STAY WITHIN YOUR BOUNDS: DISTANCE-GUIDED DECODING FOR GUARANTEED CONTEXT-FREE GRAMMAR COMPLIANCE

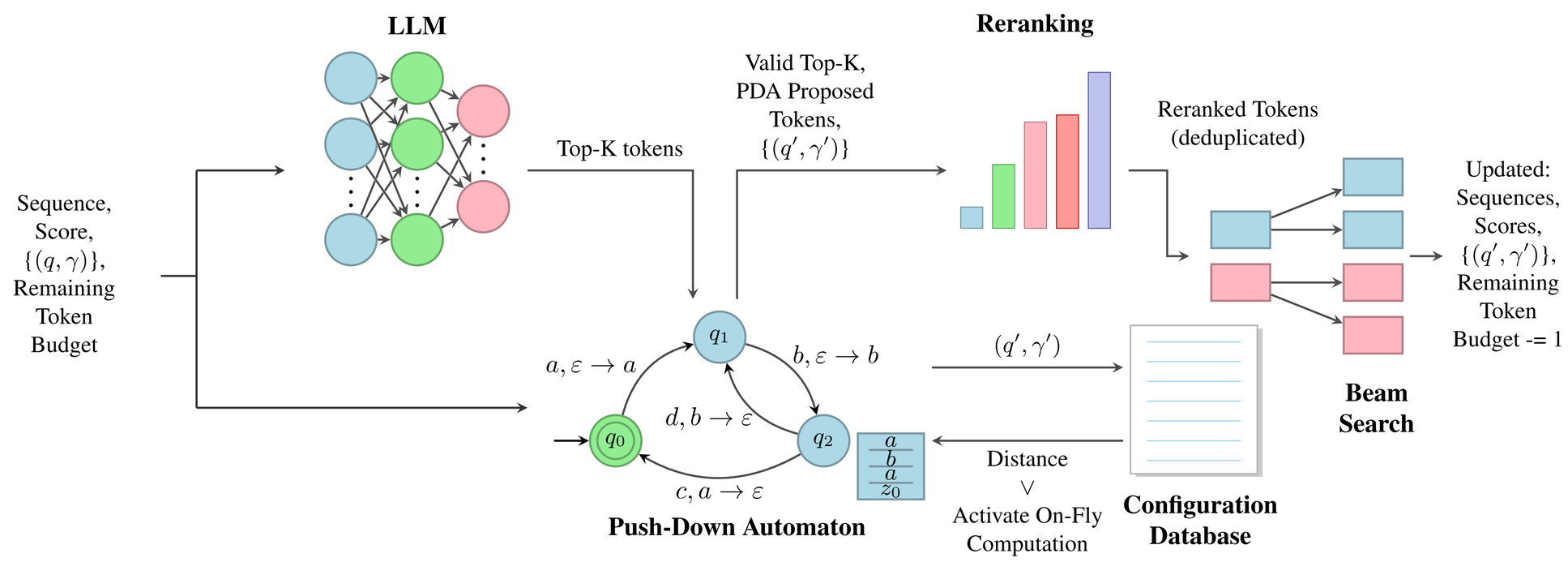
Vincenzo Collura, Karim Tit, Eleonora Giunchiglia, Mike Papadakis, Maxime Cordy – SnT SerVal Research Group, Imperial College London – Vincenzo.collura@uni.lu

INTRODUCTION

Grammar-constrained decoding helps large language models produce syntactically valid structured outputs. For context-free grammars, many practical decoders enforce local prefix feasibility: each token must keep the current prefix extendable to some valid completion. Yet, under tokenizer–grammar mismatch and finite token budgets, feasible prefixes may still fail to reach acceptance.

METHOD

We propose **SWYB** a **lookahead-guided decoding method** for context-free grammars based on pushdown automata. Offline, we compute bounded **pushdown summaries** with **reachability labels** and **upper-bound distances to acceptance**. Online, these estimates **guide horizon-aware pruning** and **beam search**. The resulting decoder is **syntactically sound**: every output is accepted by the target grammar.



RESULTS

Experiments on JSON, SQL, and Linear Temporal Logic (LTL) show both **perfect syntactic validity** (as guaranteed by our theorem) and **improved quality** over existing baselines.

JSON

Method	Llama-3.1-8B-Instruct				Llama-3.2-3B-Instruct				Qwen2.5-7B-Instruct			
	Syn.	Schema	Time	Perp.	Syn.	Schema	Time	Perp.	Syn.	Schema	Time	Perp.
Base LM	63.7±15.9	57.1±14.4	1.74±0.0	1.05	56.4±9.8	49.4±8.7	**0.96**±0.1	**1.06**	93.6±0.8	87.4±0.8	**1.47**±0.0	**1.01**
Sample-Verify	92.0±0.0	80.9±0.9	17.38±0.0	1.08	80.0±0.0	67.1±1.9	9.65±0.0	1.22	96.0±0.0	88.3±0.5	14.71±0.0	1.04
Guidance	96.0	89.0	2.88	1.08	93.0	84.0	2.76	1.12	92.0	87.0	2.85	1.02
Outlines	94.0	90.0	3.49	1.07	93.0	90.0	2.04	1.12	95.0	89.0	3.62	1.02
XGrammar	93.0	85.0	**1.57**	1.45	95.0	88.0	**0.96**	1.38	93.0	85.0	1.60	1.80
SynCode	88.0	79.0	3.05	**1.02**	88.0	77.0	2.12	1.13	94.0	88.0	1.50	**1.01**
GenLM (NP=2)	90.8±1.9	86.6±1.8	3.77±1.9	1.05	91.4±1.3	87.5±1.4	3.93±1.1	1.09	93.6±0.5	89.0±0.0	2.93±1.9	**1.01**
GenLM (NP=5)	91.6±1.1	87.2±1.3	5.59±0.3	1.03	93.0±1.0	88.4±0.9	7.36±0.4	1.09	93.8±0.4	89.0±0.0	4.07±0.2	**1.01**
▷ SWYB	**100.0**	**100.0**	3.96	1.13	**100.0**	**100.0**	3.46	1.14	**100.0**	**100.0**	4.05	1.11

SQL

Method	Llama-3.1-8B-Instruct				Llama-3.2-3B-Instruct				Qwen2.5-7B-Instruct			
	Ex.	Syn.	Time	Perp.	Ex.	Syn.	Time	Perp.	Ex.	Syn.	Time	Perp.
Base LM	53.8±4.7	92.9±7.9	**0.87**±0.0	1.61	46.3±4.0	96.2±6.7	**0.53**±0.0	2.16	61.4±0.6	99.2±0.2	**0.88**±0.0	3.48
Sample-Verify	57.5±1.1	**100.0**±0.0	8.90±0.0	1.56	47.8±0.7	99.1±0.3	5.29±0.0	3.50	61.5±0.4	99.7±0.1	8.83±0.0	10.38
SynCode	43.2	90.7	3.31	3.70	39.6	98.1	2.01	2.47	43.2	90.7	3.31	3.70
GenLM (NP=5,5,4)	59.2±0.9	99.9±0.1	10.74±8.4	1.67	49.4±0.7	99.9±0.1	4.17±0.3	**1.38**	51.9±0.4	86.1±0.6	6.77±0.1	3.78
▷ SWYB	**61.1**	**100.0**	9.92	**1.22**	**51.5**	**100.0**	6.25	1.43	**62.8**	**100.0**	13.03	**2.10**

LTL

Method	Llama-3.1-8B-Instruct				Llama-3.2-3B-Instruct				Qwen2.5-7B-Instruct			
	Acc.	Syn.	Time	Perp.	Acc.	Syn.	Time	Perp.	Acc.	Syn.	Time	Perp.
Base LM	22.1±3.3	98.8±0.4	**0.29**±0.0	1.31	9.6±3.3	94.6±3.1	**0.17**±0.0	1.57	30.9±0.5	98.4±0.4	**0.29**±0.0	1.61
Sample-Verify	22.2±0.3	**100.0**±0.0	2.86±0.0	1.31	9.3±0.2	99.9±0.0	1.64±0.0	1.56	30.6±0.2	99.7±0.0	2.89±0.0	**1.07**
SynCode	24.4	99.9	0.30	**1.21**	15.2	99.9	0.19	1.36	31.0	**100.0**	0.31	1.09
▷ SWYB	**29.6**	**100.0**	1.69	1.23	**22.1**	**100.0**	0.78	**1.27**	**31.6**	**100.0**	1.53	1.09

BUDGET SATURATION PROBLEM

The limited performance of existing grammar-constrained decoding methods is primarily caused by their reliance on **local prefix feasibility**.

- Increasing the **generation budget** does not materially improve constraint satisfaction, indicating that failures are **not primarily due to insufficient generation length**.
- The decoder can repeatedly select locally valid actions, such as **opening nested objects or arrays**, without receiving a signal that guides it toward an **accepting configuration**.
- Larger token budget may **substantially increase decoding time**.

Method	Llama-3.1-8B-Instruct				Llama-3.2-3B-Instruct				Qwen2.5-7B-Instruct			
	Syn.	Schema	Time	Perp.	Syn.	Schema	Time	Perp.	Syn.	Schema	Time	Perp.
Guidance	98.0	90.0	3.62	1.07	98.0	85.0	3.12	1.10	98.0	91.0	3.61	**1.01**
Outlines	99.0	**91.0**	3.70	1.07	95.0	92.0	1.93	1.09	**100.0**	**93.0**	3.66	1.02
SynCode	**100.0**	83.0	26.89	**1.02**	**99.0**	79.0	15.14	1.11	**100.0**	92.0	**1.60**	**1.01**
GenLM (NP=2)	96.3±0.7	89.8±0.9	5.4±1.8	1.04	96.1±0.9	90.1±1.3	4.8±1.2	**1.08**	98.0±0.2	92.9±0.6	2.6±1.4	**1.01**
GenLM (NP=5)	96.9±1.1	90.8±1.1	6.8±2.3	1.04	97.0±0.9	90.4±1.2	8.2±1.3	**1.08**	98.0±0.2	92.8±0.4	4.2±0.1	**1.01**

Method	Llama-3.1-8B-Instruct				Llama-3.2-3B-Instruct				Qwen2.5-7B-Instruct			
	Syn.	Schema	Time	Perp.	Syn.	Schema	Time	Perp.	Syn.	Schema	Time	Perp.
Guidance	98.0	90.0	**3.62**	1.07	98.0	85.0	3.13	1.10	98.0	91.0	3.48	**1.01**
Outlines	99.0	**91.0**	3.69	1.07	95.0	**92.0**	**1.94**	1.09	99.0	91.0	3.69	1.09
SynCode	**100.0**	83.0	55.84	**1.02**	**99.0**	80.0	31.65	1.11	**100.0**	92.0	**1.61**	**1.01**
GenLM (NP=2)	96.7±0.7	89.9±1.7	6.8±3.0	1.04	96.0±1.2	89.6±1.3	7.7±2.8	**1.08**	98.0±0.0	92.7±0.7	2.5±1.5	**1.01**
GenLM (NP=5)	96.5±0.5	90.2±0.8	10.3±5.9	1.04	97.1±0.7	90.2±1.5	12.7±3.9	**1.08**	98.0±0.0	**92.8**±0.6	4.0±0.2	**1.01**

Method	Llama-3.1-8B-Instruct				Llama-3.2-3B-Instruct				Qwen2.5-7B-Instruct			
	Syn.	Schema	Time	Perp.	Syn.	Schema	Time	Perp.	Syn.	Schema	Time	Perp.
Guidance	98.0	90.0	**3.57**	1.07	98.0	85.0	3.08	1.11	98.0	91.0	3.46	**1.01**
Outlines	99.0	**91.0**	3.68	1.07	95.0	**92.0**	**1.91**	1.09	100.0	**93.0**	3.60	1.02
SynCode	**100.0**	83.0	113.81	**1.02**	**99.0**	79.0	63.47	1.12	**100.0**	92.0	**1.60**	**1.01**
GenLM (NP=2)	96.4±0.5	**89.9**±0.9	22.0±8.9	1.04	95.5±1.3	89.8±1.5	20.4±15.2	**1.08**	98.0±0.0	92.6±0.5	2.4±1.5	**1.01**
GenLM (NP=5)	96.3±0.5	89.8±1.0	26.3±19.1	1.04	96.7±1.1	**90.5**±0.8	29.0±16.7	**1.08**	98.0±0.0	92.8±0.4	4.2±0.2	**1.01**

JSON results with token budget of 1024, 2048 and 4096, respectively.

CONCLUSIONS

- We introduce SWYB, a lookahead-guided decoder for context-free grammars that formally **guarantees full syntactic validity**, **overcoming the limits of local prefix feasibility**.
- By leveraging pushdown automata and precomputed distances to acceptance, the decoder steers generation toward accepting configurations without relying on larger token budgets.
- Results on JSON, SQL, and LTL confirm perfect syntactic validity alongside improved quality over existing baselines.