Uncertainty Estimation in Large Language Models to Support Biodiversity Conservation

María Mora-Cross and **Saúl Calderón-Ramírez**
Costa Rica Institute of Technology
{maria.mora, sacalderon}@itcr.ac.cr

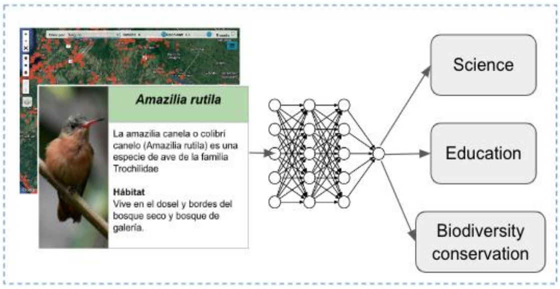
Abstract

Large Language Models (LLM) provide significant value in question answering (QA) scenarios and have practical application in complex decision-making contexts, such as biodiversity conservation. However, despite substantial performance improvements, they may still produce inaccurate outcomes. Consequently, incorporating uncertainty quantification alongside predictions is essential for mitigating the potential risks associated with their use. This study introduces an exploratory analysis of the application of Monte Carlo Dropout (MCD) and Expected Calibration Error (ECE) to assess the uncertainty of generative language models. To that end, we analyzed two publicly available language models (Falcon-7B and DistilGPT-2). Our findings suggest the viability of employing ECE as a metric to estimate uncertainty in generative LLM.

The findings from this research contribute to a broader project aiming at facilitating free and open access to standardized and integrated data and services about Costa Rica's biodiversity to support the development of science, education, and biodiversity conservation.

1 Introduction

The signatory countries of the Convention on Biological Diversity (CBD) of the United Nations (UN) have committed to safeguarding and sustainably using the planet's biodiversity (United Nations, 1992). However, countries lack comprehensive data and the application of biodiversity knowledge in decision making has been limited (Secretariat of the Convention on Biological Diversity, 2020). Much of the required data to address this need are in text format and are part of the globally available taxonomic literature.



Amazilia rutila

La amazilia canela o colibrí canelo (Amazilia rutila) es una especie de ave de la familia Trochilidae

Hábitat
Vive en el dosel y bordes del bosque seco y bosque de galería.

Figure 1: An innovative and reliable system developed to facilitate free and open access to the Costa Rica's biodiversity data often hindered by its textual format. The primary goal is to advance science, education, and biodiversity conservation.

Taxonomic literature keeps records of the planet's biodiversity and gives access to the knowledge needed for research and sustainable management. The number of publications generated is quite large: the corpus of biodiversity literature includes tens of millions of figures, taxonomic treatments, and other technical documents. Unfortunately, most of the taxonomic literature is in text format. The Biodiversity Heritage Library (Gwinn and Rinaldo, 2009), the world's largest open access digital library for biodiversity literature and archives, integrates more than 61 million digitized pages. Additionally, our local project, that works with data on the biodiversity of Costa Rica, has a database with over 32 thousand records containing information such as scientific name, morphological description, common names, species distribution, life cycle, feeding, reproduction, demography, habitat, phenology, behavior, interactions, conservation status, and legislation, among other types of content. Obtaining highly structured records from digitized text has been shown to be complex and very expensive (Cui et al., 2021; Mora and Araya, 2018). Also, there is much left to document. The scientific community has described over 1.2 million species, but studies suggest that 86% of existing species on Earth and 91% of species in the ocean still await description (Mora et al., 2011). The published descriptions synthesize observations

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track), pages 368–378
June 16-21, 2024 ©2024 Association for Computational Linguistics

made by taxonomists over centuries of research and include detailed morphological aspects (i.e., shape and structure) of species useful to identify specimens and to improve information search mechanisms. Other tasks include: data analysis of species having particular characteristics and comparison of species descriptions (Mora et al., 2023).

Generative language models are valuable in scenarios lacking a predefined answer, notably in high-risk and complex contexts like biodiversity conservation or medical diagnosis (Doi et al., 2023). Despite improvements, they may still produce inaccuracies. Incorporating uncertainty measures with predictions helps mitigate risks in decision-making and other applications (Kim et al., 2022; Jiang et al., 2021).

Uncertainty estimation refers to the process of assessing and quantifying the degree of unreliability or lack of confidence associated with a particular measurement, prediction, or decision. Uncertainty is present in all phases of the machine learning pipeline in Natural Language Processing (NLP) applications. Uncertainty can originate in the acquisition and preprocessing of data (random uncertainty) and in the design and training of the model (epistemic uncertainty). This gives rise to different ways of measuring uncertainty, depending on the aspects that are taken into account and the application area (Mena et al., 2021). Uncertainty quantification has been identified as a key unsolved challenge in LLM for text generation (Kuhn et al., 2023; Hendrycks et al, 2022; Jiang et al., 2021; Amodei et al., 2016). This field of research has advanced widely in other application areas such as image pattern recognition (Fathullah and Gales, 2022; Mena et al., 2021; Gal, 2016) or text classification (He et al., 2020; Xiao and Yang, 2018). However, there has been a notable low contribution in this field that specifically addresses calibration within a regression framework. In the field of generative language models, the challenges are unique because the outputs are presented in free-form text. In language, the semantic content of a sentence and its syntax play an important role in the meaning and there are many ways to generate a correct text (Kuhn et al., 2023).

Uncertainty estimation enhances decision-making in biodiversity conservation and other areas. In the context of QA models for biodiversity conservation, uncertainty quantification is useful to determine whether the model's reply is reliable. If the model's responses are frequently unreliable, this might suggest that a re-training or finetuning of the model is necessary. Hallucination is a frequent shortcoming of LLMs. Uncertainty quantification can help to detect whether the LLM might be hallucinating. Knowing the model uncertainty associated with a response encourages responsible use and is essential to building trust in users who are aware of the limitations of the information provided by the model.

The main objective of this paper is to present an exploratory analysis of the application of MCD and ECE to assess the uncertainty of DistilGPT-2 (Radford et al., 2019) and Falcon-7B (Almazrouei et al., 2023). Our results indicate the feasibility of utilizing ECE as a metric for assessing uncertainty in generative LLM. We evaluate the usage of two methods for uncertainty quantification Perplexity and MCD. To evaluate the reliability of the uncertainty scores, we use the ECE against the BERTScore (Zhang et al., 2020) of the generated responses compared against a ground truth dataset.

The findings from this research contribute to a broader project aiming at facilitating free and open access to standardized and integrated data and services about the biodiversity of Costa Rica to support the development of science, education, and biodiversity conservation (see Figure 1).

Our contributions can be summarized as follows:

- An exploratory analysis of the application of MCD combined with ECE to assess the reliability of uncertainty estimates in generative LLM within a closed technical domain.
- Uncertainty estimation in generative LLM applied to biodiversity data represents an emerging field of research. Currently, there are no publications available in this specific application domain. This project aims to contribute insights for establishing reliable generative QA models, fostering advancements in science, education, and biodiversity conservation.

2 Background

Recent efforts to assess uncertainty in generative LLM have introduced various algorithms. One approach, as proposed by Zhou et al. (2024) and Lin et al. (2022), involves training a model to generate an uncertainty estimate alongside the generated text. Another strategy, suggested by Kuhn et al. (2023), utilizes clustering techniques to estimate uncertainty. In their work, an

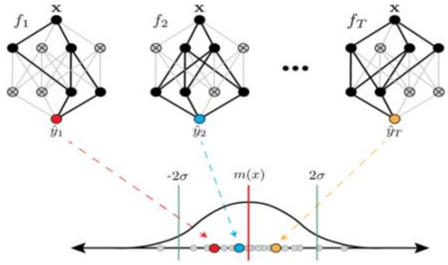


Figure 2: Diagram illustrating the operation of the MCD: each network produces a different output by randomly turning off a group of neurons (circles with x) in each forward propagation, this simulates the effect of an ensemble of models. Multiple forward passes with different Dropout settings produce a predictive distribution which approximates the posterior predictive distribution of the original network (image: Van Katwyk et al., 2023).

unsupervised method known as semantic entropy is introduced. This method leverages linguistic invariances to group texts based on shared meanings, offering an uncertainty estimation using clustering. In addition, Duan et al. (2023) have proposed a method that takes into consideration the relevance of tokens within a sentence or the relevance of sentences within a text. Their approach involves implementing an attention shift mechanism that adjusts attention based on the importance of tokens or sentences, ultimately influencing how uncertainty is estimated in the generated language model.

2.1 Uncertainty estimation

Uncertainty quantification methods according to Mena et al. (2021) can be grouped mainly into Monte Carlo Dropout, Ensemble, and Variational Inference Methods.

MCD methods estimate the conditional probability density by sampling a number of evaluations of the model with the same input and analyzing the distribution of those evaluations. Once this distribution is obtained, it is possible to use uncertainty measures such as entropy or variance to estimate the uncertainty associated with the selected response. In general, the greater the entropy or the variance of the probability distribution, the greater the uncertainty associated with the selected response (Mena et al., 2021).

2.2 Monte Carlo Dropout

Proposed originally by Gal and Ghahramani (2016), this method approximates the conditional distribution by sampling N evaluations of the model $f_{\theta_i}(x)$ with a dropout rate d. An ensemble of N models $f_{\theta_1}(x), f_{\theta_2}(x), \dots, f_{\theta_N}(x)$ is evaluated with a different set of randomly disabled neurons θ_i. In the case of the generative QA model, the logits scores for each token $y_i = f_{\theta_i}(x)$ are used to calculate the variance of the conditional distribution (Figure 2). This is used as the score $s_{MCD}(x_i)$.

2.3 Model Calibration

Calibration algorithms are designed to harmonize the predicted probabilities or confidence scores generated by a model with the real-world results. Some calibration metrics includes ECE. ECE is a metric commonly used in classification problems to assess the calibration of a probabilistic model by comparing the predicted probabilities to the true outcomes in the following way:

$$ECE = \sum_{b=1}^{B} \frac{n_b}{N} \left|acc(b) - conf(b)\right| \qquad (1)$$

The measure involves splitting the data into B equally spaced bins, where n_b corresponds to the number of predictions in bin b, N is the total number of data points, and $acc(b)$ and $conf(b)$ are the accuracy and confidence of bin b respectively. In a regression problem, where continuous values are predicted rather than discrete classes, it is possible to adapt the concept of calibration to assess the accuracy of predictions computing the absolute difference between the mean predicted value and the mean true value and calculate the weighted average of the calibration errors across all bins:

$$ECE = \sum_{b=1}^{B} \frac{n_b}{N} \left|avg(acc(b)) - avg(conf(b))\right| \qquad (2)$$

Finally we evaluate the computed calibration error to determine how well the model's predicted probabilities align with the actual outcomes. A lower Calibration Error indicates better calibration (Naeini, 2015).

In the context of this work, the confidence is meant to be estimated with MCD and perplexity methods. As our task at hand is not a classification problem, we aim to use the average BertScore using the groundtruth as reference for the model responses. The ECE then measures the linear correlation between the uncertainty score and the BERTScore in our case.

2.4 Generation Quality

Evaluating the text generated by a model is complex because there is no absolute ground truth. The best way to evaluate generated text is with the help of humans, but it is costly and does not guarantee reproducibility due to the bias introduced by individuals and the sampling process (Papineni, 2002). Some of the most used metrics for evaluating the quality of generated text include BLEU or Bilingual Evaluation Understudy (Papineni, 2002), ROUGE or Recall-Oriented Understudy for Gisting Evaluation (Lin, 2004), BERTScore, METEOR (Banerjee and Lavie, 2005), and Self-BLEU (Zhu et al., 2018).

We employ BERTScore as our primary metric for evaluating the generated text. BERTScore is a metric that leverages BERT embeddings and computes cosine similarity for each token in the candidate sentence with each token in the reference sentence. BERTScore correlates better with human judgments and, in some applications, provides better performance than existing metrics (Zhang et al., 2019).

One of the most used metrics to measure confidence in generative language models is **Perplexity.** Perplexity is a measure used to assess the quality of text generated by a probabilistic model. Quality is usually measured in terms of coherence and predictability. Perplexity quantifies how well the model predicts or represents a given dataset. The range of this metric is [0, inf). To incorporate this metric into the project, the obtained results were rescaled to fall within the range of [0,1]. The following formula compute the perplexity of the generated text:

$$PPL(X) = e^{\left(-\frac{1}{t}\sum_{i}^{t} log\, p_{\theta}(x_i|x_{j<i})\right)} \qquad (3)$$

Perplexity can be thought as how surprised the model is when evaluating a token sequence. Therefore, it can also be leveraged as an uncertainty quantifier. We evaluate its usage against MCD.

3 Methodology

3.1 Experimental Design

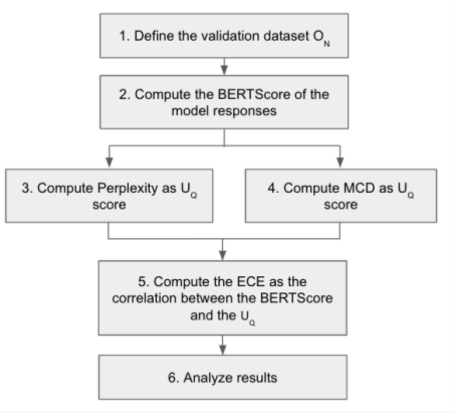
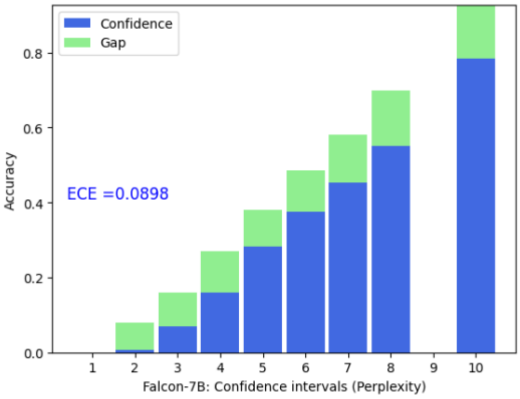

1. Define the validation dataset O_N
2. Compute the BERTScore of the model responses
3. Compute Perplexity as U_Q score
4. Compute MCD as U_Q score
5. Compute the ECE as the correlation between the BERTScore and the U_Q
6. Analyze results

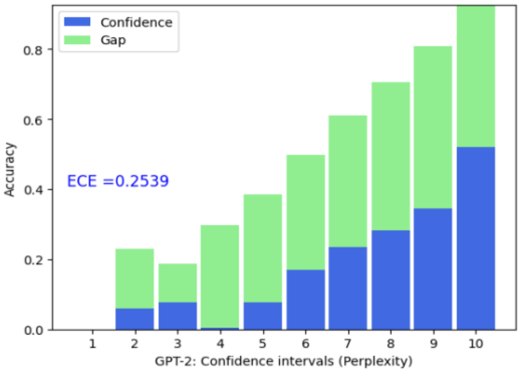
Figure 3: Project workflow: 1) The first stage involves randomly selecting 500 records from the ELI5Sci dataset. 2) For the selected records, model confidence is calculated using Perplexity and to evaluate results, BERTScore is applied. 3) 10 samples are generated for each selected record. 4) Finally, the ECE is calculated.

Dataset: The project database has over 32 thousand records with species information such as scientific name, morphological description, common names, species distribution, life cycle, feeding, reproduction, demography, habitat, phenology, behavior, interactions, conservation status, and legislation, among other types of contents. The biodiversity related texts are currently in the preparation process. Initially, we used a database that simulates the complexity of the biodiversity conservation project, and then we will apply the findings to the biodiversity texts. The experiments use the "Explain Like I'm 5" dataset's science segment (ELI5Sci). This dataset, developed by Facebook AI Research, serves as a benchmark for evaluating long-form question answering. It encompasses data across a wide range of subjects, including science, history, and general topics. The science segment has 131,778 records for training, 2,281 for validation, and 4,462 for testing (Fan et al., 2019). Creating accurate and coherent ELI5-style answers with ELI5Sci can be challenging, as it requires a deep understanding of the underlying concepts and the ability to deliver complex information in a simple way.

Models evaluated: Falcon-7B and DistilGPT-2. **Falcon-7B** is a causal decoder-only model built by the Technology Innovation Institute (TII). The model was trained on 1,500B tokens of RefinedWeb (Penedo et al., 2023) enhanced with curated corpora (Almazrouei et al., 2023). Falcon-7B was fine-tuned using Quantization of Low Rank Adapters - QLoRA (Dettemers et al., 2023), the bitsandbytes library, and Parameter-Efficient Fine-Tuning (PEFT) from Hugging Face (Wolf et al., 2020) with the ELI5 training dataset (during 300 global steps, training loss=2.27, validation loss=2.35) before performing the experiments and the results were saved locally. **DistilGPT-2** (short



a. Falcon-7B: Reliability diagram **for Perplexity**



b. DistilGPT-2: Reliability diagram **for Perplexity**

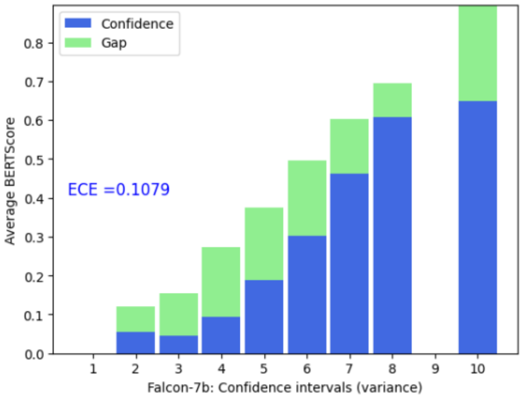
Figure 4: Uncertainty representation method based on calibration confidence for Perplexity. a) Falcon-7B reliability diagram that plots the observed probability against the predicted probability (Perplexity) for each bin, a perfectly calibrated model should have a diagonal line. b) DistilGPT-2 reliability diagram.

for Distilled-GPT2) is a compact version of the original Generative Pre-trained Transformer 2 (GPT-2) model, developed by Hugging Face. It shares the same transformer architecture and is pre-trained with the supervision of the smallest version of GPT-2 (Hugging Face, 2019). DistilGPT-2 was fine-tuned with ELI5Sci training dataset (during 6 epochs, training loss=3.42, validation loss=3.54). To perform the planned exploratory analysis, both models offer a balance between performance and resource requirements.

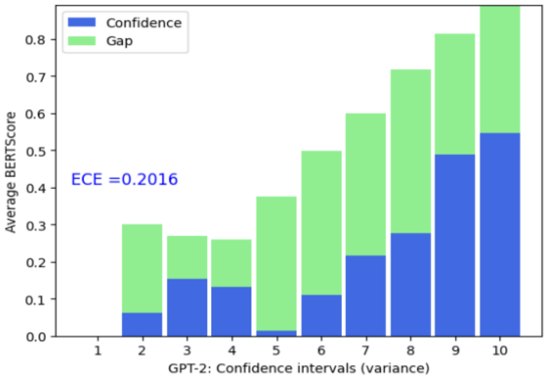
Hardware: Lenovo Legion 5i, Intel Core i7-12700, GeForce RTX 3060 (12 GB GDDR6), 64GB DDR5 RAM.

3.2 Experiment

We use the pre-trained checkpoints of DistilGPT-2, Falcon-7B, and metrics from Hugging Face. Then we fine-tune the models using training data from



a. Falcon-7B: Reliability diagram **for MCD**



b. DistilGPT-2: Reliability diagram **for MCD**

Figure 5: Reliability diagrams for MCD for each model that plots confidence using MCD against predicted probability (without MCD).

the target dataset along with the validation dataset (131,778 records for training, 2,281 validation records).

The **goal of the experiment** is to estimate the model's uncertainty by applying a sampling-based method using MCD to generate multiple different predictions and ECE to assess the uncertainty of DistilGPT-2 and Falcon-7B fine-tuned with ELI5Sci. Furthermore, we evaluate whether it is feasible to use Perplexity as a confidence measure of the models (a brief example of results is found in appendix A).

The stages in the project's workflow, as shown in Figure 3, are the following:

1. Define the validation dataset O_N - The process begins with the random selection of 500 records from the ELI5Sci test segment (for reasons of computational power, tests are not initially performed on the complete data).
2. Compute the BERTScore of the model responses - These chosen records are then used to generate text using Contrastive Search

Variable	DistilGPT-2	Falcon-7B
Dropout rate d (MCD)	0.05	0.04
ECE (PPL)	0.2539	**0.0898**
ECE (MCD)	0.2016	**0.1079**

Table 1: Parameters and comparative results by model evaluated.

(Su and Collier, 2022) with both models (without MCD) and compute BERTScore to evaluate the generated text.

3. Compute Perplexity as U_Q score - Models' confidence is computed using Perplexity. We complement our evaluation with the analysis of Perplexity to measure the coherence and predictability of the generated text. A high Perplexity indicates that the text is very unlikely and has coherence issues given the vocabulary distribution (Su et al., 2022; Holtzman et al., 2020). Therefore, perplexity can be thought as an uncertainty quantifier. To incorporate this metric into the project, the obtained results were rescaled to fall within the range of [0,1]. The re-scaling was made using the maximum value obtained for all the selected samples of the calibration data.
4. Compute MCD as U_Q score - Subsequently, each model generates 10 samples for each selected record using MCD with a dropout rate d to compute the model's accuracy (average of BERTScore) and uncertainty (variance of Perplexity and BERTScore). The dropout rates for both models (Table 1) were fine-tuned before the experiment through a rigorous evaluation of the generated text using BERTScore across various dropout rates.
5. Compute the ECE as the correlation between the BERTScore and the U_Q - The final stage involves assessing the model's calibration by computing the ECE, which helps determine how closely the model's predicted probabilities align with actual outcomes. In the context of this project, we implement ECE by partitioning the data into B equally spaced bins. Here, we specify $B = 10$ B, a commonly adopted value for binning in such analyses. This comprehensive workflow serves to analyze and fine-tune the model's performance. Figure 4 and Figure 5 show the process results.

4 Results

Ours findings indicate the feasibility of using ECE as a metric for estimating uncertainty in generative LLM. The outcomes obtained through perplexity and MCD highlight the disparity in text quality between DistilGPT-2 and Falcon-7B (Figure 4 and Figure 5). Across all experiments, DistilGPT-2 demonstrates a higher calibration error compared to Falcon-7B, with an average ECE using Perplexity of 0.2539, as opposed to Falcon-7B, which registers an error of 0.0898 (Table 1). The ECE values for MCD using BERTScore maintain a notable contrast, averaging at 0.2016 for DistilGPT-2 and 0.1079 for Falcon-7B. It is important to acknowledge that the tokenization procedure directly influences the Perplexity of a model. This factor should always be considered when comparing various models that is why we include BERTScore in the analysis. These results must be corroborated with experiments involving humans. Both Perplexity and MCD tests were conducted three times, and the results were averaged.

The calibration of the Biodiversity Project models not only involves parameters such as dropout rate but also parameters related to text generation. The quality of the texts generated by LLM rely on randomness in the decoding method, in particular through penalty alpha, top-k, and top-p variables that guide the selection of next word. The next word generated by a model is selected from the top k most probable choices in the model's vocabulary given a set of previously generated words or from the smallest set of tokens for which the cumulative probability exceeds a specified value, p. (Radford et al., 2019).

The Biodiversity Project already has a portal that provides free and open access to biodiversity data. The fine-tune LLM will complement the services available on the portal to support the development of science, education and biodiversity conservation.

For conducting tests with data from the Biodiversity Project, we have a cloud-based server (4x GPU NVIDIA® T4 and 192 GB RAM or some similar configuration) at our disposal, and we plan to use an open multilingual LLM with state-of-the-art performance (e.g. Falcon-40B) for the process.

5 Conclusion

Ensuring model calibration is vital as it guarantees not only accurate but also reliable probability estimates. Our exploratory study gives preliminary data on uncertainty estimation in generative language models using MCD, Perplexity methods for uncertainty quantification, and BERTScore. According to the results yielded for the two evaluated models, the performance of both uncertainty scoring methodologies can be thought to be similar. However, more testing is necessary, with more replicas and a calibration process using a holdout dataset. Identified areas requiring future scrutiny and enhancement include: a) Model Selection - The choice of Falcon-7B and DistilGPT-2 was constrained by computational resources. However, there is a necessity for a superior-performing, multilingual, open-source language model like Falcon-40B, particularly for testing texts written in Spanish. b) The results obtained are encouraging; however, it is necessary to involve humans in the following experiments of the project. c) Evaluating model confidence using metrics beyond Perplexity, as suggested by Meister and Cotterell (2021), who propose a framework of metrics based on language laws. d) Prompt Engineering Evaluation - Assess diverse prompt strategies designed to improve text generation.

Limitations

Our analysis and conclusions have been based only on a single language (i.e., English), a single dataset (ELI5Sci), and two transformer based models (i.e., Distilgpt2 and Falcon-7B). The generalization to other languages, data and models is yet to be verified.

Ethical statement

The proposed metric for estimating uncertainty is currently undergoing research, and additional testing is necessary to determine its viability. Its premature use may result in misleading or inaccurate information, potentially leading to adverse effects when incorporated into decision-making systems.

Acknowledgments

This work was made possible through the support provided by the Postgraduate Office of the ITCR and the Costa Rican Innovation and Research Promoter of the Ministry of Science, Innovation, Technology, and Telecommunications (MICITT) of Costa Rica.

References

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. In *Proceedings 8th International Conference on Learning Representations, ICLR 2020.* https://doi.org/10.48550/arXiv.1904.09751

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. *Language models are unsupervised multitask learners.* OpenAI blog, 1(8):9

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, Michael Auli. 2019. ELI5: Long form question answering. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* https://doi.org/10.18653/v1/P19-1346.

Camilo Mora, Derek P. Tittensor, Sina Adl, Alastair G. B. Simpson and Boris Worm. 2011. *How Many Species Are There on Earth and in the Ocean?.* PLoS Biology 9 (8). https://doi.org/10.1371/journal.pbio.1001127

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. https://aclanthology.org/W04-1013/

Clara Meister and Ryan Cotterell. 2021. Language Model Evaluation Beyond Perplexity. *In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers), pages 5328–5339, Online. Association for Computational Linguistics. DOI 10.18653/v1/2021.acl-long.414

Dan Hendrycks, Nicholas Carlini, John Schulman, Jacob Steinhardt.2022. *Unsolved Problems in ML Safety.* Computing Research Repository. https://doi.org/10.48550/arXiv.2109.13916

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, Dan Mané. 2016. Concrete Problems in AI Safety. Computing Research Repository ArXiv. https://doi.org/10.48550/arXiv.1606.06565.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, Guilherme Penedo. 2023. *The Falcon Series of Open Language Models.* Computing Research Repository https://doi.org/10.48550/arXiv.2311.16867

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei and Julien Launay. 2023. *The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only.* Computing Research Repository https://doi.org/10.48550/arXiv.2306.01116

Hong Cui, Bruce Ford, Julian Starr, James Macklin, Anton Reznicek, Noah W. Giebink, Dylan Longert, Étienne Léveillé-Bourret, Limin Zhang. 2021. *Author-Driven Computable Data and Ontology Production for Taxonomists.* Biodiversity Information Science and Standards 5. https://doi.org/10.3897/biss.5.75741

Hideyuki Doi, Takeshi Osawa, Narumasa Tsutsumida. 2023. *The role of large language models in ecology and biodiversity conservation: Opportunities and Challenges.* https://europepmc.org/article/ppr/ppr674919

HuggingFace. 2019. Distilgpt2. URL https://huggingface.co/distilgpt2.

Jeremy Nixon, Mike Dusenberry, Ghassen Jerfel, Timothy Nguyen, Jeremiah Liu, Linchuan Zhang, Dustin Tran. 2019. *Measuring Calibration in Deep Learning.* CVPR workshops. Vol. 2. No. 7. Computing Research Repository https://doi.org/10.48550/arXiv.1904.01685

Jianfeng He, Xuchao Zhang, Shuo Lei, Zhiqian Chen, Fanglan Chen, Abdulaziz Alhamadani, Bei Xiao, and ChangTien Lu. 2020. Towards More Accurate Uncertainty Estimation In Text Classification. *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 8362–8372, Online. Association for Computational Linguistics.*

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura and Kaidi Xu. 2023. *Shifting Attention to Relevance: Towards the Uncertainty Estimation of Large Language Models.* Computing Research Repository https://arxiv.org/abs/2307.01379

José Mena , Oriol Pujol and Jordi Vitrià. 2021. *A survey on uncertainty estimation in deep learning classification systems from a Bayesian perspective*. ACM Computing Surveys (CSUR) 54.9 (2021): 1-35.

Daeyoung Kim, Seongsu Bae, Seungho Kim, Edward Choi. 2022. *Uncertainty-Aware Text-to-Program for Question Answering on Structured Electronic Health Records.* https://arxiv.org/abs/2203.06918. Computing Research Repository https://doi.org/10.48550/arXiv.2203.06918

Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, Maarten Sap. 2024. *Relying on the Unreliable: The Impact of Language Models' Reluctance to Express Uncertainty*. Computing Research Repository https://doi.org/10.48550/arXiv.2401.06730

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. *BLEU: a Method for Automatic Evaluation of Machine Translation*. https://doi.org/10.3115/1073083.1073135

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. *Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation*. Computing Research Repository https://arxiv.org/abs/2302.09664.

Mahdi Pakdaman Naeini, Gregory Cooper and Milos Hauskrecht. 2015. *Obtaining well calibrated probabilities using bayesian binning.* Proceedings of the AAAI conference on artificial intelligence. Vol. 29. No. 1. 2015. https://doi.org/10.1609/aaai.v29i1.9602

Maria Mora and José Enrique Araya. 2018. Semi-automatic Extraction of Plants Morphological Characters from Taxonomic Descriptions Written in Spanish. *Biodiversity Data Journal 6*

Maria Mora, William Ulate, Brandon Retana Chacón, María Biarreta Portillo, Josué David Castro Ramírez, Jose Chavarria Madriz. 2023. Structuring Information from Plant Morphological Descriptions using Open Information Extraction. *Biodiversity Information Science and Standards* 7: e113055.

Nancy E. Gwinn and Constance Rinaldo. 2009. The Biodiversity Heritage Library: sharing biodiversity literature with the world. *IFLA Journal. Official Journal of the International.* https://doi.org/10.1177/0340035208102032

Peter Van Katwyk, Baylor Fox-Kemper, Hélène Seroussi, Sophie Nowicki and Karianne J. Bergen. 2023. *A Variational LSTM Emulator of Sea Level Contribution From the Antarctic Ice Sheet*. https://doi.org/10.1029/2023MS003899

Satanjeev Banerjee and Alon Lavie. 2005. *METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments.* Computing Research Repository https://aclanthology.org/W05-0909

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. *Teaching models to express their uncertainty in words.* Computing Research Repository arXiv preprint arXiv:2205.14334.

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. Transformers: State-of-the-art natural language processing. *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.*

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger and Yoav Artzi. 2020. BERTScore: *Evaluating Text Generation with BERT*. Computing Research Repository https://doi.org/10.48550/arXiv.1904.09675

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman and Luke Zettlemoyer. 2023. *QLoRA: Efficient Finetuning of Quantized LLMs.* Computing Research Repository https://doi.org/10.48550/arXiv.2305.14314

United Nations. 1992. *Convention on Biological Diversity*. New York. [83] p.

Secretariat of the Convention on Biological Diversity. 2020. *Global Biodiversity Outlook 5*. Montreal.

Yarin Gal. 2016. *Uncertainty in deep learning.* PhD thesis. https://www.cs.ox.ac.uk/people/yarin.gal/website/thesis/thesis.pdf

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang and Yong Yu. 2018. *Texygen: A Benchmarking Platform for Text Generation Models*. Computing Research Repository https://doi.org/10.48550/arXiv.1802.01886

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. Computing Research Repository https://arxiv.org/abs/1506.02142

Yassir Fathullah, Mark J. F. Gales. 2022. Self-distribution distillation: efficient uncertainty estimation. *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, PMLR 180:663-673.

Yijun Xiao and William Yang Wang. 2018. *Quantifying Uncertainties in Natural Language Processing Tasks*. Computing Research Repository https://arxiv.org/pdf/1811.07253.pdf

Yixuan Su, Nigel Collier. 2023. *Contrastive Search Is What You Need For Neural Text Generation*. Computing Research Repository https://doi.org/10.48550/arXiv.2210.14140

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong and Nigel Collier. 2022. A *Contrastive Framework for Neural Text Generation.* Computing Research Repository https://doi.org/10.48550/arXiv.2202.06417

Zhengbao Jiang, Jun Araki, Haibo Ding, Graham Neubig. 2021. *How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering.* Computing Research Repository https://doi.org/10.48550/arXiv.2012.00955

A. Appendix

Uncertainty Estimation Example

The main objective of this example is to evaluate whether LLM can handle queries associated with scientific names, so we selected a question from the ELI5Sci test dataset to assess whether Falcon-7B fine-tuned has the ability to respond to queries associated with scientific names.

Q1) Question: Is there such a thing as endangered plants???
The best human answer (score 20): Endangered and extinct plants exist, just like animals. An example: the Arizona Agave with less than 100 plants alive in 1984. There are also plenty of extinct plants. The reason you never hear about them is probably because people care less about plants. They are also often closely related to other species which aren't extinct.

Falcon-7B answers

Responses were generated using Falcon-7B fine-tuned with ELI5Sci. Perplexity and BertScore metrics were calculated for the responses. Texts were generated using the following parameters: penalty_alpha = 0.1, topk=5, and answer_length = 512. To enhance the diversity of responses, a dropout rate of 0.04 was applied in some cases.

Answer	**Metadata**
1) Yes, there are endangered plants. In fact, there are many. Here's an article that lists some of the endangered plants in the U.S. The U.S. has a program called the Endangered Species Act (ESA) which protects endangered plants as well as animals. I am sure that there are plants that are endangered in other parts of the world as well, but I am not sure. I know that the U.S. has some endangered plants. There are plants that are endangered because they are in danger of extinction, and there are plants that have been endangered because their habitats have been destroyed by humans. I think the same rules would apply as for animals...	Falcon-7B **without dropout** **Metrics** Perplexity normalized: 0.0679 BertScore: 0.8332
2) I've read a few articles on endangered plants, but I'm not sure if any have been declared extinct yet or not. The closest I've heard is the California Redwoods. I believe the last one was chopped down about 10 years ago, and scientists are now working to try and reestablish them…	Falcon-7B **with dropout of 0.04** Perplexity normalized: 0.0372 BertScore: 0.7851
3) Yes, there are endangered plants and many are on this planet. The biggest problem is the fact that we are destroying their habitat,which is the biggest reason they are endangered. Also, some species of plants are endangered because of the fact that we use their parts in medicine and we are taking them away from their habitat. There are many endangered plants. There is a plant in Australia, known as the Wollemi Pine, that has been extinct for over 80 million years, yet it has been rediscovered in the wild in Australia. This plant is endangered. There are many more, however…	Falcon-7B **with dropout of 0.04** Perplexity normalized: 0.0980 BertScore: 0.8263

Questions associated with scientific names

These questions were not part of the test dataset, so the BertScore cannot be measured.

Question	Answer	Metadata
Q2)Taking into account your scientific knowledge, is the species *Sequoia sempervirens* in danger in the wild?	Yes, it is. The giant sequoia is endangered because of the loss of habitat, and the destruction of its forests. The giant sequoia is also endangered because it is being killed by bark beetles and other pests…	Falcon-7B **without dropout** Perplexity normalized: 0.0414
Q3) Taking into account your scientific knowledge, is the species *Wollemia nobilis* in danger in the wild?	Wollemia Noblis is an ancient species of fern which only grows in one place in the world (Australia) and it is very rare. There is no danger in the wild, it is protected. It's a very interesting and important species for palaeontologist because we know very little about it, but it could be a clue for understanding the evolution of fern. It is very important for palaeontologist, but it is very rare so there's no danger for the species. Wollemia nobilis, is a species of fern. It is the oldest known fern, and is found in one place in the world - Australia. The species is endangered because of its habitat being destroyed.	Falcon-7B **without dropout** Perplexity normalized: 0.0504
Q4)Taking into account your scientific knowledge, is the species *Amazilia boucardi* in danger in wild?	Amazilia boucardi is a species that is not listed on the IUCN Red List. It is classified as Least Concern. The Amazilia boucardi population is considered stable, with a high reproductive potential. It is not affected by poaching. Amazilia boucardi is not in danger in the wild.	Falcon-7B **without dropout** **Perplexity normalized: 0.1257**

The initial three model answers demonstrate good quality, with an acceptable BertScore alongside low perplexity. Additionally, in responses 2) and 3), we can see that the model uses two species as examples of endangered plants (California Redwoods and Wollemi Pine). However, the model does not use the scientific names of the species; instead, it uses their common names.

We conducted the experiment to evaluate the model's ability to answer questions regarding the conservation status of various species using their scientific names. Initially, we queried the model about species it had previously referenced by their common names in responses 2) (California Redwoods) and 3) (Wollemi Pine). Subsequently, we presented the model with a question about an exceedingly rare species endemic to Costa Rica (i.e. only found within the borders of Costa Rica), known to be endangered. The experiment results were interesting: a) regarding the species for which the model had knowledge of the conservation status associated with the common name, the model was able to use that data correctly. b) Regarding the species *Amazilia boucardi*, the model likely did not have much information, resulting in an incorrect response and a higher perplexity metric compared to the other two species (i.e., *Sequoia sempervirens* and *Wollemia nobilis*).