Reranking Laws for Language Generation: A Communication-Theoretic Perspective

António Farinhas[1,2] **Haau-Sing Li**[2,3] **André F. T. Martins**[1,2,4,5]
[1]Instituto Superior Técnico, Universidade de Lisboa [2]Instituto de Telecomunicações
[3]Ubiquitous Knowledge Processing Lab, TU Darmstadt [4]ELLIS Unit Lisbon [5]Unbabel
{antonio.farinhas,andre.t.martins}@tecnico.ulisboa.pt, hli@ukp.tu-darmstadt.de

Abstract

To ensure large language models (LLMs) are used safely, one must reduce their propensity to hallucinate or to generate unacceptable answers. A simple and often used strategy is to first let the LLM generate multiple hypotheses and then employ a reranker to choose the best one. In this paper, we draw a parallel between this strategy and the use of redundancy to decrease the error rate in noisy communication channels. We conceptualize the generator as a sender transmitting multiple descriptions of a message through parallel noisy channels. The receiver decodes the message by ranking the (potentially corrupted) descriptions and selecting the one found to be most reliable. We provide conditions under which this protocol is asymptotically error-free (*i.e.*, yields an acceptable answer almost surely) even in scenarios where the reranker is imperfect (governed by Mallows or Zipf-Mandelbrot models) and the channel distributions are statistically dependent. We use our framework to obtain reranking laws which we validate empirically on two real-world tasks using LLMs: text-to-code generation with DeepSeek-Coder 7B and machine translation of medical data with TowerInstruct 13B.

1 Introduction

Large language models (LLMs) have shown remarkable performance across many tasks in natural language processing, computer vision, and speech recognition. Despite their capabilities, instances of hallucinations and other critical errors occasionally arise, casting doubt on the reliability of their predictions, without clear indication of when and how badly they might fail (Ji et al., 2023; Guerreiro et al., 2023). This is particularly concerning as these models are increasingly used in high-stakes applications such as those within the medical or legal domains (Hung et al., 2023) or as agents that can perform multiple tasks, including generating and executing code (Wang et al., 2024).

The most common mitigation strategy is to "steer" the LLM with the aid of a reward model or directly from human preferences, either at training time (Stiennon et al., 2020; Yuan et al., 2024; Rafailov et al., 2024) or during decoding (Liu et al., 2024; Huang et al., 2024). A simple and effective decoding-time strategy is first to generate multiple hypotheses and then use a reranker to select the most appropriate one. Several generation techniques used with modern LLMs, including voting procedures (Borgeaud and Emerson, 2020; Wang et al., 2023; Liévin et al., 2024; Shi et al., 2022), minimum Bayes risk decoders (Eikema and Aziz, 2020; Freitag et al., 2022), quality-aware decoders (Fernandes et al., 2022), or other types of hypothesis ensembling/reranking techniques (Farinhas et al., 2023; Ni et al., 2023; Bertsch et al., 2023; Li et al., 2024), embody this idea. An essential aspect of these procedures is that they all add **redundancy** as an intermediate step (by generating multiple hypotheses) to increase the chances of returning an acceptable answer as the final output.

The idea of adding redundancy to decrease the error rate in noisy channels is a cornerstone of **communication theory**, more specifically in forward error correction methods. In its simplest

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

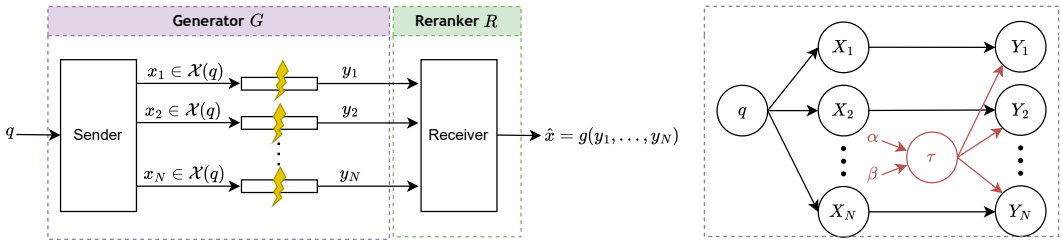

Figure 1: **Left:** A generator-reranker system (G, R) depicted as a communication system (§2). Given a query q with acceptance set $\mathcal{X}(q)$, the sender sends N descriptions through noisy channels. The receiver's goal is to decode an acceptable answer through reranking. **Right:** Graphical model of the generator G. We consider two different models: a simplified version with N independent hypotheses, represented in black (§3), and a scenario with exchangeable hypotheses, represented in red (§4).

form—repetition codes—a message block is sent multiple times, and the decoder uses some form of majority voting to recover the original message with high probability (MacKay, 2002; Cover and Thomas, 2006). The same idea underlies more sophisticated error-correcting codes (Hamming, 1950; Reed and Solomon, 1960; Gallager, 1962; Berrou et al., 1993).

In this paper, we draw a parallel between these two worlds by regarding generator-reranker LLMs as communication systems (§2 and Fig. 1, left). We conceptualize the LLM generator G as a sender transmitting N message descriptions in parallel through noisy channels, leading to N potentially corrupted hypotheses. Then, the receiver, which corresponds to the reranker R, decodes the message by ranking the potentially corrupted descriptions and selecting the one found to be most reliable. The goal is for the combined (G, R) system to have lower error rate than G alone, and for the error rate to decay quickly with N. Our main contributions are as follows:

- We show that when the channel distributions are independent, this simple protocol is asymptotically error-free (*i.e.*, it generates an acceptable answer almost surely when $N \to \infty$), even in scenarios where the reranker is imperfect, *e.g.*, governed by a Mallows or a Zipf-Mandelbrot model. In the former case, the error probability decays exponentially fast (§3).
- We show that the protocol is still asymptotically error-free if we assume that the channel distributions are statistically dependent. When they are coupled by a Beta prior, we show that the error probability decays as a power law when the reranker is perfect (§4).
- We use our framework to obtain "reranking laws", which we validate empirically on text-to-code generation with DeepSeek-Coder 7B (§5.1), on machine translation of medical data with TowerInstruct 13B (§5.2), and on mathematical and commonsense reasoning benchmarks (App. B.3).

Notation. We denote $[N] := \{1, ..., N\}$ and we use the shorthand notation $X_{1:N} := (X_1, ..., X_N)$. We use capital letters $(X, Y, ...)$ for random variables and represent probability distributions by $\mathbb{P}(X), \mathbb{P}(Y)$, etc. We denote expectations of functions f under $\mathbb{P}(X)$ by $\mathbb{E}_X[f(X)]$.

2 A Communication-Theoretic Perspective of Generator-Reranker Systems

The focus of our paper is on **generator-reranker systems**: a **generator** G (such as an LLM) is prompted with a **query** q (*e.g.*, a question to be answered, a source text to be translated, or a textual prompt for code). As a response to this query, G generates N candidate answers $y_1, ..., y_N$ (called **hypotheses**). We are agnostic about the internals of G and the way the hypotheses are generated: they could come from the same system through sampling or beam search, or they could come from an ensemble of different systems. These hypotheses are then processed by a **reranker** R, which ranks them and returns as the final output the one which is found to be the best answer. We are also agnostic about how R is built—it could be an external system or it could be part of (or share parameters with) the generator. Commonly used rerankers are quality estimators (Fernandes et al., 2022), energy-based models (Bhattacharyya et al., 2021), reward models (Li et al., 2022), and minimum Bayes risk decoders (Kumar and Byrne, 2002; Eikema and Aziz, 2020; Freitag et al., 2022; Shi et al., 2022).

Regardless of specific design decisions, the goal of the generator-reranking system (G, R) is to leverage the reranker R to produce better answers (according to some quality metric) than the ones which would be obtained through G alone (*e.g.*, a single sample). In this paper, we show that the propensity for this combined system to generate unacceptable outputs, such as those containing critical errors or hallucinations, decays quickly enough with N under mild assumptions on G and R.

We draw an analogy with communication theory as follows. Let Σ be an underlying alphabet and $\Sigma^* := \bigcup_{i=0}^{\infty} \Sigma^i$ its Kleene closure, *i.e.*, the set of strings from Σ. Given the query q, we denote by $\mathcal{X}(q) \subseteq \Sigma^*$ the set of **acceptable answers**.[1] We assume the communication system depicted in Fig. 1 (left), a form of **multiple description source coding** (Ozarow, 1980; Gamal and Cover, 1982; Laneman et al., 2005). In this framework, the sender transmits N acceptable answers (called **descriptions**) $x_1, ..., x_N \in \mathcal{X}(q)^N$ in parallel through noisy channels. These descriptions are corrupted according to a distribution $\mathbb{P}(y_1, ..., y_N | x_1, ..., x_N)$, so that some hypotheses y_i may become unacceptable ($y_i \in \Sigma^* \setminus \mathcal{X}(q)$). This "channel noise" is a way to conceptualize the imperfections of the generator G. On the receiver side, a decoder processes the (potentially) corrupted descriptions and estimates $\hat{x} = g(y_1, ..., y_N)$ using some decoding function g. The overarching goal is to achieve a low error probability $P_{\text{err}}(N; q) := \mathbb{P}(\hat{X} \notin \mathcal{X}(q) \mid q)$ for any query q. By bounding the maximal probability of error (over all queries), the average error probability is automatically bounded (Cover and Thomas, 2006, §8). In this paper, we focus on rerankers as the decoding functions, where $g(y_1, ..., y_N)$ returns the top ranked answer, *i.e.*, $g(y_1, ..., y_N) = y_i$ for some $i \in [N]$.

We formalize this construction by considering different models for G and R in the following sections, studying the conditions under which the resulting protocol is **asymptotically error-free**:

Definition 1. *A protocol is asymptotically error-free if, for any query q, the probability of the decoder outputting an unacceptable answer approaches zero as N tends to infinity, i.e.,*

$$\lim_{N \to \infty} \underbrace{\mathbb{P}(g(Y_1, ..., Y_N) \notin \mathcal{X}(q) \mid q)}_{:= P_{\text{err}}(N; q)} = 0. \tag{1}$$

For simplicity, we assume that $X_1, ..., X_N$ are conditionally independent given the query q, *i.e.*, that $\mathbb{P}(x_1, ..., x_N | q) = \prod_{i=1}^{N} \mathbb{P}(x_i | q)$.[2] We also assume that $Y_{1:N}$ are independent from q given $X_{1:N}$ such that $q \to X_{1:N} \to Y_{1:N}$ forms a Markov chain. Taken together, these two assumptions mean that $\mathbb{P}(x_{1:N}, y_{1:N} | q) = \mathbb{P}(x_{1:N} | q) \mathbb{P}(y_{1:N} | x_{1:N}) = \left(\prod_{i=1}^{N} \mathbb{P}(x_i | q) \right) \mathbb{P}(y_{1:N} | x_{1:N})$.

3 Generator-Reranker Systems with Independent Hypotheses

We first consider the case where the corrupted descriptions $Y_{1:N}$ are conditionally independent and identically distributed (i.i.d.) given $X_{1:N}$ and where Y_i depends only on X_i, that is, $\mathbb{P}(y_{1:N} | x_{1:N}) = \prod_{i=1}^{N} \mathbb{P}(y_i | x_i)$. Conceptually, this is the scenario where the parallel channels do not interfere, and it corresponds to the graphical model shown in Fig. 1 (right) without the part in red. While this case may not be very realistic in practice—for example, when the hypotheses produced by the generator are all sampled from the same model—it makes the analysis simpler. We will show later in §4 how the analysis can be extended when this assumption does not hold, reusing the results from this section.

In the sequel, given a query q, we let ϵ denote the probability of a hypothesis being unacceptable, $\epsilon := \mathbb{P}(Y_i \notin \mathcal{X}(q) \mid X_i = x_i, q) = \mathbb{P}(Y_i \notin \mathcal{X}(q) \mid X_i = x_i)$.

3.1 Perfect and random rerankers

We start by assuming that R is a **perfect reranker**, which implies that it produces an acceptable output when presented with a set of N hypotheses if and only if at least one of them is acceptable. In

[1] A key difference between our framework and most lossless communication systems is that there is no need to communicate a *specific* message—any answer in the equivalence class $\mathcal{X}(q)$ is acceptable, hence, if the decoder recovers any message in this set, the communication is considered successful. This is a natural conceptualization in problems involving natural language (where a paraphrase of a correct answer is still correct) or code (where multiple programs might lead to the same execution).

[2] In fact, all results in this paper still hold if there are dependencies between $X_1, ..., X_N$.

this case, the error probability becomes

$$P_{\text{err}}(N;q) = \mathbb{P}(g(Y_1,...,Y_N) \notin \mathcal{X}(q) \mid q) = \mathbb{E}_{X_{1:N}|q}\big[\mathbb{P}(g(Y_1,...,Y_N) \notin \mathcal{X}(q) \mid X_{1:N}, q)\big]$$
$$= \mathbb{E}_{X_{1:N}|q}\big[\mathbb{P}(Y_i \notin \mathcal{X}(q),\ \forall i \in [N] \mid X_{1:N})\big]$$
$$= \mathbb{E}_{X_{1:N}|q}\left[\prod_{i=1}^{N} \underbrace{P(Y_i \notin \mathcal{X}(q) \mid X_i)}_{=\epsilon}\right] = \epsilon^N. \quad (2)$$

Thus, $P_{\text{err}}(N;q)$ goes to zero exponentially fast with N for any $\epsilon \in [0,1)$, indicating that when the hypotheses are independent and the reranker is perfect, the protocol is error-free.

On the other end of the spectrum, if the reranker is **random**—*i.e.*, if it selects one of the N hypotheses uniformly at random, we obtain

$$P_{\text{err}}(N;q) = \mathbb{P}(g(Y_1,...,Y_N) \notin \mathcal{X}(q) \mid q) = \mathbb{E}_{X_{1:N}|q}\big[\mathbb{P}(g(Y_1,...,Y_N) \notin \mathcal{X}(q) \mid X_{1:N}, q)\big]$$
$$= \mathbb{E}_{X_{1:N}|q}\Big[\mathbb{E}_i\big[\mathbb{P}(Y_i \notin \mathcal{X}(q) \mid X_{1:N}, i)\big]\Big] = \epsilon, \quad (3)$$

that is, we obtain the same error probability as the generator alone, as expected.

3.2 Imperfect reranker: Mallows model

We consider now more realistic rerankers. A statistical ranking model widely used in machine learning applications is the **Mallows model** (Klementiev et al., 2008, 2009; Chierichetti et al., 2018; Tang, 2019). Let Π denote the set of permutations over N elements, and let $d : \Pi \times \Pi \to \mathbb{R}_+$ be a distance function between permutations. In this paper, we use the Kendall-tau distance $d(\pi, \pi')$, which returns the number of adjacent transpositions needed to turn π into π'. Given a location parameter $\pi_0 \in \Pi$ and a scale parameter $\lambda \in \mathbb{R}_+$, the probability of a ranking π according to the Mallows model is $\mathbb{P}(\pi; \pi_0, \lambda) = \exp(-\lambda d(\pi, \pi_0))/Z(\lambda)$, where $Z(\lambda)$ is the partition function.

In our setting, we assume that π_0 is the ground truth (oracle) ranking[3] of the hypotheses $y_1, ..., y_N$ and π is the ranking obtained by the reranker model, so that $\mathbb{P}(\pi; \pi_0, \lambda)$ expresses how imperfect the reranker might be. Note that the family of Mallows models include both perfect and random rerankers as limit cases, respectively as $\lambda \to +\infty$ and as $\lambda = 0$.[4]

Let η_j denote the marginal probability that the reranker places at the top the j^{th} highest ranked hypothesis according to the oracle, *i.e.*, $\eta_j = \mathbb{P}(\pi_0(\pi^{-1}(1)) = j)$. When K out of the N hypotheses are unacceptable, the reranker will pick an unacceptable hypothesis with probability $\sum_{j=N-K+1}^{N} \eta_j$. Combining this with the fact that the probability of G generating K unacceptable hypotheses is a binomial distribution, the error probability becomes

$$P_{\text{err}}(N;q) = \mathbb{P}(g(Y_1,...,Y_N) \notin \mathcal{X}(q) \mid q) = \mathbb{E}_{X_{1:N}|q}\big[\mathbb{P}(g(Y_1,...,Y_N) \notin \mathcal{X}(q) \mid X_{1:N}, q)\big]$$
$$= \sum_{K=0}^{N} \left[\binom{N}{K} \epsilon^K (1-\epsilon)^{N-K} \sum_{j=N-K+1}^{N} \eta_j\right]. \quad (4)$$

Note that (4) holds for **any reranker** with top-1 (marginal) probability mass function $\boldsymbol{\eta} = [\eta_1, ..., \eta_N]$, not only Mallows models. Naively determining $\boldsymbol{\eta}$ would require marginalizing $\mathbb{P}(\pi; \pi_0, \lambda)$ by summing over all permutations π satisfying $\pi_0(\pi^{-1}(1)) = j$, which is intractable due to the factorial number of terms involved. Fortunately, tractable combinatorial expressions exist for Mallows models (Fligner and Verducci, 1986; Lebanon and Mao, 2008): the partition function has the compact expression $Z(\lambda) = \prod_{j=1}^{N}(1 - e^{-\lambda j})/(1 - e^{-\lambda})$, and we have (Lebanon and Mao, 2008, Prop. 5):

$$\eta_j = Z^{-1}(\lambda) \sum_{\pi: j=\pi_0(\pi^{-1}(1))} e^{-\lambda d(\pi,\pi_0)} = \frac{e^{-\lambda(j-1)}}{\sum_{r=1}^{N} e^{-\lambda(r-1)}}. \quad (5)$$

[3] More specifically, we assume that the hypotheses are ranked according to some quality metric compatible with $\mathcal{X}(q)$, that is, unacceptable answers should be ranked after acceptable answers.

[4] Notably, $e^{-\lambda}$ strictly between 0 and 1 correspond to imperfect rerankers that are better than random. Lower values indicate higher-quality rerankers, making $e^{-\lambda}$ an inverse measure of reranker quality.

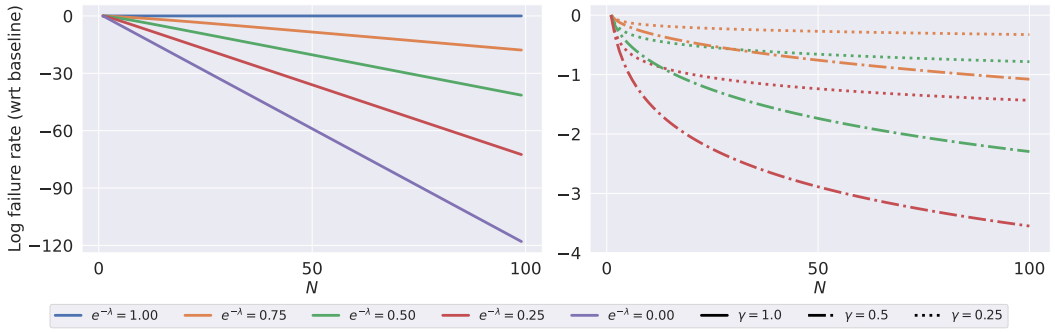

Figure 2: Log of the failure rate (difference with respect to the baseline rate $\log \epsilon$) as a function of the number of generated independent hypotheses N for several values of $e^{-\lambda}$ and $\epsilon = 0.3$. **Left:** Mallows model (§3.2). **Right:** Zipf-Mandelbrot model (§3.3).

Plugging (5) into (4), invoking the binomial theorem, and simplifying, we obtain

$$P_{\text{err}}(N; q) = \mathbb{P}(g(Y_1, ..., Y_N) \notin \mathcal{X}(q) \mid q) = \begin{cases} \epsilon & \text{if } \lambda = 0 \\ \frac{[e^{-\lambda}(1-\epsilon)+\epsilon]^N - e^{-\lambda N}}{1 - e^{-\lambda N}} & \text{otherwise.} \end{cases} \tag{6}$$

Notably, when $\lambda \to +\infty$ (perfect reranking), the failure probability becomes ϵ^N, as expected (see (2)), demonstrating the model's ability to interpolate between scenarios of random reranking ($\lambda = 0$) with a failure probability of ϵ (see (3)), and optimal reranking ($\lambda \to +\infty$) with a failure probability of ϵ^N. A plot is shown in Fig. 2 (left), for several values of $e^{-\lambda} \in [0, 1]$.

Our next result, proved in App. A.1, shows that, even with an imperfect reranker, an asymptotically error-free protocol is possible:

Proposition 1. *When R is a Mallows reranker, for any $\lambda > 0$, the protocol is asymptotically error-free and the error probability decays exponentially fast, $P_{\text{err}}(N; q) = \mathcal{O}((e^{-\lambda}(1-\epsilon)+\epsilon)^N)$.*

This result shows that $P_{\text{err}}(N; q)$ converges Q-linearly to zero with rate of convergence $e^{-\lambda}(1-\epsilon) + \epsilon > \epsilon$. Therefore, **Mallows rerankers behave asymptotically as a perfect reranker but where the generator has an increased error probability.**

Given this result, one might wonder whether any reranker "slightly better than random" suffices to obtain an asymptotically error-free protocol. This it **not** the case, as the next counter-example shows.

Example 1. *Assume a reranker with probability mass function $\eta_j \propto (N - j + 1)$. The resulting protocol is not asymptotically error-free; we have $P_{\text{err}}(N; q) = \mathcal{O}(\epsilon^2)$. Therefore, the error is reduced from $\mathcal{O}(\epsilon)$ to $\mathcal{O}(\epsilon^2)$ but it is not eliminated. More generally, if $\eta_j \propto (N - j + 1)^r$ for a fixed positive integer r, we have $P_{\text{err}}(N; q) = \mathcal{O}(\epsilon^{r+1})$. See App. A.2 for a proof and plots.*

Next, we present a class of rerankers weaker than Mallows which still lead to error-free protocols.

3.3 Imperfect reranker: Zipf-Mandelbrot model

For Mallows models (using the Kendall-tau distance), the marginal probabilities (5) can be written as $\boldsymbol{\eta} = \text{softmax}(-\lambda \boldsymbol{z})$, where $\boldsymbol{z} = [0, 1, ..., N-1]^\top$. We now consider transformations that yield distributions with heavier tails, which we will see later in §5 to be a better empirical fit in several applications. A known extension to softmax is the **γ-entmax** (Peters et al., 2019),[5] a family of transformations parametrized by $\gamma \geq 0$,

$$\gamma\text{-entmax}(\boldsymbol{z}) := [1 + (\gamma - 1)(\boldsymbol{z} - \tau \mathbf{1})]_+^{1/(\gamma-1)}, \tag{7}$$

which recovers softmax as a limit case when $\gamma \to 1$. In (7), τ is a constant which ensures that $\gamma\text{-entmax}(\boldsymbol{z})$ is normalized. When $\gamma > 1$, γ-entmax can return sparse distributions (Blondel et al., 2020). Conversely, when $\gamma < 1$, γ-entmax leads to heavy-tailed distributions (see App. A.3).

[5] Peters et al. (2019) call this α-entmax; we use γ instead not to clash the notation to be introduced in §4.

Let us now consider $\boldsymbol{\eta} = \gamma\text{-entmax}(-\lambda \boldsymbol{z})$, where $\boldsymbol{z} = [0, 1, ..., N-1]^\top$, instead of (5). Letting $p := 1/(1-\gamma)$, $b = \lambda/p$, and $a = \frac{p+\tau}{\lambda} - 1$ (where a is seen here as a normalizing constant that replaces τ), and assuming $a > -1$ and $\gamma < 1$, we can write the γ-entmax model as $\eta_j = b^{-p}(a+j)^{-p}$. Note that $\gamma < 1$ is equivalent to $p > 1$. This is called a **Zipf-Mandelbrot model** (Zipf, 1932; Mandelbrot, 1965). This model generalizes the famous Zipf's law, which applies empirically to many practical contexts, such as the frequency table of words in a corpus of natural language (Powers, 1998). The constant a is determined to satisfy $\sum_{j=1}^{N}(a+j)^{-p} = b^p$. When $N \to \infty$, the left hand side becomes the Hurwitz zeta function (Hurwitz, 1882), which equals the Riemann's zeta when $a = 0$,

$$\zeta(p, a+1) := \sum_{j=1}^{\infty} \frac{1}{(a+j)^p} = \frac{1}{\Gamma(p)} \int_0^\infty dt \frac{t^{p-1}}{e^{(a+1)t}(1-e^{-t})}. \tag{8}$$

The following result, proved in App. A.4, shows that Zipf-Mandelbrot rerankers (which are weaker than Mallows rerankers and become the latter when $\gamma \to 1$) still ensure error-free protocols. The proof makes use of the integral representation of the Hurwitz zeta function (8) and of the dominated convergence theorem, reusing the result for Mallows models in Proposition 1.

Proposition 2. *When R is a Zipf-Mandelbrot reranker, for any $\lambda > 0$ and $\gamma < 1$, the protocol is asymptotically error-free.*

Fig. 2 (right) shows how this model differs from the one presented in §3.2. Since the reranker is weaker, the error curves bend causing the error decrease to be slower, but still convergent to zero.

4 Generator-Reranker Systems with Dependent Hypotheses

We assume now a more realistic scenario where the independence assumption of §3 might not hold. For example, $(X_1, Y_1), ..., (X_N, Y_N)$ might be only **exchangeable**—this is the case, for example, when the hypotheses are generated from G by sampling from a given model, conditioned on the query. In communication theory parlance, this assumes the presence of channel "interference" that introduces dependencies between the errors at the various channels, although permuting the messages at each channel does not change the joint distribution. By de Finetti's theorem (Diaconis and Freedman, 1980), exchangeability implies that there is some mixture variable $h \in \mathcal{H}$ such that $\mathbb{P}(x_{1:N}, y_{1:N}) = \int_{\mathcal{H}} d\mathbb{P}(h) \prod_{i=1}^{N} \mathbb{P}(x_i|h)\mathbb{P}(y_i|x_i, h)$.

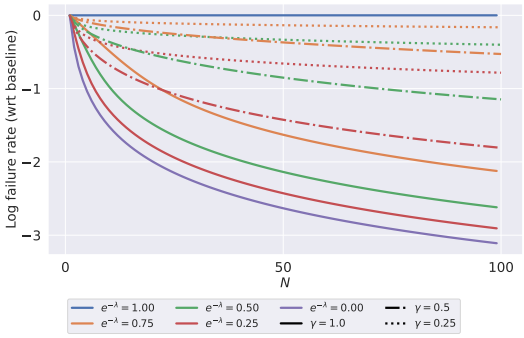

Figure 3: Log of the failure rate as a function of the number of generated **exchangeable** hypotheses N for several values of γ, $e^{-\lambda}$, and $\epsilon = \alpha = 0.3$.

We assume further that $h = (q, \tau)$ can be decoupled into the query variable q, which conditions x, and a random variable τ, which conditions y, such that $\mathbb{P}(x_i|h) := \mathbb{P}(x_i|q)$ and $\mathbb{P}(y_i|x_i, h) := \mathbb{P}(y_i|x_i, \tau)$. This corresponds to the graphical model in Fig. 1 (right), including the part in red. We let τ be a continuous random variable in $[0, 1]$ such that $\mathbb{E}[\tau] = \epsilon = \mathbb{P}(Y_i \neq \mathcal{X}(q) \mid X_i)$. A convenient choice is a Beta distribution with parameters α and β, $p(\tau; \alpha, \beta) := \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \tau^{\alpha-1}(1-\tau)^{\beta-1}$.

Perfect reranker and Beta coupling. If R is a perfect reranker, the error probability is

$$\begin{aligned} P_{\text{err}}(N; q) &= \mathbb{P}(g(Y_1, ..., Y_N) \notin \mathcal{X}(q) \mid q) = \mathbb{E}_{X_{1:N}|q}\big[\mathbb{P}(g(Y_1, ..., Y_N) \notin \mathcal{X}(q) \mid X_{1:N}\big] \\ &= \mathbb{E}_{X_{1:N}}\left[\int_0^1 d\tau\ p(\tau) \prod_{i=1}^{N} \underbrace{\mathbb{P}(Y_i \notin \mathcal{X}(q) \mid X_i, \tau)}_{=\tau}\right] = \mathbb{E}_\tau[\tau^N]. \end{aligned} \tag{9}$$

When $\tau \sim \text{Beta}(\tau; \alpha, \beta)$, the N^{th}-raw moment (9) has a closed form, leading to $P_{\text{err}}(N; q) = \prod_{i=1}^{N} \frac{\alpha+i-1}{\alpha+\beta+i-1}$. The next result, proved in App. A.5 using Gautschi's inequality (Gautschi, 1959) and the Stirling's formula, shows that we still obtain an error-free protocol, albeit the error decays slower than in the independent case—no longer exponentially but rather following a power law.

Proposition 3. *When $\tau \sim \text{Beta}(\tau; \alpha, \beta)$ and with a perfect reranker, the protocol is error-free and the error probability decays as a power law, $P_{\text{err}}(N; q) = \mathcal{O}(N^{-\beta})$. Furthermore, for $\beta < 1$, we have $P_{\text{err}}(N; q) \in \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)} (\alpha + \beta + N)^{-\beta}, \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)} (\alpha + \beta + N - 1)^{-\beta} \right)$.*

Imperfect reranker. When $\tau \sim \text{Beta}(\tau; \alpha, \beta)$, the probability of exactly K out of N messages being corrupted is (due to the conjugacy between the Beta prior and the binomial distribution) $\binom{N}{K} \int_0^1 d\tau \; p(\tau; \alpha, \beta) \tau^K (1-\tau)^{N-K} = \binom{N}{K} \frac{\prod_{i=1}^{K} (\alpha+i-1) \prod_{i=1}^{N-K} (\beta+i-1)}{\prod_{i=1}^{N} (\alpha+\beta+i-1)}$. Therefore, using the reranker marginals $\boldsymbol{\eta}$ as in (4), we get

$$P_{\text{err}}(N; q) = \sum_{K=0}^{N} \binom{N}{K} \frac{\prod_{i=1}^{K} (\alpha + i - 1) \prod_{i=1}^{N-K} (\beta + i - 1)}{\prod_{i=1}^{N} (\alpha + \beta + i - 1)} \sum_{j=N-K+1}^{N} \eta_j, \tag{10}$$

which leads to the plot in Fig. 3 for Mallows and Zipf-Mandelbrot models.[6]

The next result, proved in App. A.6, shows that the dependencies considered in this subsection do not compromise the error-free protocol when it exists for any density $p(\tau)$ which is finite in $(0, 1)$ (not necessarily a Beta distribution). The proof invokes the dominated convergence theorem to enable commuting the limit with the integral sign.

Proposition 4. *Let G_τ be a generator producing independent hypotheses (§3) where each hypothesis is acceptable with probability $1 - \tau$. Let the reranker R be such that (G_τ, R) has error probability $P_{\text{err}}^{\text{indep}}(N; q, \tau) \to 0$ for every $\tau \in (0, 1)$ (i.e., it is asymptotically error-free). Assume that the function $\tau \mapsto P_{\text{err}}^{\text{indep}}(N; q, \tau)$ is measurable for every $N \in \mathbb{N}$. Then, when R is used with a generator G which produces exchangeable hypotheses with arbitrary distribution $p(\tau)$, finite in $(0, 1)$, the system (G, R) is still asymptotically error-free.*

This result has important implications: it tells us that, to design error-free protocols, it is sufficient to verify if they are error-free in the simpler case where hypotheses are independent.

5 Experiments

In this section, we demonstrate the validity of our reranking laws on two different tasks:[7] text-to-code generation (§5.1) and machine translation of medical data (§5.2). Following existing literature on scaling laws for language modeling, we fit all curves on the development set using least squares (Ghorbani et al., 2022, App. E) and plot them on the *unseen* test set.[8] In all cases, we consider the generalized model presented in §4 with parameters α, β, and a Zipf-Mandelbrot reranker with parameters γ, and $e^{-\lambda}$, which becomes a Mallows reranker when $\gamma \to 1$. This is done in two steps: first, we fit α and β using the data for the perfect reranker ($e^{-\lambda} = 0$). Then, we fit γ and $e^{-\lambda}$ using the already estimated α and β and the data for the imperfect reranker. Our code is available at https://github.com/deep-spin/reranking-laws.

5.1 Code generation

We use a sanitized version of the MBPP dataset (Austin et al., 2021; Liu et al., 2023), a widely used benchmark for evaluating code LLMs, which includes 400 programming problems in Python. For each problem, the dataset includes ground-truth programs and three test cases with input and ground-truth output. We split the dataset in two equally sized parts to get development and test splits.

We generate 200 hypotheses with DeepSeek-Coder 7B (Guo et al., 2024) using a sampling temperature of 1 (see App. B.1 for the prompt template). As in previous work, for simplicity, we use only one test case for each problem (Shi et al., 2022), and select one candidate by taking a **majority vote** over the

[6] Since $\tau \sim \text{Beta}(\tau; \alpha, \beta)$, we have $\mathbb{E}[\tau] = \alpha/(\alpha + \beta)$, which we set to ϵ to match the independent setting from §3, resulting in $\beta = (\epsilon^{-1} - 1)\alpha$. Hence, α is our only new free parameter. As $\alpha \to 0^+$, the hypotheses become maximally dependent and reranking is hopeless; as $\alpha \to \infty$, the scenario reverts to full independence.

[7] App. B.3 contains additional experiments on mathematical and commonsense reasoning benchmarks.

[8] We use `scipy.optimize.least_squares`.

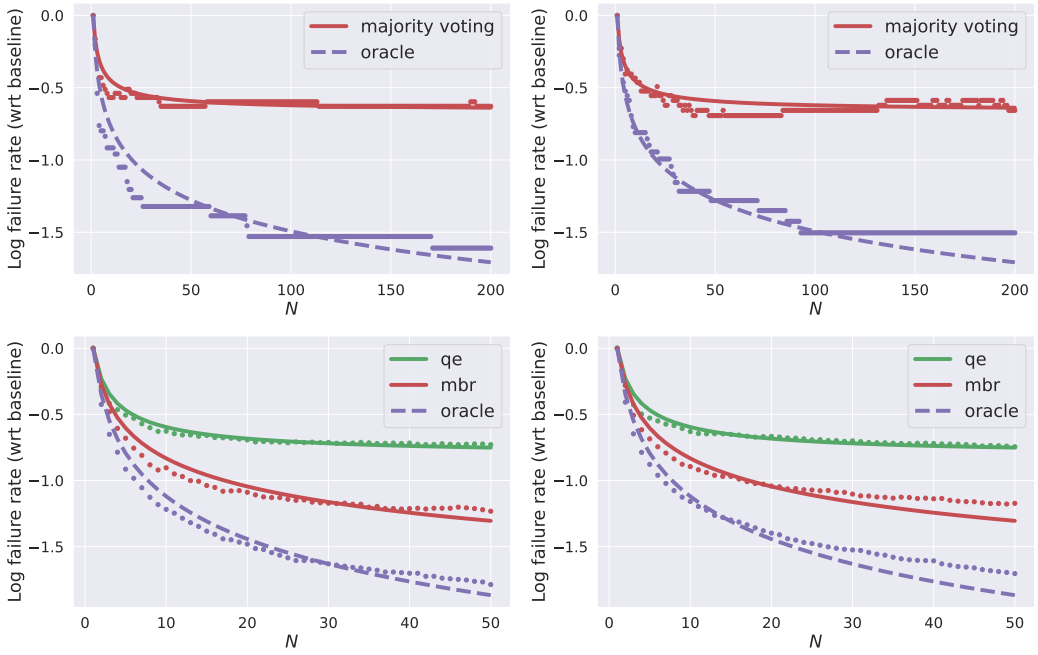

Figure 4: Log of the failure rate as a function of N. The empirical data is represented with dots (**left:** dev, **right:** test set) and our fitted models with solid and dashed lines (imperfect and perfect reranker, respectively). **Top:** text-to-code generation (§5.1). **Bottom:** machine translation (§5.2).

execution results, dismissing hypotheses that fail to execute on the test case (Wang et al., 2023). A hypothesis is considered unacceptable if the result of at least one test case (out of three) is different from the ground truth.

Fig. 4 (top) shows the log failure rate on the dev and test sets (left and right, respectively) as a function of N. Even though the oracle fit is not perfect, we get $\alpha = .1$, $\beta = .309$, $\gamma = .001$, and $e^{-\lambda} = .003$ for the imperfect reranker with majority voting, which fits the data well, as shown by the red curve.

5.2 Machine translation

We use the TICO-19 dataset (Anastasopoulos et al., 2020), which includes 3071 English sentences in the medical domain (*i.e.*, COVID-19 related content) translated into 38 languages. We use the official splits, which contain 971 examples for development and 2100 for testing, focusing on translating from English (EN) to Portuguese (PT), Spanish (ES), and Russian (RU).

For each source sentence, we sample 50 translation hypotheses with a temperature of 1 from TowerInstruct 13B (Alves et al., 2024) using the prompt template in App. B.2.[9] Folowing Farinhas et al. (2023), we consider two reranking strategies: selecting the best candidate with **MBR decoding** using COMET-22 as the utility metric (Eikema and Aziz, 2020; Rei et al., 2022a) and **reranking based on quality estimation** using the reference-free CometKiwi (Fernandes et al., 2022; Rei et al., 2022b). Since we cannot afford to collect human evaluation scores for each sampled hypothesis, we consider a translation to have a critical mistake (*i.e.*, to be unacceptable) if its COMET-22 score is below 0.85, and an **oracle** (perfect) reranker that picks the translation with the highest COMET-22 score.

We follow the described procedure using the data from all language pairs together. Fig. 4 (bottom) shows the log failure rate on the dev and test sets as a function of N. We get $\alpha = 0.1$ and $\beta = 0.46$. Additionally, we have $\gamma = 0.182$ and $e^{-\lambda} = 0.001$ for MBR decoding and $\gamma = 0.001$ and $e^{-\lambda} = 0.005$ for QE reranking. See App. B.2 for additional plots showing these curves when the data

[9]This model outperforms all existing open-source alternatives (even of larger scales) for translating content between the supported languages and is also competitive with GPT-4 (OpenAI et al., 2023), especially when combined with MBR decoding (Alves et al., 2024, App. A).

from each language pair is used to fit a separate model. Again, we see a reasonable fit, especially for the imperfect rerankers, with MBR decoding leading to lower failure rates than reranking with QE.

6 Discussion and Related Work

We believe the communication-theoretic perspective introduced in this paper might inspire the design of new protocols for increasing the quality and safety of LLMs. The generator-reranker system studied in this paper bears resemblance with repetition codes, a very naive (and inefficient) class of error-correcting codes. Can more powerful designs (Hamming, 1950; Reed and Solomon, 1960; Gallager, 1962; Berrou et al., 1993) inspire more efficient protocols? In machine translation, other forms of adding redundancy, such as lattice generation (Singhal et al., 2023) and hypothesis recombination (Vernikos and Popescu-Belis, 2024), suggest that more efficient designs are indeed possible.

Recent work also suggests that **LLM-based evaluators** could be used as highly effective rerankers in specific tasks (Kim et al., 2024). While LLMs are not yet ready to fully replace human evaluators across diverse NLP tasks (Bavaresco et al., 2024), in some cases, they can even provide fine-grained assessments in addition to single scores (Kocmi and Federmann, 2023; Fernandes et al., 2023a).

Another class of communication systems allow for **feedback**, *e.g.*, in "automatic repeat request" protocols (Lin et al., 1984), where the receiver has a backchannel to request the sender to retransmit missing bits of information. This framework can be useful to analyze LLM protocols where the generator generates a varying number of hypotheses interactively, relying on feedback from another module, such as a reward model or a confidence estimator, as in Quach et al. (2023). Communication with feedback was also used recently by Jung et al. (2024) for summarization when the generator error probability ϵ is large—our mild conditions for asymptotically error-free protocols (Propositions 1–4) suggest that "bootstrapping" a correct answer is possible even in scenarios where G is very weak. Additionally, recent work has shown that LLMs may struggle with planning or self-verification, advocating instead for tighter integration between LLMs and external model-based verifiers (Kambhampati et al., 2024). This supports our view that using external feedback models can improve LLMs by enabling interactive, error-correcting communication.

We provide **reranking laws**, which allow us to predict how many hypotheses are necessary to achieve a desired error probability. This links to a rich body of literature aiming to predict the performance of deep learning models in terms of fundamental parameters, such as the model size or the amount of compute and data used to train them (Hestness et al., 2017, 2019). These so called "neural scaling laws" have been studied in the context of language modeling (Kaplan et al., 2020; Hoffmann et al., 2022) and machine translation (Ghorbani et al., 2022; Fernandes et al., 2023b), where we observe a power-law scaling for the performance as a function of each fundamental parameter. Our paper complements this line of work by considering the decoding dimension for generator-reranker systems.

The analysis and theoretical results of this paper focus on binary acceptable/unacceptable decisions; however it is possible to extend our framework to consider also **continuous quality metrics** (such as COMET scores for translation (Rei et al., 2020)) by replacing the notion of "asymptotically error-free" protocol (Definition 1) by a more general concept associated to a quality target. A possible path is to posit a probability *density* for the continuous quality metric (instead of a Bernoulli error probability) for each hypothesis coming from the generator, such as a Gaussian or uniform distribution with some input-dependent parameters. For a perfect reranker and independent hypotheses, the resulting output after reranking would be distributed according to the corresponding *extreme value distribution* (this models the distribution of the *highest* evaluation metric score among the N hypotheses). Extreme value distributions are an important subject of study in order statistics (David and Nagaraja, 2004) and their densities have closed form expressions in some restricted cases: for example, the Gaussian assumption above yields a Gumbel distribution, and a uniform assumption yields a Beta distribution. The asymptotic case ($N \to \infty$) corresponds to one of Gumbel, Fréchet or Weibull families (this is a consequence of the Fisher–Tippett–Gnedenko theorem (David and Nagaraja, 2004)). From the extreme value distribution, we can obtain the *expected* evaluation metric score or the probability of a quality score being below an acceptable threshold. However, the generalization to imperfect rerankers (such as the Mallows or Zipf-Mandelbrot rerankers described in § 3.2 and 3.3) seems harder than in the binary case and requires further investigation.

7 Conclusions

We presented a communication-theoretic perspective of generator-reranker LLMs, where the generator G is conceptualized as a sender transmitting N descriptions in parallel through noisy channels, and the reranker R decodes the message by selecting the most appropriate description. Under mild conditions, the combined system (G, R) yields an acceptable answer almost surely when $N \to \infty$. Experiments on text-to-code generation and machine translation with LLMs validate our framework.

8 Limitations and Broader Impacts

We regard our paper as a first step connecting communication theory and LLMs, as discussed in §6. However, it should be noted that our work has several limitations. First, the guarantees of error-free protocol in Propositions 1–4 are only asymptotic, and in certain cases a large N may be necessary to achieve a large enough error decrease. We provide convergence rates only for Mallows rerankers (with independent hypotheses and also in the dependent case, when combined with a Beta prior). Second, there is no simple recipe to determine if the Mallows and Zipf-Mandelbrot reranker models are a good empirical fit to concrete rerankers. The same applies to the prior distribution $p(\tau)$ which makes hypotheses dependent. Third, while our experiments in §5 suggest a reasonable fit in two tasks (code generation and machine translation), the fit is not perfect. A challenge is that, for large N, errors are rare events, and therefore prone to statistical inaccuracies (this is visible in the "steps" observed in the code generation plots). Finally, although our framework focuses on binary acceptable/unacceptable decisions, it can be extended to continuous evaluation metrics, but this would require modifications to some concepts (*e.g.*, the notion of asymptotically error-free protocols). Despite these limitations, the binary case remains highly relevant in practice—for example, in code generation, where the output either executes correctly or it does not. We expect future work to overcome some of these limitations.

In considering the broader impact of our work, it is crucial to acknowledge its early stage and predominantly theoretical nature, which lends the discussion a speculative quality. We believe that our research can significantly enhance the reliability of LLMs by facilitating the identification of potential system failures, holding promise in fields such as natural language processing and computer vision, where robustness and error prediction are paramount. While not directly addressing environmental concerns shared across different LLMs (Strubell et al., 2019), our work could indirectly contribute to energy efficiency efforts by quantifying the efficiency of reranking methods, potentially reducing computational requirements while maintaining requisite quality thresholds during inference.

Acknowledgments

We would like to thank Ben Peters, Duarte Alves, Marcos Treviso, Mário Figueiredo, Sweta Agrawal, and the SARDINE lab team for helpful discussions. This work was supported by EU's Horizon Europe Research and Innovation Actions (UTTER, contract 101070631), by the project DECOLLAGE (ERC-2022-CoG 101088763), by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Center for Responsible AI), and by Fundação para a Ciência e Tecnologia through contract UIDB/50008/2020.

References

Pranjal Aggarwal, Aman Madaan, Yiming Yang, and Mausam. Let's sample step by step: Adaptive-consistency for efficient reasoning and coding with LLMs. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12375–12396, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.761. URL https://aclanthology.org/2023.emnlp-main.761.

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. Tower: An open multilingual large language model for translation-related tasks, 2024.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. TICO-19: the translation initiative for COvid-19. In Karin Verspoor, Kevin Bretonnel Cohen, Michael Conway, Berry de Bruijn, Mark Dredze, Rada Mihalcea, and Byron Wallace, editors, *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online, December 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.nlpcovid19-2.5. URL https://aclanthology.org/2020.nlpcovid19-2.5.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models, 2021.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks, 2024. URL https://arxiv.org/abs/2406.18403.

Claude Berrou, Alain Glavieux, and Punya Thitimajshima. Near shannon limit error-correcting coding and decoding: Turbo-codes. 1. In *Proceedings of ICC'93-IEEE International Conference on Communications*, volume 2, pages 1064–1070. IEEE, 1993.

Amanda Bertsch, Alex Xie, Graham Neubig, and Matthew Gormley. It's MBR all the way down: Modern generation techniques through the lens of minimum Bayes risk. In Yanai Elazar, Allyson Ettinger, Nora Kassner, Sebastian Ruder, and Noah A. Smith, editors, *Proceedings of the Big Picture Workshop*, pages 108–122, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.bigpicture-1.9. URL https://aclanthology.org/2023.bigpicture-1.9.

Sumanta Bhattacharyya, Amirmohammad Rooshenas, Subhajit Naskar, Simeng Sun, Mohit Iyyer, and Andrew McCallum. Energy-based reranking: Improving neural machine translation using energy-based models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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 4528–4537, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.349. URL https://aclanthology.org/2021.acl-long.349.

Mathieu Blondel, AndrÃ© F.T. Martins, and Vlad Niculae. Learning with fenchel-young losses. *Journal of Machine Learning Research*, 21(35):1–69, 2020. URL http://jmlr.org/papers/v21/19-021.html.

Sebastian Borgeaud and Guy Emerson. Leveraging sentence similarity in natural language generation: Improving beam search using range voting. In Alexandra Birch, Andrew Finch, Hiroaki Hayashi, Kenneth Heafield, Marcin Junczys-Dowmunt, Ioannis Konstas, Xian Li, Graham Neubig, and Yusuke Oda, editors, *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 97–109, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.ngt-1.11. URL https://aclanthology.org/2020.ngt-1.11.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri,

Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021.

Flavio Chierichetti, Anirban Dasgupta, Shahrzad Haddadan, Ravi Kumar, and Silvio Lattanzi. Mallows models for top-k lists. *Advances in Neural Information Processing Systems*, 31, 2018.

Gonçalo M. Correia, Vlad Niculae, and André F. T. Martins. Adaptively sparse transformers. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2174–2184, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1223. URL https://aclanthology.org/D19-1223.

Gonçalo Correia, Vlad Niculae, Wilker Aziz, and André Martins. Efficient marginalization of discrete and structured latent variables via sparsity. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11789–11802. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/887caadc3642e304ede659b734f79b00-Paper.pdf.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006. ISBN 0471241954.

H.A. David and H.N. Nagaraja. *Order Statistics*. Wiley Series in Probability and Statistics. Wiley, 2004. ISBN 9780471654018. URL https://books.google.pt/books?id=bdhzFXg6xFkC.

Persi Diaconis and David Freedman. Finite exchangeable sequences. *The Annals of Probability*, pages 745–764, 1980.

Bryan Eikema and Wilker Aziz. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.398. URL https://aclanthology.org/2020.coling-main.398.

António Farinhas, José de Souza, and Andre Martins. An empirical study of translation hypothesis ensembling with large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11956–11970, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.733. URL https://aclanthology.org/2023.emnlp-main.733.

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. Quality-aware decoding for neural machine translation. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.100. URL https://aclanthology.org/2022.naacl-main.100.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.100. URL https://aclanthology.org/2023.wmt-1.100.

Patrick Fernandes, Behrooz Ghorbani, Xavier Garcia, Markus Freitag, and Orhan Firat. Scaling laws for multilingual neural machine translation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10053–10071. PMLR, 23–29 Jul 2023b. URL https://proceedings.mlr.press/v202/fernandes23a.html.

M. A. Fligner and J. S. Verducci. Distance based ranking models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 48(3):359–369, 1986. ISSN 00359246. URL http://www.jstor.org/stable/2345433.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. High quality rather than high model probability: Minimum bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825, 2022.

Robert Gallager. Low-density parity-check codes. *IRE Transactions on information theory*, 8(1): 21–28, 1962.

A.E. Gamal and T. Cover. Achievable rates for multiple descriptions. *IEEE Transactions on Information Theory*, 28(6):851–857, 1982. doi: 10.1109/TIT.1982.1056588.

Walter Gautschi. Some elementary inequalities relating to the gamma and incomplete gamma function. *J. Math. Phys*, 38(1):77–81, 1959.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 04 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00370. URL https://doi.org/10.1162/tacl_a_00370.

Behrooz Ghorbani, Orhan Firat, Markus Freitag, Ankur Bapna, Maxim Krikun, Xavier Garcia, Ciprian Chelba, and Colin Cherry. Scaling laws for neural machine translation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=hR_SMu8cxCV.

Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. Hallucinations in Large Multilingual Translation Models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517, 12 2023. ISSN 2307-387X. doi: 10.1162/tacl_a_00615. URL https://doi.org/10.1162/tacl_a_00615.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Y. Wu, Y. K. Li, Fuli Luo, Yingfei Xiong, and Wenfeng Liang. Deepseek-coder: When the large language model meets programming – the rise of code intelligence, 2024.

Richard W Hamming. Error detecting and error correcting codes. *The Bell system technical journal*, 29(2):147–160, 1950.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically, 2017.

Joel Hestness, Newsha Ardalani, and Gregory Diamos. Beyond human-level accuracy: computational challenges in deep learning. In *Proceedings of the 24th Symposium on Principles and Practice of Parallel Programming*, PPoPP '19, page 1–14, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362252. doi: 10.1145/3293883.3295710. URL https://doi.org/10.1145/3293883.3295710.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack William Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=iBBcRUlOAPR.

James Y Huang, Sailik Sengupta, Daniele Bonadiman, Yi-an Lai, Arshit Gupta, Nikolaos Pappas, Saab Mansour, Katrin Kirchoff, and Dan Roth. Deal: Decoding-time alignment for large language models. *arXiv preprint arXiv:2402.06147*, 2024.

Chia-Chien Hung, Wiem Ben Rim, Lindsay Frost, Lars Bruckner, and Carolin Lawrence. Walking a tightrope – evaluating large language models in high-risk domains. In Dieuwke Hupkes, Verna Dankers, Khuyagbaatar Batsuren, Koustuv Sinha, Amirhossein Kazemnejad, Christos Christodoulopoulos, Ryan Cotterell, and Elia Bruni, editors, *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 99–111, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.genbench-1.8. URL https://aclanthology.org/2023.genbench-1.8.

Adolf Hurwitz. Einige eigenschaften der dirichlet'schen funktionen, die bei der bestimmung der klassenanzahlen binärer quadratischer formen auftreten. *Zeitschrift fur Mathematik und Physik*, 1882.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12), mar 2023. ISSN 0360-0300. doi: 10.1145/3571730. URL https://doi.org/10.1145/3571730.

Jaehun Jung, Peter West, Liwei Jiang, Faeze Brahman, Ximing Lu, Jillian Fisher, Taylor Sorensen, and Yejin Choi. Impossible distillation: from low-quality model to high-quality dataset & model for summarization and paraphrasing. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2024.

Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant Bhambri, Lucas Paul Saldyt, and Anil B Murthy. Position: LLMs can't plan, but can help planning in LLM-modulo frameworks. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 22895–22907. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/kambhampati24a.html.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=8euJaTveKw.

Alexandre Klementiev, Dan Roth, and Kevin Small. Unsupervised rank aggregation with distance-based models. In *Proceedings of the 25th international conference on Machine learning*, pages 472–479, 2008.

Alexandre Klementiev, Dan Roth, Kevin Small, and Ivan Titov. Unsupervised rank aggregation with domain-specific expertise. In *Twenty-First International Joint Conference on Artificial Intelligence*, 2009.

Tom Kocmi and Christian Federmann. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.64. URL https://aclanthology.org/2023.wmt-1.64.

Shankar Kumar and William Byrne. Minimum Bayes-risk word alignments of bilingual texts. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 140–147. Association for Computational Linguistics, July 2002. doi: 10.3115/1118693.1118712. URL https://aclanthology.org/W02-1019.

J.N. Laneman, E. Martinian, G.W. Wornell, and J.G. Apostolopoulos. Source-channel diversity for parallel channels. *IEEE Transactions on Information Theory*, 51(10):3518–3539, 2005. doi: 10.1109/TIT.2005.855578.

Guy Lebanon and Yi Mao. Non-parametric modeling of partially ranked data. *Journal of Machine Learning Research*, 9(79):2401–2429, 2008. URL http://jmlr.org/papers/v9/lebanon08a.html.

Haau-Sing Li, Patrick Fernandes, Iryna Gurevych, and André F. T. Martins. Doce: Finding the sweet spot for execution-based code generation, 2024. URL https://arxiv.org/abs/2408.13745.

Zichao Li, Prakhar Sharma, Xing Han Lu, Jackie Cheung, and Siva Reddy. Using interactive feedback to improve the accuracy and explainability of question answering systems post-deployment. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 926–937, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.75. URL https://aclanthology.org/2022.findings-acl.75.

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. Can large language models reason about medical questions? *Patterns*, 5(3):100943, 2024. ISSN 2666-3899. doi: https://doi.org/10.1016/j.patter.2024.100943. URL https://www.sciencedirect.com/science/article/pii/S2666389924000424.

Shu Lin, Daniel J Costello, and Michael J Miller. Automatic-repeat-request error-control schemes. *IEEE Communications magazine*, 22(12):5–17, 1984.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and LINGMING ZHANG. Is your code generated by chatGPT really correct? rigorous evaluation of large language models for code generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=1qvx610Cu7.

Tianlin Liu, Shangmin Guo, Leonardo Bianco, Daniele Calandriello, Quentin Berthet, Felipe Llinares, Jessica Hoffmann, Lucas Dixon, Michal Valko, and Mathieu Blondel. Decoding-time realignment of language models. *arXiv preprint arXiv:2402.02992*, 2024.

David J. C. MacKay. *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, USA, 2002. ISBN 0521642981.

Benoît Mandelbrot. Information theory and psycholinguistics. *BB Wolman and E*, 1965.

André Martins, António Farinhas, Marcos Treviso, Vlad Niculae, Pedro Aguiar, and Mario Figueiredo. Sparse and continuous attention mechanisms. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20989–21001. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f0b76267fbe12b936bd65e203dc675c1-Paper.pdf.

André F. T. Martins, Marcos Treviso, António Farinhas, Pedro M. Q. Aguiar, Mário A. T. Figueiredo, Mathieu Blondel, and Vlad Niculae. Sparse continuous distributions and fenchel-young losses. *Journal of Machine Learning Research*, 23(257):1–74, 2022. URL http://jmlr.org/papers/v23/21-0879.html.

Pedro Henrique Martins, Vlad Niculae, Zita Marinho, and André F. T. Martins. Sparse and structured visual attention. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 379–383, 2021. doi: 10.1109/ICIP42928.2021.9506028.

Ansong Ni, Srini Iyer, Dragomir Radev, Veselin Stoyanov, Wen-Tau Yih, Sida Wang, and Xi Victoria Lin. LEVER: Learning to verify language-to-code generation with execution. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 26106–26128. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/ni23b.html.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2023.

L. Ozarow. On a source-coding problem with two channels and three receivers. *The Bell System Technical Journal*, 59(10):1909–1921, 1980. doi: 10.1002/j.1538-7305.1980.tb03344.x.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple math word problems? In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.168. URL https://aclanthology.org/2021.naacl-main.168.

Ben Peters and André F. T. Martins. Smoothing and shrinking the sparse Seq2Seq search space. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy,

Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2642–2654, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.210. URL https://aclanthology.org/2021.naacl-main.210.

Ben Peters, Vlad Niculae, and André F. T. Martins. Sparse sequence-to-sequence models. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1504–1519, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1146. URL https://aclanthology.org/P19-1146.

David MW Powers. Applications and explanations of zipf's law. In *New methods in language processing and computational natural language learning*, 1998.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S Jaakkola, and Regina Barzilay. Conformal language modeling. In *The Twelfth International Conference on Learning Representations*, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Irving S Reed and Gustave Solomon. Polynomial codes over certain finite fields. *Journal of the society for industrial and applied mathematics*, 8(2):300–304, 1960.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.213. URL https://aclanthology.org/2020.emnlp-main.213.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid), December 2022a. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.52.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid), December 2022b. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.60.

Freda Shi, Daniel Fried, Marjan Ghazvininejad, Luke Zettlemoyer, and Sida I. Wang. Natural language to code translation with execution. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3533–3546, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.231. URL https://aclanthology.org/2022.emnlp-main.231.

Prasann Singhal, Jiacheng Xu, Xi Ye, and Greg Durrett. Eel: Efficiently encoding lattices for reranking. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9299–9316, 2023.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650,

Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1355. URL https://aclanthology.org/P19-1355.

Wenpin Tang. Mallows ranking models: maximum likelihood estimate and regeneration. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6125–6134. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/tang19a.html.

Giorgos Vernikos and Andrei Popescu-Belis. Don't rank, combine! combining machine translation hypotheses using quality estimation. *arXiv preprint arXiv:2401.06688*, 2024.

Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. Executable code actions elicit better LLM agents. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024. URL https://openreview.net/forum?id=8oJyuXfrPv.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. Rrhf: Rank responses to align language models with human feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

George Kingsley Zipf. *Selected studies of the principle of relative frequency in language*. Harvard university press, 1932.

A Proofs and Visualizations

A.1 Proof of Proposition 1

Let $\lambda_\epsilon := -\log\left(e^{-\lambda}(1-\epsilon)+\epsilon\right)$ and define $F(N) = \log P_{\text{err}}(N;q) = \frac{e^{-\lambda_\epsilon N}-e^{-\lambda N}}{1-e^{-\lambda N}}$. Observe that $0 < \lambda_\epsilon < \lambda$ for any $\lambda > 0$ and $\epsilon \in (0,1)$. We extend the domain of F to the real numbers in $[1,+\infty)$. We will prove that $F(N)$ is decreasing and that $\lim_{N\to\infty} F'(N) = -\lambda_\epsilon$. This shows that $P_{\text{err}}(N;q) \to 0$ at asymptotic rate $e^{-\lambda_\epsilon}$. We have

$$F'(N) = \frac{(e^{-\lambda_\epsilon N}-e^{-\lambda N})'}{e^{-\lambda_\epsilon N}-e^{-\lambda N}} - \frac{(1-e^{-\lambda N})'}{1-e^{-\lambda N}} = \frac{-\lambda_\epsilon e^{-\lambda_\epsilon N}+\lambda e^{-\lambda N}}{e^{-\lambda_\epsilon N}-e^{-\lambda N}} - \frac{\lambda e^{-\lambda N}}{1-e^{-\lambda N}} \le 0,$$

hence $F(N)$ is decreasing. Since the second term tends to zero, the limit is given by the first term:

$$\lim_{N\to\infty} F'(N) = \lim_{N\to\infty} \frac{-\lambda_\epsilon e^{-\lambda_\epsilon N}+\lambda e^{-\lambda N}}{e^{-\lambda_\epsilon N}-e^{-\lambda N}} = \lim_{N\to\infty} \frac{-\lambda_\epsilon}{1-e^{(-\lambda+\lambda_\epsilon)N}} + \frac{\lambda}{e^{(-\lambda_\epsilon+\lambda)N}-1} = -\lambda_\epsilon,$$

where we used the fact that $e^{(-\lambda+\lambda_\epsilon)N} \to 0$ and $e^{(-\lambda_\epsilon+\lambda)N} \to +\infty$. This proves the desired claim, that is, the error probability decreases exponentially fast with rate $e^{-\lambda_\epsilon}$. Note that, for a perfect reranker ($\lambda \to \infty$), we get $e^{-\lambda_\epsilon} = \epsilon$ and we recover the rate ϵ^N seen in §3.1.

A.2 Proof of Example 1

We first provide a proof for $r = 1$. We have $\sum_{j=N-K+1}^{N} \eta_j = \sum_{j=1}^{K} \eta_{N-K+j} = \frac{\sum_{j=1}^{K} j}{\sum_{j=1}^{N} j} = \frac{K(K+1)}{N(N+1)}$. Plugging this into Eq. (4), we obtain

$$\begin{aligned} P_{\text{err}}(N;q) &= \sum_{K=0}^{N} \binom{N}{K} \epsilon^K (1-\epsilon)^{N-K} \frac{K^2+K}{N^2+N} = \frac{\mathbb{E}_{K\sim B(N,\epsilon)}[K^2+K]}{N^2+N} \\ &= \frac{N\epsilon(1-\epsilon)+N^2\epsilon^2+N\epsilon}{N(N+1)} = \frac{\epsilon(1-\epsilon)+N\epsilon^2+\epsilon}{N+1}, \end{aligned} \tag{11}$$

where $B(N,\epsilon)$ denotes the binomial distribution with parameters N and ϵ and we use the facts that $\mathbb{E}_{K\sim B(N,\epsilon)}[K] = N\epsilon$ and $\mathbb{E}_{K\sim B(N,\epsilon)}[K^2] = N\epsilon(1-\epsilon)+N^2\epsilon^2$. Therefore, we obtain $\lim_{N\to\infty} P_{\text{err}}(N;q) = \epsilon^2$.

We now prove the general case $r \ge 1$. From Faulhaber's formula, we have $\sum_{j=1}^{K} j^r = \frac{1}{r+1}\sum_{j=0}^{r}\binom{r+1}{j} B_j K^{r-j+1}$, where $B_j = \sum_{\ell=0}^{j} \frac{1}{\ell+1}\sum_{m=0}^{\ell}\binom{\ell}{m}(-1)^m (m+1)^j$ denotes the j^{th} Bernoulli number. Therefore, we get

$$\sum_{j=N-K+1}^{N} \eta_j = \sum_{j=1}^{K} \eta_{N-K+j}^r = \frac{\sum_{j=1}^{K} j^r}{\sum_{j=1}^{N} j^r} = \frac{\sum_{j=0}^{r}\binom{r+1}{j} B_j K^{r-j+1}}{\sum_{j=0}^{r}\binom{r+1}{j} B_j N^{r-j+1}}. \tag{12}$$

Plugging this into Eq. (4), we obtain

$$\begin{aligned} P_{\text{err}}(N;q) &= \sum_{K=0}^{N} \binom{N}{K} \epsilon^K (1-\epsilon)^{N-K} \frac{\sum_{j=0}^{r}\binom{r+1}{j} B_j K^{r-j+1}}{\sum_{j=0}^{r}\binom{r+1}{j} B_j N^{r-j+1}} \\ &= \frac{\sum_{j=0}^{r}\binom{r+1}{j} B_j \mathbb{E}_{K\sim B(N,\epsilon)}[K^{r-j+1}]}{\sum_{j=0}^{r}\binom{r+1}{j} B_j N^{r-j+1}}. \end{aligned} \tag{13}$$

We now use the fact that the raw moments of the binomial distribution $B(N,\epsilon)$ are given by $\mathbb{E}_{K\sim B(N,\epsilon)}[K^m] = \sum_{\ell=0}^{m} \left\{ {m \atop \ell} \right\} N^{\underline{\ell}} \epsilon^\ell$, where $\left\{ {m \atop \ell} \right\} := \frac{1}{\ell!}\sum_{i=0}^{\ell}(-1)^{\ell-i}\binom{\ell}{i} i^m$ are the Stirling numbers of the second kind, and $N^{\underline{\ell}} := \frac{N!}{(N-\ell)!}$ is the ℓ^{th} falling power of N. Therefore, when $N \to \infty$, (13) becomes

$$\lim_{N\to\infty} P_{\text{err}}(N;q) = \lim_{N\to\infty} \frac{\binom{r+1}{0} B_0 \overbrace{\left\{ {r+1 \atop r+1} \right\}}^{=1} N^{\underline{r+1}} \epsilon^{r+1}}{\binom{r+1}{0} B_0 N^{r+1}} = \epsilon^{r+1}. \tag{14}$$

The plots in Fig. 5 show examples for several values of r.

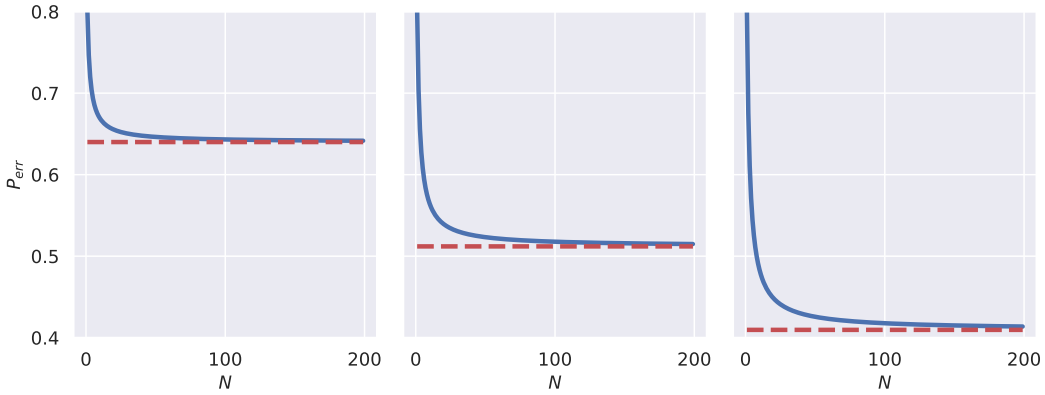

Figure 5: P_{err} using rerankers with probability mass function $\eta_j \propto (N - j + 1)^r$ with $r = \{1, 2, 3\}$ (from left to right) and $\epsilon = 0.8$. The resulting protocol is not asymptotically error-free: the horizontal asymptotes in red correspond to ϵ^{r+1}, according to Eq. (14).

A.3 Entmax

When $\gamma > 1$, γ-entmax can return sparse distributions (Blondel et al., 2020). This case has been extensively studied as a way to, *e.g.*, filter large output spaces (Correia et al., 2020; Peters and Martins, 2021) or to produce more interpretable predictions (Correia et al., 2019; Martins et al., 2020, 2021, 2022). Conversely, when $\gamma < 1$, γ-entmax leads to distributions with heavier tails, which is the case of our interest, as described in §3.3. See Fig. 6 for an illustration of γ-entmax for different values of γ in the two-dimensional case. For $\gamma > 1$, all mappings saturate at $z = \pm 1/\gamma - 1$; this does not happen for $\gamma \leq 1$.

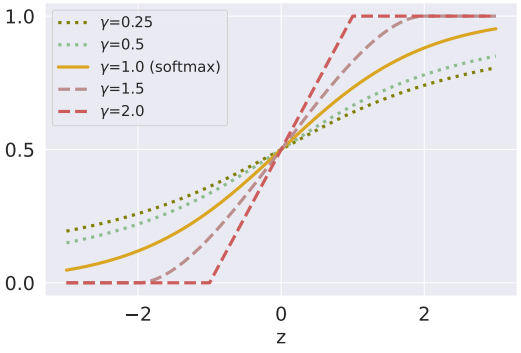

Figure 6: Two-dimensional γ-entmax$([z, 0])_1$.

A.4 Proof of Proposition 2

Note that we can write

$$\sum_{j=1}^{N} \frac{1}{(a+j)^p} = \zeta(p, a+1) - \zeta(p, a+N+1)$$

and

$$\begin{aligned}
\sum_{j=N-K+1}^{N} \eta_j &= b^{-p}(\zeta(p, a+1) - \zeta(p, a+N+1) - \zeta(p, a+1) + \zeta(p, a+N-K+1)) \\
&= b^{-p}(\zeta(p, a+N-K+1) - \zeta(p, a+N+1)) \\
&= \frac{1}{b^p \Gamma(p)} \int_0^\infty dt \frac{t^{p-1}}{e^{(a+N+1)t}(1-e^{-t})}(e^{Kt} - 1). \qquad (15)
\end{aligned}$$

The error probability is

$$
\begin{aligned}
P_{\text{err}}(N;q) &= \sum_{K=0}^{N} \left[\binom{N}{K} \epsilon^K (1-\epsilon)^{N-K} \sum_{j=N-K+1}^{N} \eta_j \right] \\
&= \frac{1}{b^p \Gamma(p)} \int_0^\infty dt \frac{t^{p-1}}{e^{(a+N+1)t}(1-e^{-t})} \underbrace{\sum_{K=0}^{N} \binom{N}{K} \epsilon^K (1-\epsilon)^{N-K} (e^{Kt}-1)}_{=(e^t\epsilon+1-\epsilon)^N-1} \\
&= \frac{1}{b^p \Gamma(p)} \int_0^\infty dt \frac{t^{p-1}[(e^t\epsilon+1-\epsilon)^N-1]}{e^{(a+N+1)t}(1-e^{-t})} \\
&= \frac{1}{b^p \Gamma(p)} \int_0^\infty dt \frac{t^{p-1}}{e^{(a+1)t}(1-e^{-t})} \frac{(e^t\epsilon+1-\epsilon)^N-1}{e^{tN}} \\
&= \frac{1}{b^p \Gamma(p)} \int_0^\infty dt \frac{t^{p-1}}{e^{(a+1)t}(1-e^{-t})} \underbrace{[((1-\epsilon)e^{-t}+\epsilon)^N - e^{-tN}]}_{:=f_N(t)\to 0}.
\end{aligned} \tag{16}
$$

Since a is the normalizing constant such that $1 = \zeta(p, a+1) = \frac{1}{b^p\Gamma(p)} \int_0^\infty dt \frac{t^{p-1}}{e^{(a+1)t}(1-e^{-t})}$ (cf. Eq. (8)), we can interpret the expression above as the expectation of $f_N(t) := ((1-\epsilon)e^{-t}+\epsilon)^N - e^{-tN}$ under the probability distribution on $(0,\infty)$ with density $\pi(t) := \frac{1}{b^p\Gamma(p)} \frac{t^{p-1}}{e^{(a+1)t}(1-e^{-t})}$. Since $f_N(t) \to 0$ pointwise for $t \in]0,\infty[$ and it is bounded in that interval, we can invoke the dominated convergence theorem to commute the limit and integral sign. We then have that $P_{\text{err}}(N;q) \to 0$.

A.5 Proof of Proposition 3

Let us consider first the case where $\beta = 1$. Then,

$$
P_{\text{err}}(N;q) = \prod_{i=1}^{N} \frac{\alpha+i-1}{\alpha+\beta+i-1} = \frac{\alpha}{\alpha+1}\frac{\alpha+1}{\alpha+2}\cdots\frac{\alpha+N-1}{\alpha+N} = \frac{\alpha}{\alpha+N} \to 0.
$$

Now consider the case where $\beta > 1$. We have for each term in the product that $\frac{\alpha+i-1}{\alpha+\beta+i-1} < \frac{\alpha+i-1}{\alpha+i}$, hence we must have $P_{\text{err}}(N;q) < \frac{\alpha}{\alpha+N}$. Since the sequence is positive (since all terms are positive) and decreasing (since all terms are < 1), we must also have $P_{\text{err}}(N;q) \to 0$ when $\beta > 1$.

Finally, let us analyze the case where $0 < \beta < 1$. From (9), we have

$$
\begin{aligned}
P_{\text{err}}(N;q) = \mathbb{E}_\tau[\tau^N] &= \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \tau^{\alpha-1}(1-\tau)^{\beta-1}\tau^N \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)} \frac{\Gamma(\alpha+N)}{\Gamma(\alpha+\beta+N)} \underbrace{\int_0^1 \frac{\Gamma(\alpha+\beta+N)}{\Gamma(\alpha+N)\Gamma(\beta)} \tau^{\alpha+N-1}(1-\tau)^{\beta-1}}_{=1} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)} \frac{\Gamma(\alpha+N)}{\Gamma(\alpha+\beta+N)}.
\end{aligned} \tag{17}
$$

We invoke Gautschi's inequality, which states that $x^{1-s} < \frac{\Gamma(x+1)}{\Gamma(x+s)} < (x+1)^{1-s}$ for any x and $s \in (0,1)$. We set $s := 1-\beta$ and $x := \alpha+\beta+N-1$, from which we obtain the desired result.

To show that the error probability decays as a power law for any $\beta > 0$, we use Stirling's formula, which states that

$$
\Gamma(z) = \sqrt{\frac{2\pi}{z}} \left(\frac{z}{e}\right)^z \left(1 + \mathcal{O}\left(\frac{1}{z}\right)\right). \tag{18}
$$

Therefore,

$$\begin{aligned}
\lim_{N\to\infty} \frac{\Gamma(\alpha+N)}{\Gamma(\alpha+\beta+N)} &= \lim_{N\to\infty} \frac{\sqrt{\frac{2\pi}{\alpha+N}}\left(\frac{\alpha+N}{e}\right)^{\alpha+N}}{\sqrt{\frac{2\pi}{\alpha+\beta+N}}\left(\frac{\alpha+\beta+N}{e}\right)^{\alpha+\beta+N}} \\
&= \lim_{N\to\infty} \underbrace{\sqrt{\frac{\alpha+\beta+N}{\alpha+N}}}_{\to 1} \underbrace{\left(\frac{\alpha+N}{\alpha+\beta+N}\right)^{\alpha+N}}_{\to e^{-\beta}} \left(\frac{\alpha+\beta+N}{e}\right)^{-\beta} \\
&= \lim_{N\to\infty} (\alpha+\beta+N)^{-\beta} = \mathcal{O}(N^{-\beta}). \qquad (19)
\end{aligned}$$

A.6 Proof of Proposition 4

Let $P_{\text{err}}^{\text{indep}}(N; q, \tau)$ denote the error probability of the generator-reranker system when the hypotheses are independent and each hypothesis produced by G has error probability τ. The error probability of the generator-reranker system with exchangeable hypotheses is given by

$$\begin{aligned}
P_{\text{err}}(N; q) &= \mathbb{P}(g(Y_1, ..., Y_N) \notin \mathcal{X}(q) \mid q) = \mathbb{E}_{X_{1:N}|q}\left[\mathbb{P}(g(Y_1, ..., Y_N) \notin \mathcal{X}(q) \mid X_{1:N}, q)\right] \\
&= \mathbb{E}_{X_{1:N}|q}\left[\int_0^1 d\tau\ p(\tau)\ \mathbb{P}(g(Y_1, ..., Y_N) \notin \mathcal{X}(q) \mid X_{1:N}, \tau)\right] \\
&= \int_0^1 d\tau\ p(\tau)\ \underbrace{\mathbb{E}_{X_{1:N}|q}\left[\mathbb{P}(g(Y_1, ..., Y_N) \notin \mathcal{X}(q) \mid X_{1:N}, \tau)\right]}_{=P_{\text{err}}^{\text{indep}}(N; q, \tau)}. \qquad (20)
\end{aligned}$$

Therefore, $\lim_{N\to\infty} P_{\text{err}}(N; q) = \lim_{N\to\infty} \int_0^1 d\tau\ p(\tau) P_{\text{err}}^{\text{indep}}(N; q, \tau)$. Since $P_{\text{err}}^{\text{indep}}(N; q, \tau) \in [0, 1]$ for any $N \in \mathbb{N}$ and $\tau \in [0, 1]$, we have that $p(\tau) P_{\text{err}}^{\text{indep}}(N; q, \tau) \in [0, p(\tau)]$, and therefore the integrand is bounded by $p(\tau)$, which integrates to 1. Therefore we can invoke the dominated convergence theorem and switch the limit and integral signs. Since by assumption $\lim_{N\to\infty} P_{\text{err}}^{\text{indep}}(N; q, \tau) = 0$ pointwise for any $\tau \in (0, 1)$, we obtain $\lim_{N\to\infty} P_{\text{err}}(N; q) = \int_0^1 d\tau\ p(\tau) \lim_{N\to\infty} P_{\text{err}}^{\text{indep}}(N; q, \tau) = 0$.

B Experimental Details

B.1 Text-to-code generation

Licenses. We use DeepSeek-Coder 7B (Guo et al., 2024), which is available under a permissive license that allows for both research and unrestricted commercial use. We report results on the MBPP dataset (Austin et al., 2021; Liu et al., 2023), released under an Apache license.

Prompt template. We generate hypotheses with DeepSeek-Coder 7B (Guo et al., 2024) using the following prompt template:

> You are an AI programming assistant, utilizing the DeepSeek Coder model, developed by DeepSeek Company, and you only answer questions related to computer science. For politically sensitive questions, security and privacy issues, and other non-computer science questions, you will refuse to answer.
>
> ### Instruction:
>
> Please complete the following Python function in a markdown style code block:
>
> ```python
> [prompt]
> ```
>
> ### Response:
>
> ```python

MBR-exec. We use MBR-exec, an approach proposed by Shi et al. (2022) that consists of *(1)* sampling programs from an LLM, *(2)* executing each program on one test case, and *(3)* selecting the example with the minimal execution result-based Bayes risk. We use a 0/1 matching loss between execution results, and the Bayes risk of a program is defined by the sum of the loss between itself and the other sampled programs (the ground-truth program output is not used). We break ties by selecting the program with the smallest sampling index, corresponding to a random selection. See Shi et al. (2022, Section 3) for more details.

B.2 Machine translation

Licenses. We use TowerInstruct 13B (Alves et al., 2024), which is released under a CC-BY-NC-4.0 license. We report results on the TICO-19 dataset (Anastasopoulos et al., 2020), publicly available through a Creative Commons CC0 license.

Prompt template. We generate hypotheses with TowerInstruct 13B (Alves et al., 2024) using the following prompt template:

```
<|im_start|>user
Translate the following [source language] source text to [target language]:
[source language]: [source sentence]
[target language]: <|im_end|>
<|im_start|>assistant
```

Visualizations. In §5.2 we obtained a single reranking law for the all language pairs; we now fit different models for each language pair. Fig. 7 shows the log failure rate on the dev and test sets as a function of N for EN-PT, EN-ES, and EN-RU. While the fits on the dev set are good, there is some degradation on the test set, especially for EN-ES (oracle and MBR decoding), possibly due to a shift in the distribution of scores/errors. We leave the investigation of more robust techniques and how to adapt to these cases for future work.

B.3 Mathematical and commonsense reasoning

Our approach is fully general and can be useful in other domains other than code and language generation. In this subsection, we present additional experiments on mathematical and commonsense reasoning benchmarks, as prior work has shown that generating multiple hypotheses as an intermediate step is also advantageous in these scenarios (Wang et al., 2023).

We use samples generated by Aggarwal et al. (2023) with code-davinci-002, a GPT-3-based model with 175 billion parameters (Brown et al., 2020) which is part of the Codex series (Chen et al., 2021) (please refer to their Section 4 for more details; these samples were made publicly available by the authors at https://github.com/Pranjal2041/AdaptiveConsistency). We apply self-consistency over diverse reasoning paths (Wang et al., 2023), selecting the most frequent answer in the candidate set, and report results on the SVAMP (Patel et al., 2021) and StrategyQA (Geva et al., 2021) datasets. Following §5.1, we split the datasets in two equally sized parts to get development and test splits.

Similarly to Fig. 4, Fig. 8 shows the log failure rate on the dev and test sets (left and right, respectively) as a function of N, confirming that the same trends hold also for these two additional tasks.

B.4 Computing infrastructure

Our insfrastructure consists of 2 machines, each equipped with 8 NVIDIA RTX A6000 GPUs (46GB) and 12 Intel Xeon Gold 6348 CPUs (2.60GHz, 1TB RAM). The machines were used interchangeably, and all experiments were executed on a single GPU.

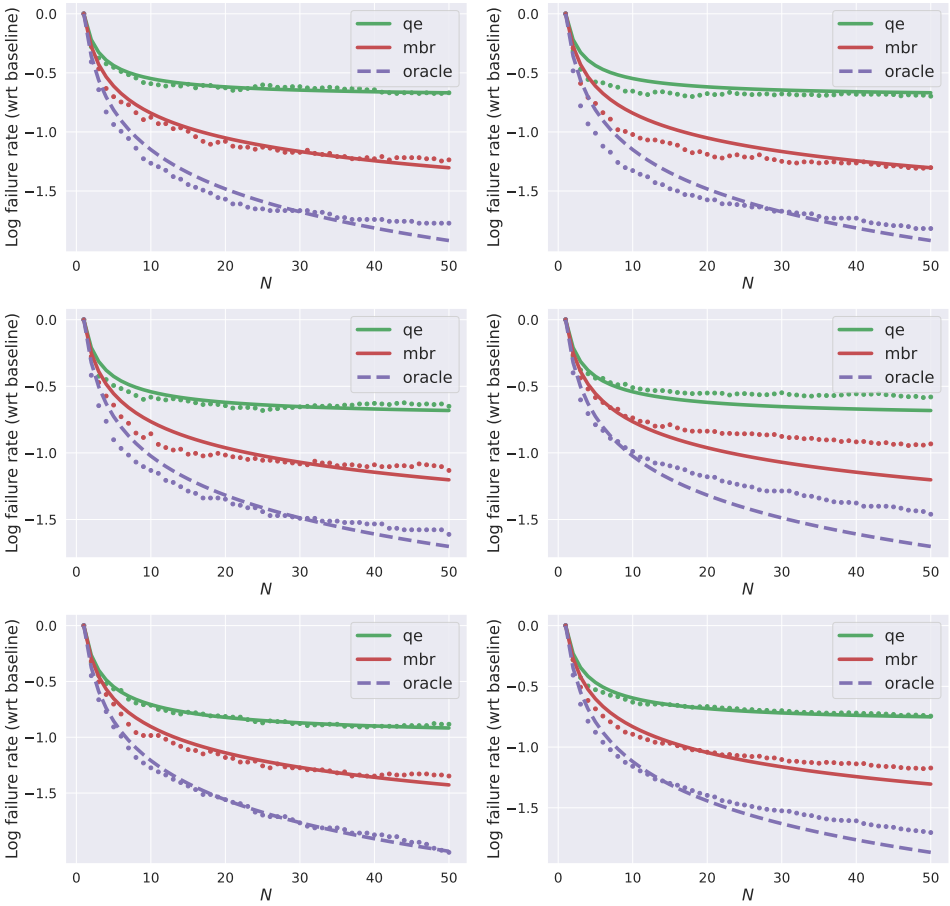

Figure 7: Log of the failure rate as a function of N. The empirical data is represented with dots (**left:** dev, **right:** test set) and our fitted models with solid and dashed lines (imperfect and perfect reranker, respectively). In this case, we fit separate models for each language pair (**from top to bottom:** EN-PT, EN-ES, and EN-RU).

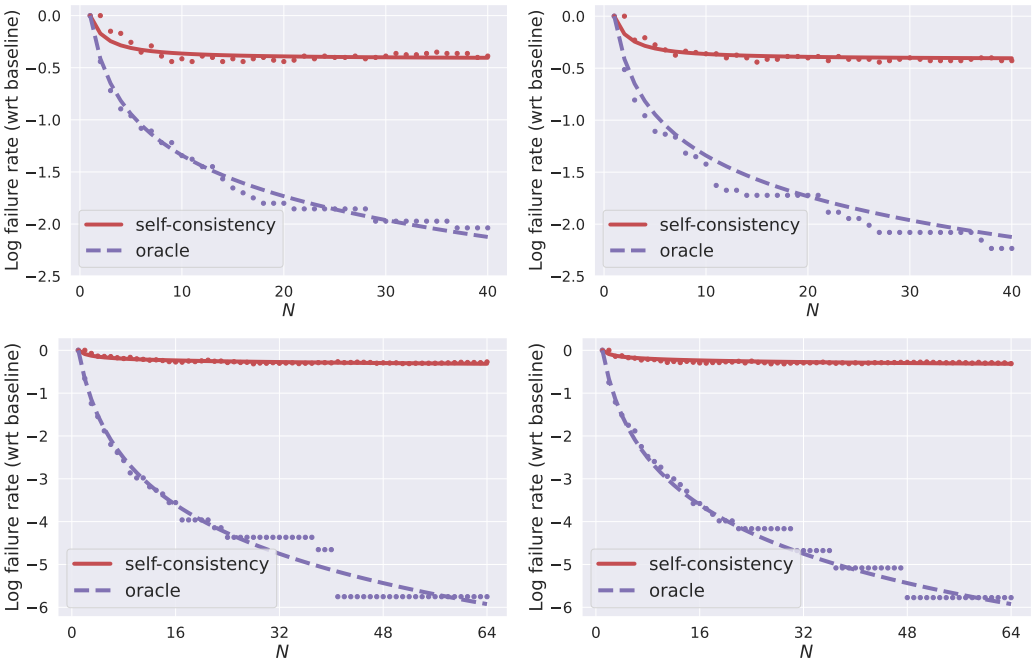

Figure 8: Log of the failure rate as a function of N. The empirical data is represented with dots (**left:** dev, **right:** test set) and our fitted models with solid and dashed lines (imperfect and perfect reranker, respectively). **Top:** mathematical reasoning on SVAMP . **Bottom:** commonsense reasoning on StrategyQA.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and the introduction state our assumptions and contributions. We summarize them in bullet points in §1.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our work in §8.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We state the full set of assumptions of our theoretical results (Propositions 1–4) and include complete proofs in App. A. We provide short proof sketches in the main paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide all the information needed to reproduce our results in §5 and App. B. Our code is available at https://github.com/deep-spin/reranking-laws.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We use publicly available data in our experiments (although we are not introducing new datasets). Our code is available at https://github.com/deep-spin/reranking-laws, with a CC-BY 4.0 license.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide all the details in §5 and App. B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: While we do not report error bars, for machine translation, we consider multiple language pairs and report individual and combined results (see §5.2 and App. B.2).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We include the type of computing resources used in our experiments in App. B.4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Our research conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the broader impact of our work in §8.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We are not releasing any data or models.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the creators of the datasets/models used in our paper, including the name of the corresponding licenses in App. B.1 and App. B.2.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our code is available at https://github.com/deep-spin/reranking-laws, with a CC-BY 4.0 license.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.