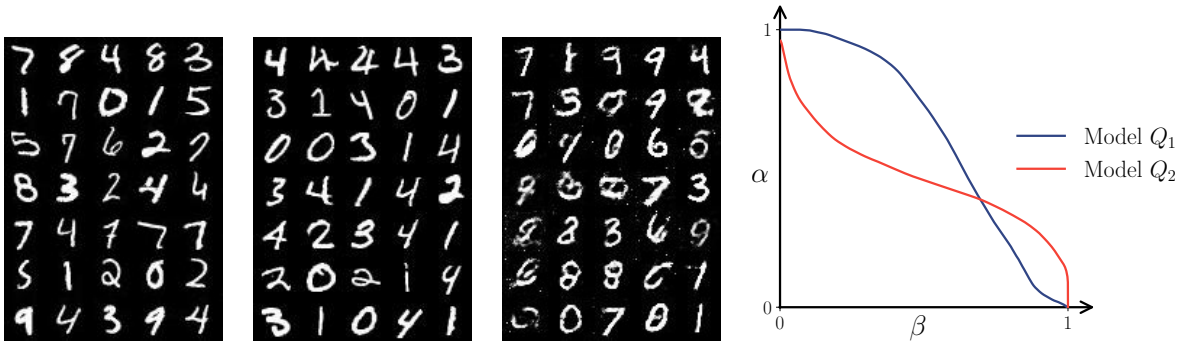
Theorem 1 (PR-Curve (Simon et al., 2019))**.** *The set of Pareto-optimal trade-offs, that is, the PR-Curve, can be represented as the set of points $(\alpha_\lambda, \beta_\lambda)_{\{\lambda\in[0,\infty]\}} \in [0,1]^2$ such that:*

$$\begin{cases} \alpha_\lambda = \mathbb{E}_Q\left[\min\left(1, \lambda\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right)\right], \\ \beta_\lambda = \mathbb{E}_P\left[\min\left(1, \frac{1}{\lambda}\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\right)\right]. \end{cases}$$

Intuitively, for any given trade-off parameter $\lambda \in [0, +\infty]$, the precision α_λ diminishes when $q(\boldsymbol{x})$ is smaller than $\lambda p(\boldsymbol{x})$ at certain points $\boldsymbol{x}$, indicating overestimation of regions by the distribution Q. On the contrary, the recall β_λ declines when $p(\boldsymbol{x})$ is smaller than $\lambda q(\boldsymbol{x})$ for specific points $\boldsymbol{x}$, suggesting underestimation of regions by Q. Although the PR-Curve is valuable for evaluating the quality and diversity of generative models, its interpretation can be challenging.

More concretely, an example of PR-Curves is depicted in Figure A.2. Figure A.2a shows samples drawn from the reference distribution P. Here, Q_1 represents a distribution capable of generating high-quality images, albeit only for the first 5 digits, while Q_2 represents a distribution that generates noisy images on all labels. The PR-Curve of Q_1 surpasses that of Q_2 for high values of λ (top left), indicating the superior quality of Q_1. Conversely, for low values of λ (lower right), the PR-Curve of Q_2 exceeds that of Q_1, indicating greater diversity in Q_2.

For a two-number summary of these curves, Sajjadi et al. (2018) has proposed the F_γ-scores: $\mathrm{F}_\gamma = \max_\lambda (1+\gamma)^2\alpha_\lambda\beta_\lambda/(\gamma^2\alpha_\lambda + \beta_\lambda)$. The $\mathrm{F}_{1/8}$ is employed to assess the quality of the model, and F_8 score is used to assess the diversity. Abusing the notation, the scores $\mathrm{F}_{1/8}$ and F_8 are often refered as the Precision and the Recall. Note that in this work, we use another two-number summary of the PR-Curve - $\alpha_\infty = Q(\mathrm{Supp}(P))$ and $\beta_0 = P(\mathrm{Supp}(Q))$ - which is practically easier to estimate.



(a) Reference distribution P. (b) Distribution Q_1. (c) Distribution Q_2. (d) PR-Curves of Q_1 and Q_2 w.r.t P.

Figure A.2: PR-Curves for the MNIST dataset. The target distribution P is shown in (a), and the distributions Q_1 and Q_2 are shown in (b) and (c) respectively. The PR-Curves of Q_1 and Q_2 are shown in (d). The PR-Curve of Q_1 indicates high quality and limited diversity and the PR-Curve of Q_2 indicates limited quality but high diversity.

A.2 Connection with the MAUVE score

The PR-Curve bears a close relationship with the MAUVE score (Pillutla et al., 2021, 2023), which represents the area under the curve of a modified version of the PR-Curve introduced by Djolonga et al. (2020). On our AG News experiment, fully described in Section 5.2, we can exemplify some limitations of this approach. Specifically, Figure A.3 shows that the underlying curves used for computing the MAUVE score can hardly differentiate the discrepancy highlighted in Figure 2, when the original PR-Curves from Sajjadi et al. (2018) can. When adapting the $\mathrm{F}_{1/8}$ and F_8 scores to the MAUVE curves, and plotting the results on Figure A.4 we find that these adapted scores are substantially less meaningful than those from (Sajjadi et al., 2018).

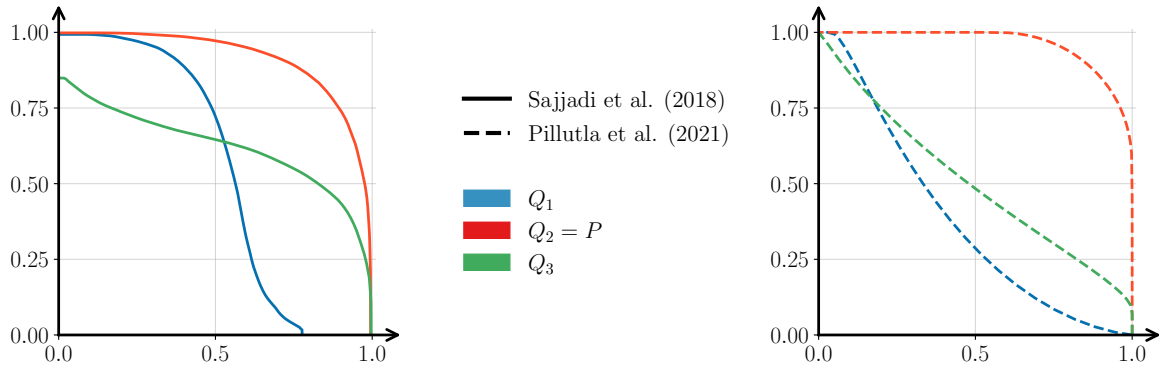


Figure A.3: Comparison of the PR-Curve and the MAUVE score for the AG News dataset. The PR-Curve is a finer representation of the MAUVE score. However, the PR-Curves introduced by Sajjadi et al. (2018) clearly reflects the lack of diversity or lack of quality.

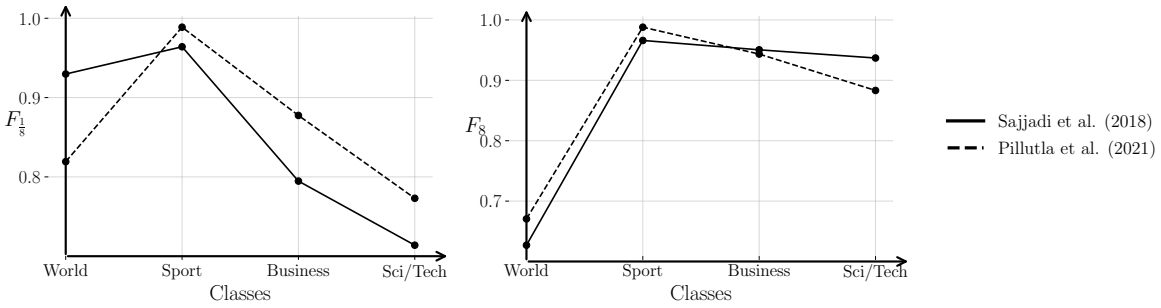


Figure A.4: The $\mathrm{F}_{1/8}$ and F_8 scores for the AG News dataset. The reference consists of texts of `World, Sport` topics. We plot the evolution of the scores with respect to this reference as we add topics to the candidate distribution. Scores from Pillutla et al. (2021) do not identify the lack of recall of when topics are missing from the candidate distribution.

Intuitively, we could use the definition of PR-Curve introduced by Sajjadi et al. (2018) and use the $\mathrm{F}_{1/8}$ and the F_8 scores to assess the quality and diversity of the models. However, while being a powerful theoretical tool to assess the quality and diversity of generative models, the PR-Curve is 1) more difficult to interpret and 2) encapsulating the PR-Curves with F_γ can fail to distinguish the quality and diversity of the models. In particular, Kynkäänniemi et al. (2019) illustrates how these scores can fail to capture the discrepancy. Also, we illustrated on generation on WebText, detailed in Section 5.3, we show how the score $\mathrm{F}_{1/8}$ and F_8 are not able to differentiate the models.

While the PR-Curve, as defined by Sajjadi et al. (2018), coupled with $\mathrm{F}1/8$ and $\mathrm{F}8$ scores, provides a robust theoretical framework for assessing both the quality and diversity of generative models, its interpretation can be challenging. Furthermore, encapsulating the PR-Curves with F_γ scores may fail to adequately distinguish the quality and diversity of models. Notably, Kynkäänniemi et al. (2019) have demonstrated instances where these scores can overlook significant discrepancies.

In our examination of generation on WebText (Section 5.3), we offer a practical illustration of these limitations, illustrated on Figure A.5. Specifically, $\mathrm{F}_{1/8}$ and F_8 correlate more with Precision than with Recall and strongly correlate with each other.

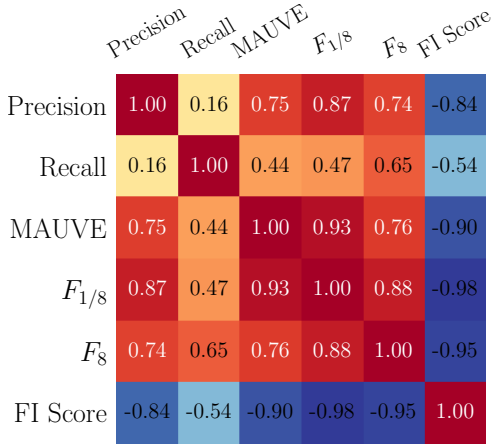


Figure A.5: Correlation between the Precision, the Recall, the MAUVE score (Pillutla et al., 2021) and the $\mathrm{F}_{1/8}$ and F_8 scores Sajjadi et al. (2018) on WebText. $\mathrm{F}_{1/8}$ and F_8 correlate more with Precision than with Recall and strongly correlate with each other.

B Experiments

This section gives full details on the experimental protocol we adopted in the paper.

Texts input length. All tested models do not share the same tokenization. To avoid a potential bias due to different generations lengths, input and output length constraints are expressed in terms of words rather than tokens. The average number of token per words has been calculated for each model on the train set of the WebText dataset[3] and reported in Table B.1.

Hardware and software. Generations were conducted on either A100 GPUs with 80Gb memory or on V100 with 32Gb memory. Models were run with a bfloat16 precision on A100 and with float32 precisions on V100, except for the PYTHIA models, which have been ran with float16 mixed precision, following the original setup (Biderman et al., 2023). For models with $\approx 7B$ parameters, generating 4000 samples takes approximately 5 hours on 2 A100s, it takes about 10 hours for models with $\approx 13B$ parameters on 2 A100, and about 40 hours for models with $\approx 70B$ parameters on 4 A100. Generating the samples required approximately 4200 GPUs hours. Code is built on PyTorch (Paszke et al., 2019), with the HuggingFace library (Wolf et al., 2020).

Computational cost of Precision and Recall. Since Precision and Recall are based on a k-NN computations, computing these metrics is especially fast, even across large datasets. Once the features are computed, computing Precision and Recall takes less than 5 seconds on a standard laptop for 4000 samples, instead computing MAUVE takes approximately 40s.

Other metrics implementations. For computing SelfBLEU, Disctinct-N and MAUVE we relied on the official implementations found in `https://github.com/koadman/mauve`, or `https://github.com/krishnap25/mauve-experiments`.

Implementations. The code used to run and evaluate the generations, as well as the exact datasets is provided in the supplementary materials.

Model	Average Tokens per Word
MISTRAL	1.313
PYTHIA	1.183
LLAMA-2	1.360
VICUNA	1.360

Table B.1: Average number of tokens per word for different families of models. VICUNA is based on LLAMA-2.

B.1 WebText Generation

Data. Reference datasets is extracted from the official OpenAI WebText dataset at `https://github.com/openai/gpt-2-output-dataset`, under a MIT license and is composed of articles extracted from WEB urls. We extracted 15k samples from this datasets for our experiments.

Input formatting. Non-chat models are simply prompted with the first 10 words of random WebText articles. For chat models, they are explicitly asked to continue the prompt. We input the last 10th word out of the instruction tokens to avoid idiomatic expressions such as *Sure! Here is* ... and ensure fair comparison between models. The different prompts are illustrated in Table B.2.

Generation setup. For this task, we used the default recommended generation setup, i.e. nucleus sampling for all models, with parameters displayed in Table B.3. LLAMA-2 default parameters are from the official repository `https://github.com/facebookresearch/llama/blob/main/example_text_completion.py` and MISTRAL's ones from their official API `https://docs.mistral.ai/api/`. For PYTHIA's models, we did not find any mention of recommenced generation parameters and hence used a temperature and nucleus p of 1.0.

[3]Available at `https://github.com/openai/gpt-2-output-dataset`

Table B.2: Input Formats for Chat and Non-Chat Models

Model	WebText input Template

LLAMA-2 CHAT, MISTRAL CHAT	`[INST] Continue the following text:`
	
	`{{first 9 words}} [/INST] {{10th word}}`

VICUNA	`USER: Continue the following text:`
	`{{first 9 words}}`
	`ASSISTANT: {{10th word}}`

Non-chat models	`{{first 10 words}}`

Table B.3: Generation Parameters Summary

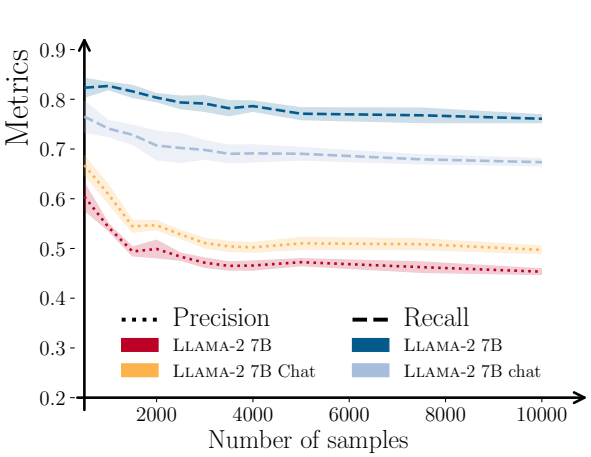
Model	Max New Tokens	Nucleus P	Temperature	Repetition Penalty
Llama2, Vicuna	448	0.9	0.6	1.18
Mistral	432	1.0	0.7	1.18
Pythia	390	1.0	1.0	1.18

Evaluation setup. To determine the optimal number of samples and the parameter k for estimating Precision and Recall, we analyze the behavior of these metrics with varying N and k. We employ the WebText generation task with LLAMA-2 7B Chat as a reference. Specifically, we generate $10,000$ samples per seed and compute the average Precision and Recall for N ranging from 100 to $10,000$, and k ranging from 1 to 30. The trends of Precision and Recall with respect to the number of samples N, while k is fixed at 4, are depicted in Figure B.1a. Conversely, the trends of Precision and Recall with respect to the parameter k, while N is fixed at $4,000$, are shown in Figure B.1b. We observe that Precision and Recall stabilize as N increases, reaching a plateau around $N = 3,000$. Thus, we set the number of samples at $N = 4,000$ for subsequent experiments. Additionally, Precision and Recall increase towards 1 as k increases, aligning with the intuition that the estimated support is more covering as k increases. Therefore, we set $k = 4$ as it facilitates a clear differentiation between the two models. We also include the standard deviation for all models described in Table B.3. Standard deviations are computed by varying the output set, the target set, and both. These results are summarized in Table B.4. We recommend consistent references for all models to ensure a fair comparison, as the standard deviation can significantly impact the results.

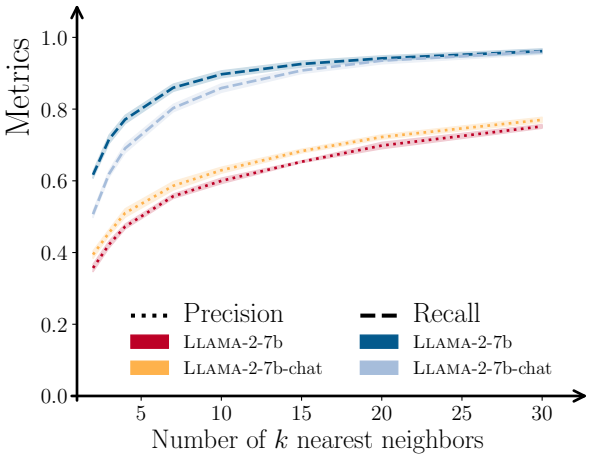
Additional results. We report MAUVE scores on WebText on Figure B.2. Contrary to what is reflected by Precision and Recall, pre-trained and instruction-tuned models exhibit similar MAUVE scores.

Table B.4: Standard deviation for seeding different sets of generated samples, different set of references samples. The standard deviation are averaged over the different models and their standard deviation are given.

	Varying Q	Varying P	Varying P and Q
$\sqrt{\mathrm{Var}(\mathrm{Precision})}$	0.005 ± 0.001	0.024 ± 0.010	0.019 ± 0.008
$\sqrt{\mathrm{Var}(\mathrm{Recall})}$	0.011 ± 0.003	0.006 ± 0.002	0.013 ± 0.004



(a) Evolution of the metrics Precision and Recall as the number of samples N increases. The shaded area represents the standard deviation computed over 5 random seeds for the set ouputs. The averages reach a plateau around $N = 3000$.



(b) Evolution of the metrics Precision and Recall as the parameter k increases. The shaded area represents the standard deviation computed over 5 random seeds for the set ouputs.

Figure B.1: Evolution of the metrics Precision and Recall as the number of samples N and the parameter k increases.

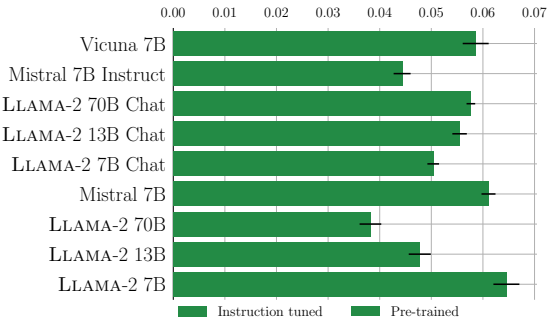


Figure B.2: MAUVE score on WebText. Dark bars represents standard error over the generation and reference seeds. No clear distinction between pre-trained and instruction-tuned models.

B.2 Wikipedia Biographies Generation

Data. We took inspiration from the WikiBio dataset (Lebret et al., 2016) and extracted Wikipedia articles corresponding to humans. We kept only the articles with a "Good" or "Featured" badge, which are considered as the best quality articles. We then extracted the summaries of these pages to use as reference biographies. Finally, we kept only articles with at least 80 words and at most 350 words, for a total of 6637 texts. Our processed dataset and the code use for its construction is available in the supplementary materials.

Input formatting. Pre-trained and chat models are prompted with a small instruction and a variable number of in-context examples. The different prompts are illustrated in Table B.5.

Generation setup. We use the default model generation parameters, as described in Appendix B.1. However, we set the maximum number of tokens generated for all models to 448 for LLAMA-2-based models (including VICUNA), 432 for MISTRAL and 390 for PYTHIA, respecting the average number of tokens per word reported in Table B.1.

B.3 Creative texts generation

Data. The full list of creative instructions is presented in Listing 1.

Listing 1: List of Creative Prompts

```
Write about a dream you had.
```

Table B.5: Input Formats for Chat and Non-Chat Models

Model	Wikipedia Biographies input Template

LLAMA-2 CHAT, MISTRAL CHAT	`[INST] Write biographies of various people.`
	
	`Here are a few examples:`
	`- Biography of {{name}}: {{content}}`
	`{{n times}}`
	`[/INST]`
	`- Biography of`

VICUNA	`USER: Write biographies of various people.`
	`Here are a few examples:`
	`- Biography of {{name}}: {{content}}`
	`{{n times}}`
	`ASSISTANT:`
	`- Biography of`

Non-chat models	`Write biographies of various people.`
	`- Biography of {{name}}: {{content}}`
	`{{n times}}`

```
Create and write about a new character.
Write about a place you'd love to visit.
Write about an important life event.
Write about life 100 years from now.
Write a story where magic exists in everyday life.
Write a poem about a personal experience.
Write a speech for a cause you believe in.
Write a short mystery story.
Write a modern day fairy tale.
Write a story set in a historical period.
Write a story about a technological advancement.
Write a letter to your future self.
Write a story from an animal's perspective.
Write a week's worth of diary entries for a character.
Write a short story using mythological characters.
Write a conversation between two characters.
Write about a day in your life.
Write a series of Haikus about seasons.
Describe a place without naming it.
Write a news article about an event in your town.
Write a recipe with a story.
Write a one-act play.
Write a story set in a dystopian future.
Write a satirical essay on a trending topic.
Write a humorous comic strip.
Write a piece inspired by nature.
Write a short story about your favorite fictional character.
Write instructions for an invented machine.
Write a story using only metaphors.
Write a script for a short film.
```

```
Write a conversation in text message format.
Write a limerick about a funny event.
Write a story about a journey to space.
Write a script for a documentary on a subject of your choice.
Write a eulogy for a character from your favorite book.
Write a script for a radio show.
Write a song about a memorable event.
Write a story with a hidden meaning.
Write a fable with a moral.
Write a comedic monologue.
Write an opinion piece on a current event.
Write a parody of a popular book or movie.
Write a story where a character has a magical power.
Write a story about time travel.
Write a story based on an object in a mystery box.
Write a ghost story.
Describe a haunted house without using cliche descriptions.
Write a story about a conspiracy theory.
Write a story about a non-traditional superhero.
```

C License

Models weights. Llama-2 weights are released under the licence available at https://ai.meta.com/llama/license/. Mistral models and weights are released an Apache 2.0 licence. Vicuna models and weights are released under an Apache 2.0 licence. Pythia models and weights are released under an Apache 2.0 licence.

Data. The WebText dataset we used is released under a MIT License. Wikipedia's articles are under the reative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA) and the GNU Free Documentation License (GFDL). AG news is available at https://huggingface.co/datasets/ag_news but no licence are disclosed.

D Wikipedia Biographies samples

Albert Einstein:
Albert Einstein (born March 14, 1879 - died April 18, 1955) was a German-born theoretical physicist who developed the theory of relativity, one of the two pillars of modern physics (the other being quantum mechanics). His work is also known for its influence on the philosophy of science. He is best known to the general public for his mass-energy equivalence formula E = mc2, which has been dubbed "the world's most famous equation". He received the 1921 Nobel Prize in Physics "for his services to Theoretical Physics, and especially for his discovery of the law of the photoelectric effect", a crucial step in understanding quantum phenomena.
Einstein was born into a Jewish family in Ulm, Wurttemberg, Germany. In 1896, he enrolled at the Swiss Federal Polytechnic in Zurich to study mathematics and physics. After graduating, he worked as a technical assistant in the patent office in Bern while working on his theories in his free time. He published several papers during this period that brought him to the attention of scientific communities in Europe. In 1905, Einstein published four groundbreaking papers that changed our understanding of space, time, energy, and matter. These papers included special relativity, the photoelectric effect, Brownian motion, and mass-energy equivalence.
Throughout his career, Einstein held academic positions in Switzerland, Germany, and the United States. He became a professor at the University of Berlin in 1914 but left two years later when Adolf Hitler came to power. He moved to the United States and joined the faculty of the Institute for Advanced Study in Princeton, New Jersey, where he lived until his death. Throughout his life, Einstein continued to publish influential works in both pure and applied physics, including the development of unified field theory, which attempted to explain gravitational forces using electromagnetic principles.

Marie Curie:
Marie Sklodowska-Curie (November 7, 1867 - July 4, 1934) was a Polish and naturalized-French physicist and chemist who conducted pioneering research on radioactivity. She was the first woman to win a Nobel Prize, the only person to win twice in multiple sciences, and the only person to win in multiple sciences for a single discovery. Born Maria Sklodowska in Warsaw, Poland, she studied mathematics and physics in Paris, where she met her future husband Pierre Curie. Together they discovered the elements polonium and radium, for which they were awarded the Nobel Prize in Physics in 1903. After Pierre's death in 1906, Marie continued her research alone and won the Nobel Prize in Chemistry in 1911 for her discoveries regarding the elements. During the First World War, she established the Radium Institute in Paris to develop mobile radiography units for soldiers. She also worked tirelessly to establish the Curie Institutes in Paris and Warsaw. She continued her scientific work until her death from aplastic anemia, believed to be caused by her long-term exposure to radiation.

Marie Curie:
Marie Sklodowska-Curie (November 7, 1867 - July 4, 1934) was a Polish and naturalized-French physicist and chemist who conducted pioneering research on radioactivity. She was the first woman to win a Nobel Prize, the only person to win two Nobel Prizes in different sciences, and the only person to win a Nobel Prize twice in the same field. Born in Warsaw, Poland, Curie moved to Paris at age 24 to study physics and mathematics at the Sorbonne University. She met her future husband, Pierre Curie, while working together in the laboratory. They discovered the elements polonium and radium and shared the 1903 Nobel Prize in Physics with Henri Becquerel for their work on radioactivity. Marie went on to receive the 1911 Nobel Prize in Chemistry for her discovery of the elements radium and polonium and their isolation in pure form. In addition to her Nobel Prizes, Curie was also named a Commander of the Legion of Honor and received numerous other honors throughout her career. She became the first professor at the Sorbonne in 1934, just one month before her death from aplastic anemia, believed to be caused by long-term exposure to radiation. Her achievements paved the way for future developments in nuclear science and technology.

Marie Curie:
Marie Sklodowska-Curie (7 November 1867 - 4 July 1934; French: [maRi sklo'doska]; Polish: [ma'rje skwo'dofska]), born Maria Salomea Sklodowska, was a Polish and naturalized-French physicist and chemist who conducted pioneering research on radioactivity. She is the first woman to win a Nobel Prize, the only person to win two Nobel Prizes in different sciences, and the only person to win in multiple sciences (Physics and Chemistry). Born into a Polish family in Warsaw, she studied physics and mathematics at Warsaw University before moving to Paris to continue her studies in science. There, she earned her higher education degrees and conducted groundbreaking research on radioactivity alongside her husband Pierre Curie, whom she married in 1895. Their work led to the discovery of the elements polonium and radium. In 1903, they received the Nobel Prize in Physics for their joint research on radiation. Marie continued her work alone following Pierre's death in 1906 and went on to discover the element actinium in 1908. She won the Nobel Prize in Chemistry in 1911 for this achievement. During World War I, she established the Radium Institute in Paris and developed mobile X-ray units for use in treating wounded soldiers on the battlefield. These units became known as Little Curies and saved countless lives. Following the war, Marie moved back to Poland and spent time teaching and conducting research at the University of Warsaw. However, due to political instability and financial difficulties, she returned to France in 1934. Tragically, she died there just three months later due to aplastic anemia caused by prolonged exposure to radium. Her achievements and legacy have inspired generations of scientists and continue to shape

Table D.1: Generated biographies from Mistral Instruct v0.2, with 1 in-context example. The model generates always the same biography, illustrated by a low Recall. The precision is also very low, because Marie-Curie and Einstein are scientific personalities, who are not widely represented in the reference dataset, compared to other occupations.

Herman Wirth:
Hermann Wirth (1 September 1885 - 12 March 1981) was a German historian and philologist. He was professor of Indo-European studies at Bonn University from 1920 to 1933, where he taught among others Hans Freudenthal, Ernst Haefliger, Erich Schenk, Wilhelm Schulze, Johannes Schmidt, Werner Voigt, Heinrich Vietzen, Wolfgang Golther, Martin Doutte, Fritz Graf, Paul Maas, Karl Meisenheim, Rudolf Much, Otto Neugebauer, Bruno Obermaier, Helmut Paech, Richard Pittioni, Gerhard Pommeranz-Wardbor, Ulrich Raulff, Albert Siebeck, Kurt Steinmann, Alfred Thayer Mahan, Max Vasmer, Eduard Frauwallner, Robert Havemann, and Hermann Schneider. From 1933 to 1935 he directed the Kaiser Friedrich Museum in Berlin. During World War II he continued his research in Paris under the pseudonym "Dr. Schmitt". He was the founder of the European School of Folklore Studies (Europaische Hochschule fur Volkskunde) in Salzburg. In 1928 Wirth founded the journal Zeitschrift fur vergleichende Sprachforschung und verwandte Gebiete ("Journal for Comparative Philology"). Its editorial board included prominent linguists such as Georges Dumezil, Leo Wiener, Karl Brugmann, Johann Christian Bekker, Emil Siegfried Engelsvater, Alexander Kappelmacher, Carl Darling, Johannes Droys, Jacob Grimm, Hugo Koch, Georg Curtius, Julius Pokorny, Gustav Reichen, and Bernhard Wiese.
Other languages
* Arabic:

George Cunningham:
George John Cunningham (23 February 1905 - 22 December 1995) was a Scottish footballer who spent all but two seasons of his senior football career playing for Manchester City. He scored 12 goals from 32 appearances in representative matches for Scotland, and was inducted into the English Football Hall of Fame in 2002. He went on to manage clubs such as Burnley and Derby County, and was part of the management team when the latter won promotion to the Premier League in 2007.
Sports

Yuriy Kuleshov:
Yury Petrovich Kuleshov (Ukrainian: Iurii Petrovich Kuleshov; 2 January 1922 - 17 February 2009) was a Ukrainian football defender. He was a squad member for the Soviet Union national football team at the 1952 Summer Olympics. He was a former Dynamo Kyiv - 1942, FC Dynamo 1943, 1942, 1944, Dynamo Kyiv 1945. He was awarded the title Master of Sport USSR 1956.

Danielle Bregoli:
Danielle Michelle Bregoli (/'breg@li:/, BRUG-@-lee; born March 26, 2003), better known online as Bhad Bhabie /baed b@bi:/, is an American rapper, singer, songwriter, internet personality and businesswoman. Born in Boynton Beach, Florida, Bregoli gained fame through her appearance on Dr Phil in September 2016, wherein she told the audience that she wanted to "clap back" at people who disrespected her mother. This phrase subsequently became popularized and has been referenced in numerous media outlets, songs, TV shows and films since then. Since appearing on Dr Phi, Bregoli has achieved commercial success as a recording artist, having reached number one on the Billboard Hot 100 chart with "These Heaux". As of May 2022, Bregoli has amassed over 1 billion views across her YouTube channel. In addition to music, Bregoli has launched clothing lines and beauty brands, as well as acting roles.

Table D.2: Generated biographies from Llama-2-7b, with 11 in-context examples. The samples are of low quality, illustrated by a low Precision. On the contrary, the Recall is high, as the model generates a wide variety of biographies.