Reranking Laws for Language Generation: A Communication-Theoretic Perspective

NeurIPS 2024 (spotlight)

António Farinhas Instituto de Telecomunicações, Instituto Superior Técnico, Lisbon

Haau-Sing Li Ubiquitous Knowledge Processing Lab, TU Darmstadt

André Martins Instituto de Telecomunicações, Instituto Superior Técnico & Unbabel, Lisbon

https://github.com/deep-spin/reranking-laws antonio-farinhas.github.io

LLMs are great, but…

LLMs show remarkable performance across many tasks
in natural language processing, computer vision, speech recognition, …

LLMs are great, but...

LLMs show remarkable performance across many tasks
in natural language processing, computer vision, speech recognition, …

models generate critical mistakes/hallucinations
instances of hallucinations and other critical errors occasionally arise

LLMs are great, but...

LLMs show remarkable performance across many tasks
in natural language processing, computer vision, speech recognition, …

models generate critical mistakes/hallucinations
instances of hallucinations and other critical errors occasionally arise

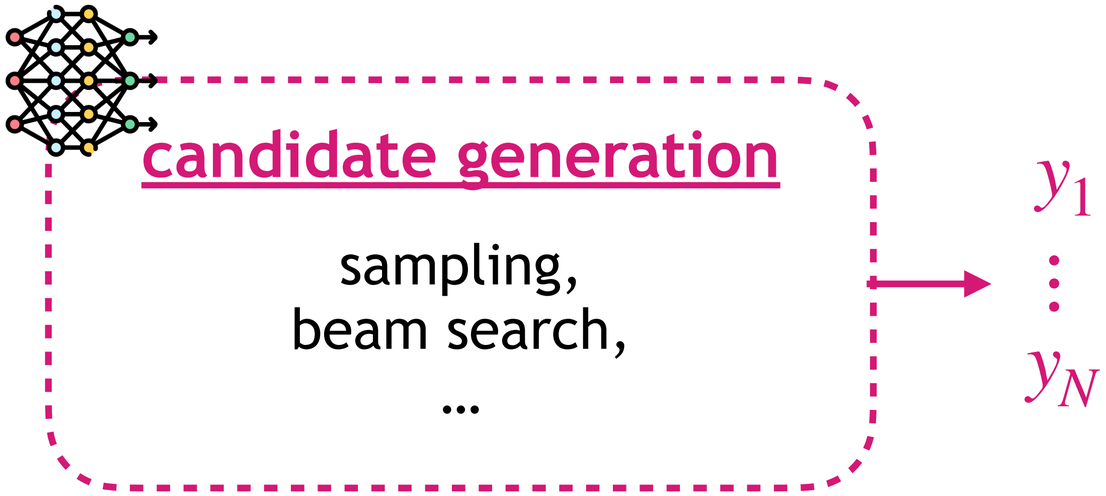
unreliable predictions
no clear indication of when and how badly models might fail

the most common mitigation strategy is to *steer* the LLM with the aid of a reward model or directly from human preferences

adding redundancy to improve quality

a simple decoding-time strategy:

(1) an LLM generates multiple hypotheses

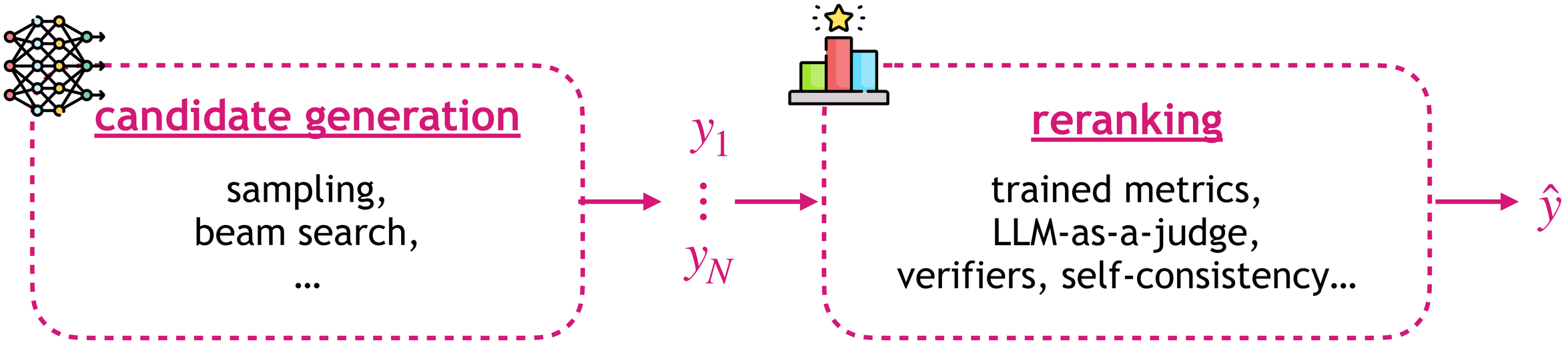


adding redundancy to improve quality

a simple decoding-time strategy:

(1) an LLM generates multiple hypotheses

(2) a reranker selects the most appropriate one



adding redundancy to improve quality

a simple decoding-time strategy:

(1) an LLM generates multiple hypotheses

(2) a reranker selects the most appropriate one

redundancy

adding redundancy as an intermediate step increases the chances of returning an acceptable answer

… also important in communication theory

adding redundancy to decrease the error rate in noisy channels is a cornerstone of communication theory

MacKay, 2002; Cover and Thomas, 2012; Hamming, 1950; Reed and Solomon, 1960; Gallager, 1962; Berrou et al., 1993

… also important in communication theory

adding redundancy to decrease the error rate in noisy channels is a cornerstone of communication theory

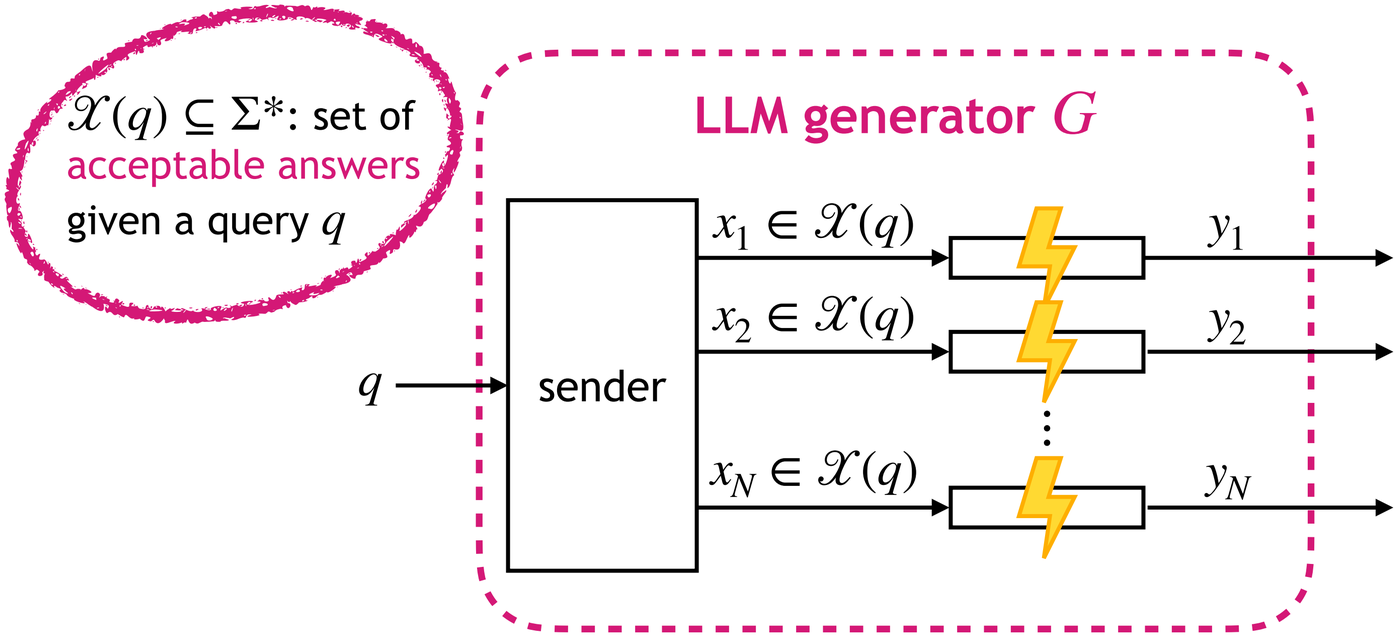
repetition codes

a message block is sent multiple times, the decoder uses majority voting to recover the original message with high probability

the same idea underlies more sophisticated error-correcting codes

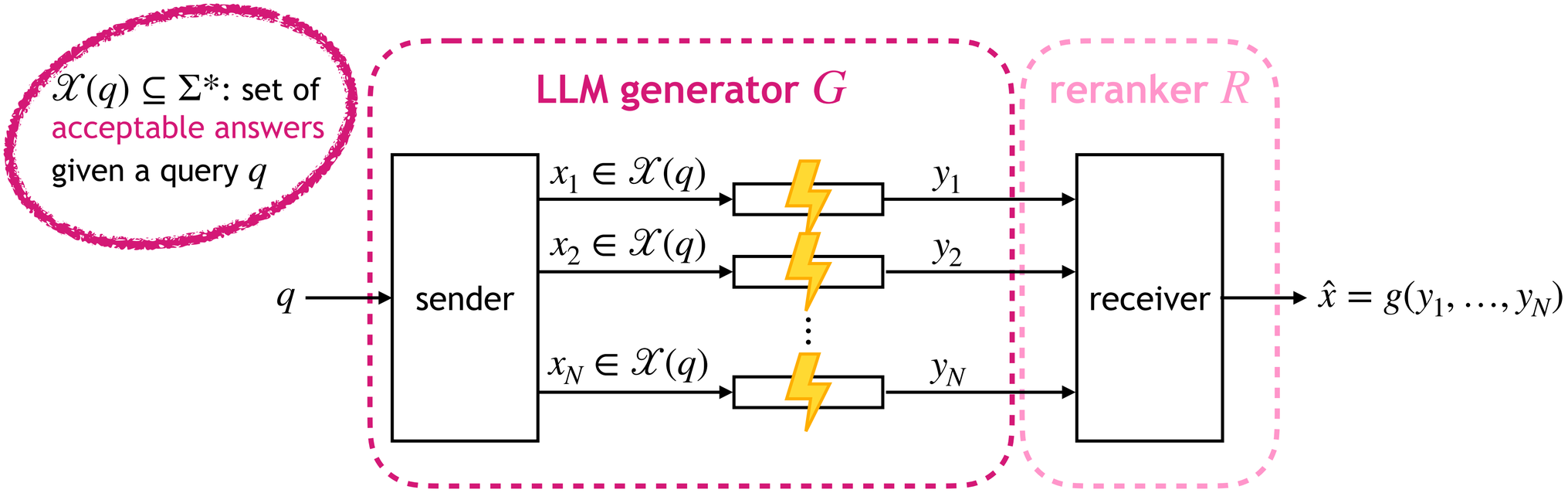
MacKay, 2002; Cover and Thomas, 2012; Hamming, 1950; Reed and Solomon, 1960; Gallager, 1962; Berrou et al., 1993

we draw a parallel between these two worlds



the sender transmits N message descriptions in parallel through noisy channels, leading to N potentially corrupted hypotheses

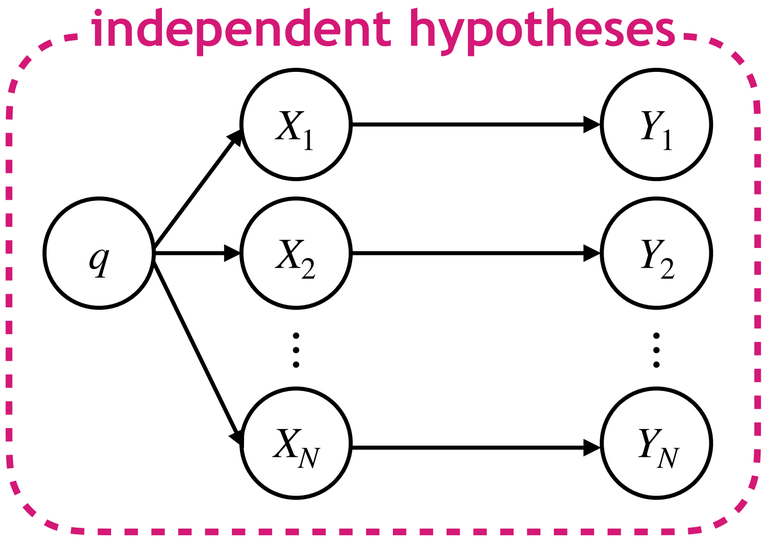
we draw a parallel between these two worlds



the sender transmits N message descriptions in parallel through noisy channels, leading to N potentially corrupted hypotheses

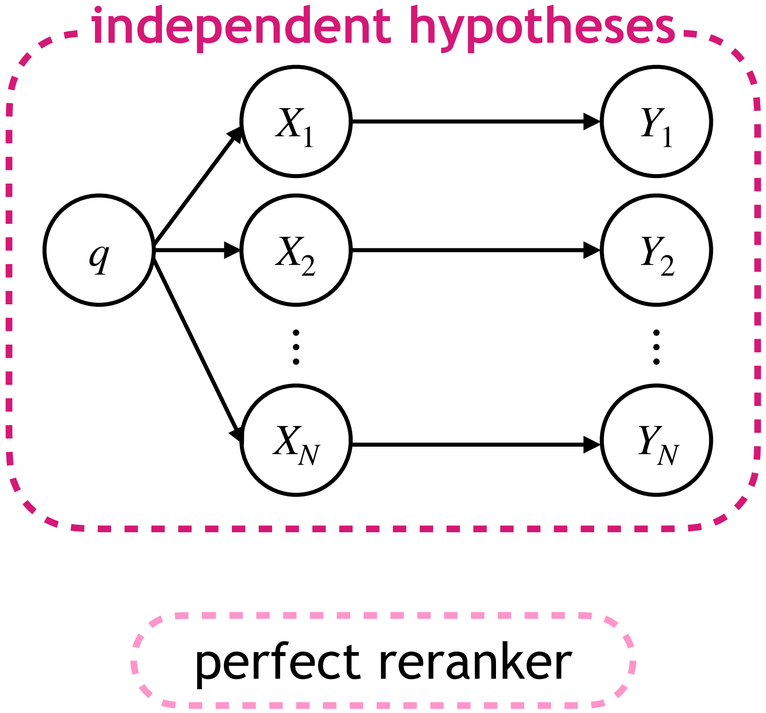
the receiver decodes the message by ranking the descriptions and selecting the one found to be most reliable

a simple case: independent hypotheses, perfect reranker



perfect reranker

a simple case: independent hypotheses, perfect reranker



perfect reranker

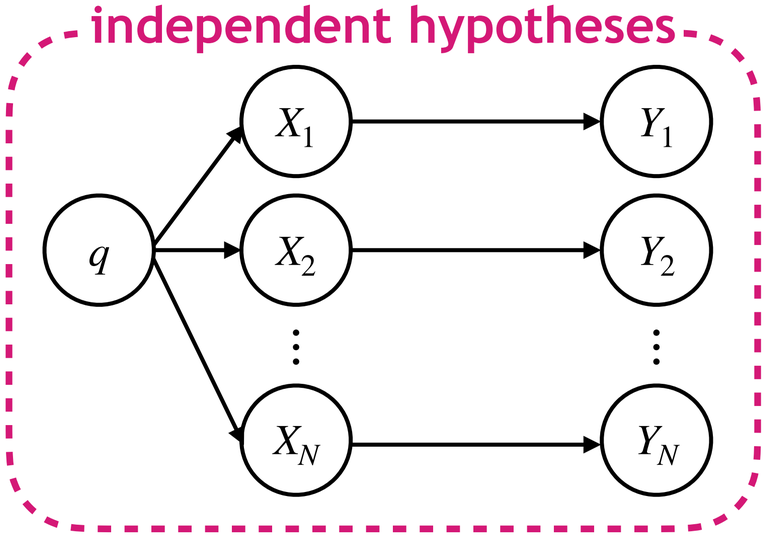
$$P_{\text{err}}(N; q) = \mathbb{P}(g(Y_{1:N}) \notin \mathcal{X}(q) \mid q)$$

$$= \mathbb{E}_{X_{1:N}|q}\left[\prod_{i=1}^{N} \underbrace{P(Y_i \notin \mathcal{X}(q) \mid X_i)}_{=\epsilon}\right]$$

$$= \epsilon^N \to 0$$

asymptotically error-free (AEF)

a simple case: independent hypotheses, perfect reranker



perfect reranker

$$P_{\text{err}}(N; q) = \mathbb{P}(g(Y_{1:N}) \notin \mathcal{X}(q) \mid q)$$

$$= \mathbb{E}_{X_{1:N}|q}\left[\prod_{i=1}^{N} \underbrace{P(Y_i \notin \mathcal{X}(q) \mid X_i)}_{=\epsilon}\right]$$

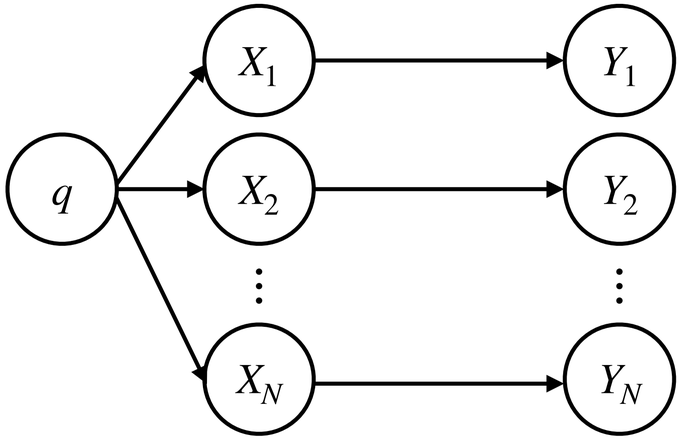
$$= \epsilon^N \to 0$$

asymptotically error-free (AEF)

reality is more complex: rerankers are not perfect, hypotheses are not independent

independent hypotheses, mallows reranker

independent hypotheses



mallows reranker

π_0 is the gold ranking and $d(\pi, \pi_0)$ the Kendall tau distance

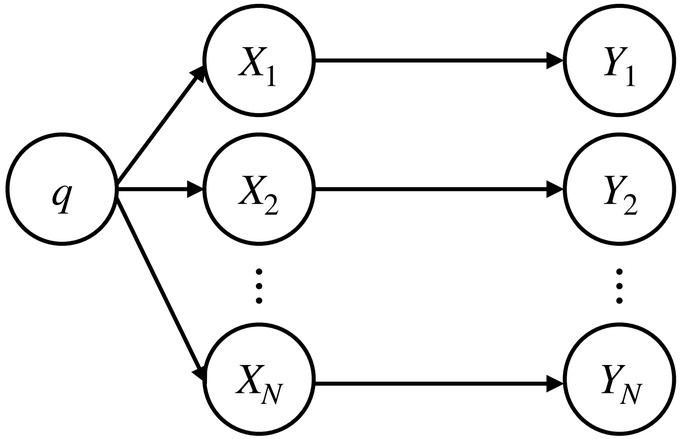
$$\mathbb{P}(\pi; \pi_0, \lambda) = \exp(-\lambda d(\pi, \pi_0))/Z(\lambda)$$

λ is a scale parameter

$\lambda \to \infty$ perfect reranker, $\lambda \to 0$ random reranker

independent hypotheses, mallows reranker

independent hypotheses



mallows reranker

π_0 is the gold ranking and $d(\pi, \pi_0)$ the Kendall tau distance

$$\mathbb{P}(\pi; \pi_0, \lambda) = \exp(-\lambda d(\pi, \pi_0))/Z(\lambda)$$

λ is a scale parameter

$\lambda \to \infty$ perfect reranker, $\lambda \to 0$ random reranker

$$P_{\text{err}}(N; q) = \begin{cases} \epsilon & \text{if } \lambda = 0 \\ \dfrac{[e^{-\lambda}(1-\epsilon)+\epsilon]^N - e^{-\lambda N}}{1 - e^{-\lambda N}} & \text{otherwise} \end{cases}$$

$$= \mathcal{O}((e^{-\lambda}(1-\epsilon)+\epsilon)^N) \to 0$$

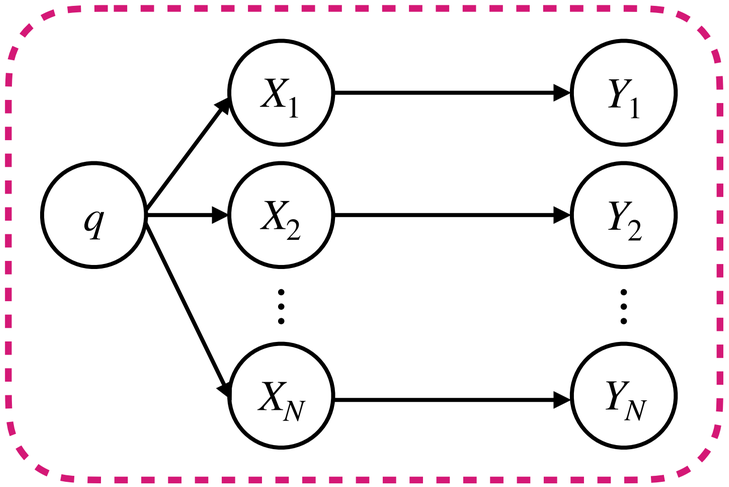
still AEF!

beyond perfect rerankers

we provide conditions under which this protocol is asymptotically error-free

independent hypotheses

- perfect — **err** $\to 0$ **(exponentially fast!)**
- Mallows — **err** $\to 0$ **(exponentially fast!)**
- Zipf-Mandelbrot — **err** $\to 0$



beyond independent hypotheses

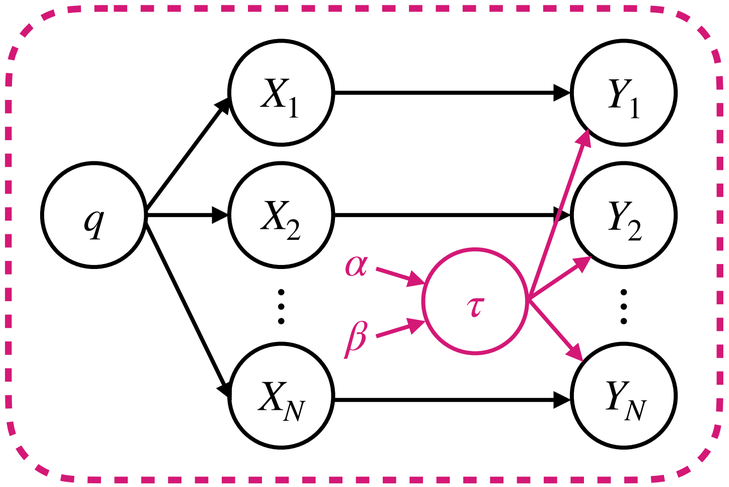
we provide conditions under which this protocol is asymptotically error-free

independent hypotheses

- perfect — **err** $\to 0$ **(exponentially fast!)**
- Mallows — **err** $\to 0$ **(exponentially fast!)**
- Zipf-Mandelbrot — **err** $\to 0$

dependent hypotheses (Beta prior)

- perfect — **err** $\to 0$ **(as a power law!)**
- Mallows — **err** $\to 0$
- Zipf-Mandelbrot — **err** $\to 0$



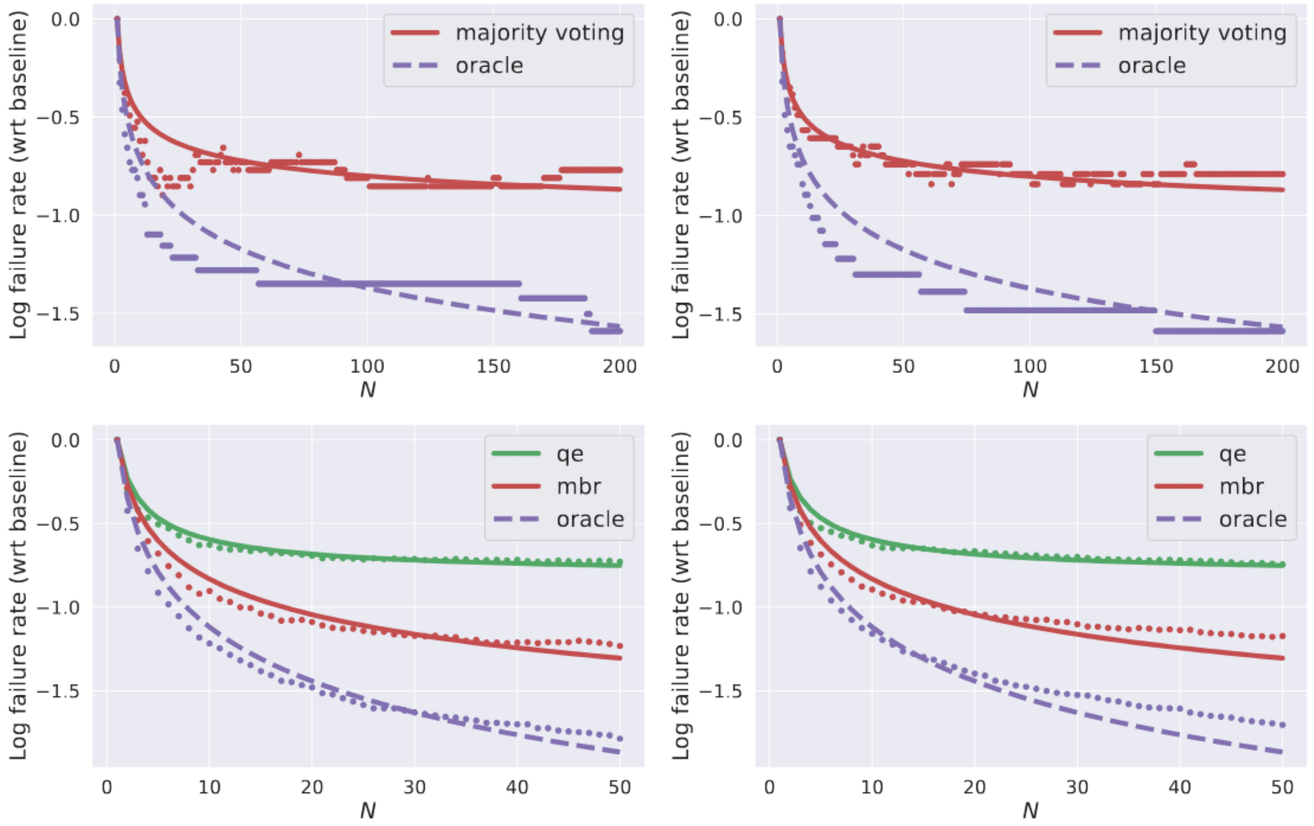
to design error-free protocols, it is sufficient to verify if they are error-free in the simpler case where hypotheses are independent

(proposition 4)

we validate our reranking laws empirically

	LLM generator	reranker	datasets
code generation	DeepSeek-Coder 7B	MBR-exec	MBPP
machine translation	TowerInstruct 13B	MBR decoding, QE reranking	TICO-19
math/commonsense reasoning	code-davinci-002	self-consistency	SVAMP, StrategyQA

code generation and machine translation



thank you

https://github.com/deep-spin/reranking-laws