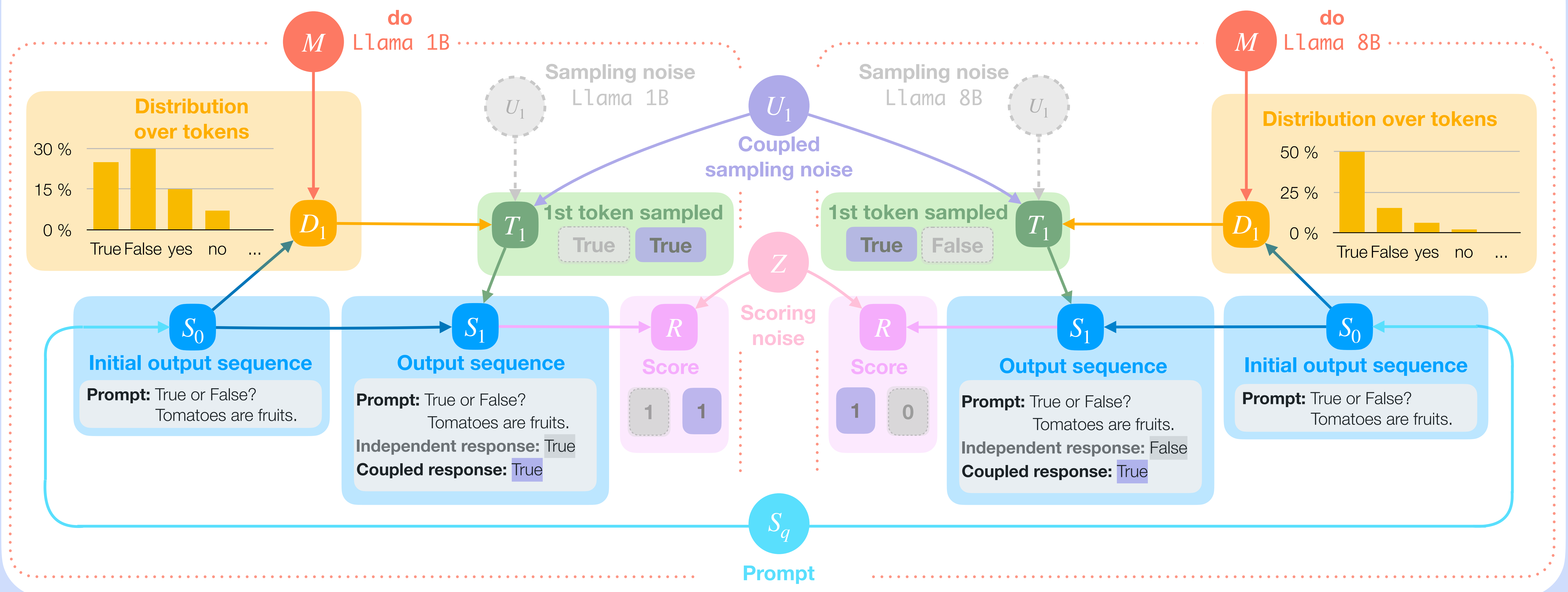
Evaluation of Large Language Models via Coupled Token Generation

Nina Corvelo Benz, Stratis Tsirtsis, Eleni Straitouri, Ivi Chatzi, Ander Artola Velasco,
Suhas Thejaswi, and Manuel Gomez-Rodriguez

MAX PLANCK INSTITUTE FOR SOFTWARE SYSTEMS

ETH zürich

A causal view of LLM evaluation



Prompt: True or False? Tomatoes are fruits.

Prompt: True or False? Tomatoes are fruits.
Independent response: True
Coupled response: True

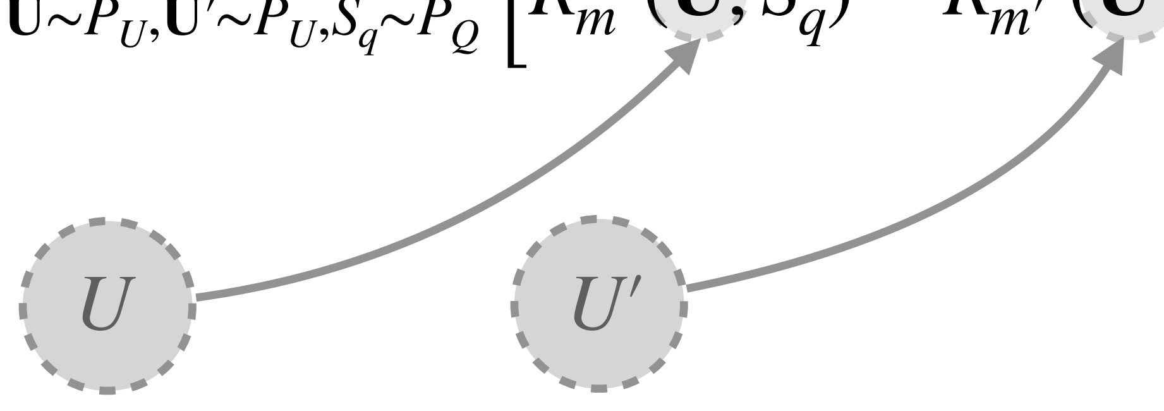
Prompt: True or False? Tomatoes are fruits.
Independent response: False
Coupled response: True

Prompt: True or False? Tomatoes are fruits.

Evaluation based on benchmark datasets

Independent evaluation reduces to estimating

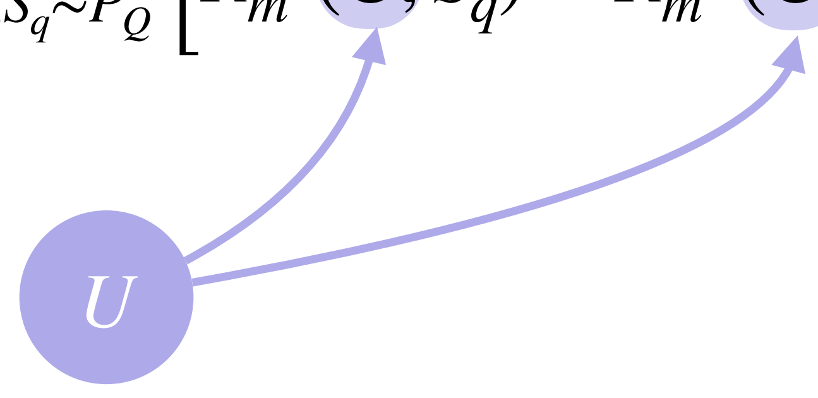
$$\mathbb{E}_{\mathbf{U}\sim P_U, \mathbf{U}'\sim P_U, S_q\sim P_Q}\left[R_m(\mathbf{U}, S_q) - R_{m'}(\mathbf{U}', S_q)\right]$$



Independent noise values

Coupled evaluation reduces to estimating

$$\mathbb{E}_{\mathbf{U}\sim P_U, S_q\sim P_Q}\left[R_m(\mathbf{U}, S_q) - R_{m'}(\mathbf{U}, S_q)\right]$$



Coupled noise values

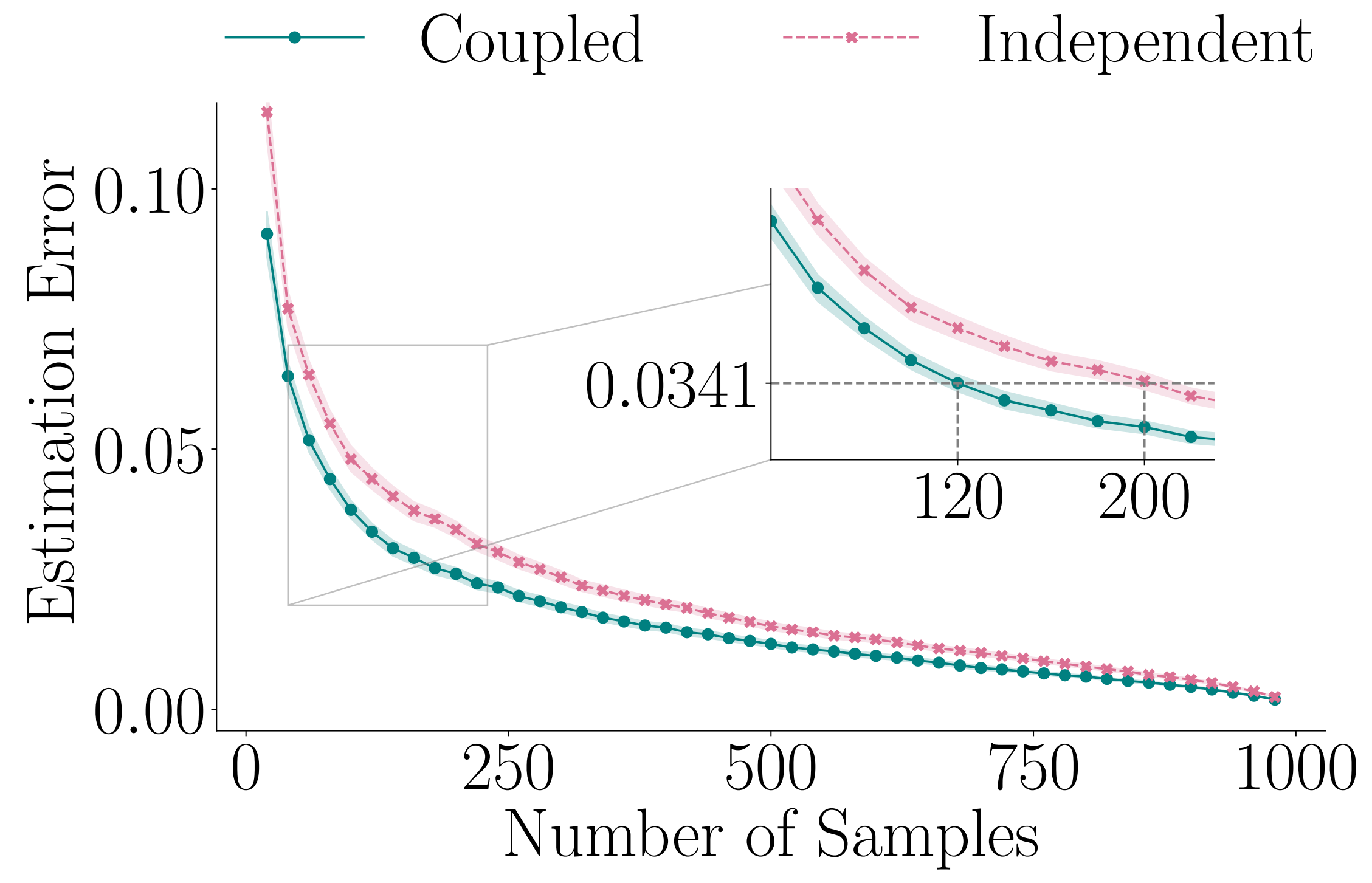
Proposition

For any pair of LLMs $m, m' \in \mathcal{M}$, it holds that

$$\mathrm{Var}[R_m(\boldsymbol{U}, S_q) - R_{m'}(\boldsymbol{U}', S_q)] = \mathrm{Var}[R_m(\boldsymbol{U}, S_q) - R_{m'}(\boldsymbol{U}, S_q)] + 2\cdot\mathrm{Cov}[R_m(\boldsymbol{U}, S_q), R_{m'}(\boldsymbol{U}, S_q)]$$

Theory: Covariance is positive when models have similar token distributions

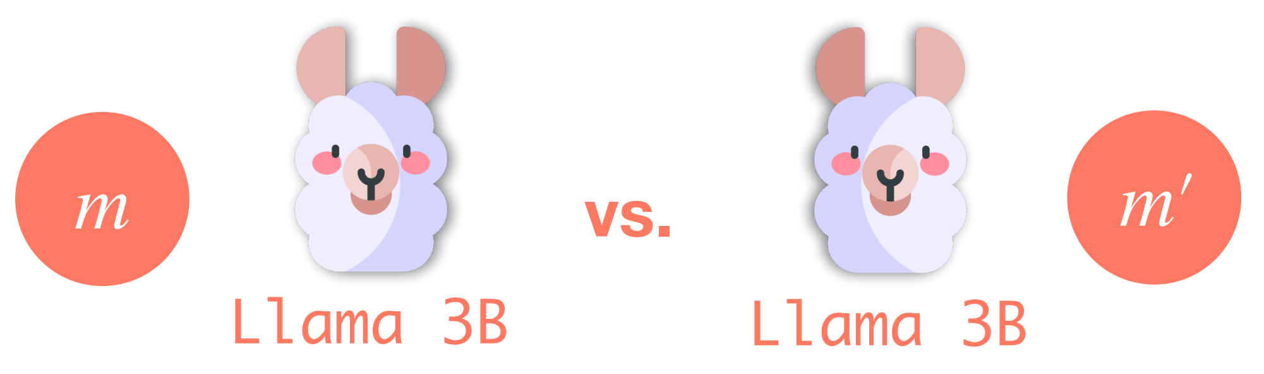
Experiments: Coupled token generation reduces the number of samples needed for evaluation



Llama 3.1 1B vs **3B** on **MMLU** dataset

Evaluation based on pairwise comparisons

Example: Comparing identical models, we expect that m and m' will be tied



Independent evaluation reduces to estimating

$$\mathbb{E}_{\mathbf{U}\sim P_U, \mathbf{U}'\sim P_U, S_q\sim P_Q}\left[\mathbb{1}\left\{R_m(\mathbf{U}, S_q) > R_{m'}(\mathbf{U}', S_q)\right\}\right]$$

Independent noise values – outputs may differ

Model m may win over m' and vice versa

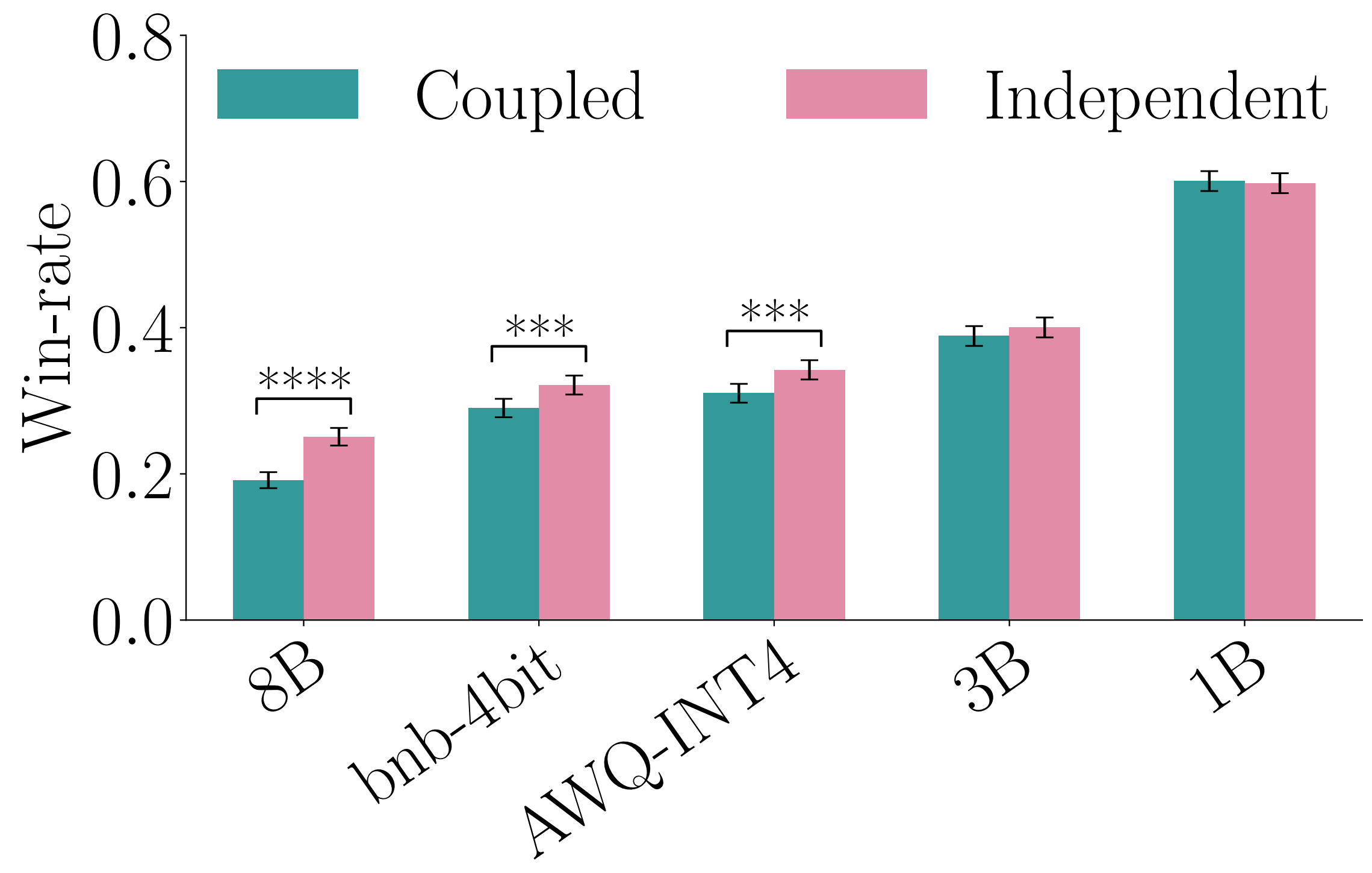
Coupled evaluation reduces to estimating

$$\mathbb{E}_{\mathbf{U}\sim P_U, S_q\sim P_Q}\left[\mathbb{1}\left\{R_m(\mathbf{U}, S_q) > R_{m'}(\mathbf{U}, S_q)\right\}\right]$$

Coupled noise values – outputs always identical

Models m and m' are always tied

Experiments: Coupled token generation leads to lower win rates (due to ties) and different rankings



Llama bnb-8bit vs other **Llama** models on **LMSYS-Chat-1M** dataset

	Coupled		Independent	
LLM	Rank	Avg. win-rate	Rank	Avg. win-rate
8B	1	0.3670 ±0.0020	1	0.3863 ±0.0020
bnb-8bit	2	0.3562 ±0.0020	1	0.3825 ±0.0020
bnb-4bit	3	0.3339 ±0.0020	3	0.3463 ±0.0020
AWQ-INT4	4	0.3164 ±0.0019	4	0.3310 ±0.0019
3B	5	0.2787 ±0.0019	5	0.2828 ±0.0019
1B	6	0.1650 ±0.0015	6	0.1664 ±0.0015

Ranking **Llama** models on **LMSYS-Chat-1M** dataset