A Survey on LLM Inference-Time Self-Improvement

Xiangjue Dong,* Maria Teleki,* James Caverlee

NeuroLogic* (Lu et al., 2022)

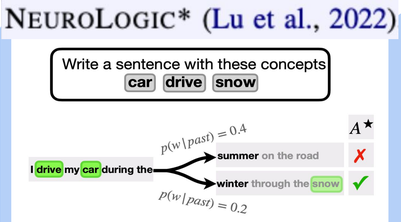


Figure 1: NeuroLogic★ leverages *lookahead heuristics* to guide generations towards those that satisfy the given task-specific constraints. In this example from the CommonGen task, although **summer** is a *more likely* next word given the already-generated *past*, NeuroLogic★ looks ahead to see that selecting **winter** results in a generation that incorporates unsatisfied constraint **snow** with a higher probability later on. Thus, **winter** is preferred despite being lower probability than **summer**.

Penalty Decoding (Zhu et al., 2023)

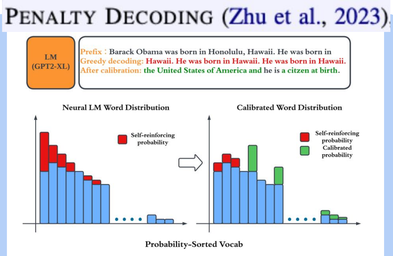


Figure 1: The predicted distribution of a neural language model (LM) can be regarded as a reinforced version of the model itself, as illustrated in the left part of the figure. To ensure high-quality text generation, penalties can be applied to the reinforced tokens, thereby correcting the distribution and improving the generated output.

DoLa (Chuang et al., 2024)

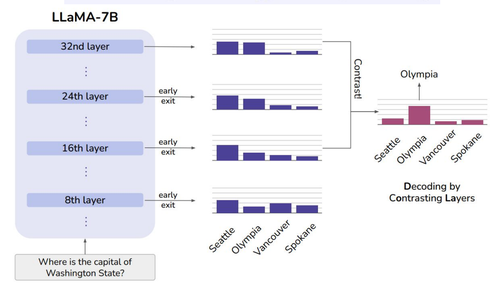
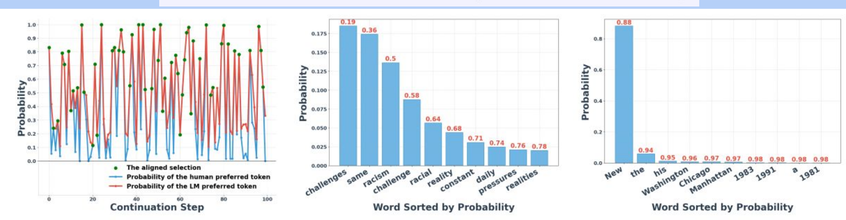


Figure 1: Illustration of an LLM progressively incorporates factual information along layers. While the next-word probabilities of "*Seattle*" remain similar throughout different layers, the probabilities of the correct answer "*Olympia*" gradually increase from lower to higher layers. DoLa uses this fact to decode by contrasting the difference between layers to sharpen an LLM's probability towards factually correct outputs.

Adaptive Decoding (Zhu et al., 2024)



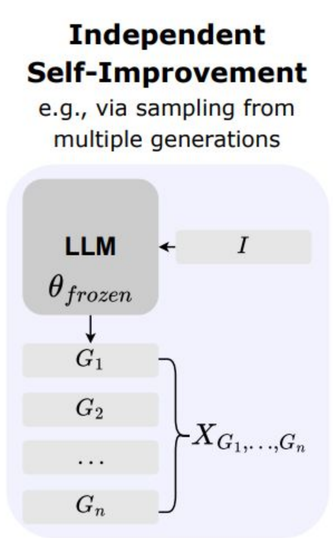
(a) The probability of human-preferred token and LM-preferred token.

(b) Head distribution given the **prefix**: *"Growing up, Obama faced the challenges"*

(c) Head distribution given the **prefix**: *"attended Columbia University in New"*

Figure 1. Human written text: "*Barack Obama was born in 1961. He was raised in Hawaii by his mother and grandparents. Growing up, Obama faced the challenges of being biracial, with a Kenyan father and an American mother. Despite these challenges, he excelled academically and eventually attended Columbia University in New York City.*" We provide this human-written text for GPT2-XL and use teacher-forcing decoding.

Taxonomy of ITI Methods



- **Large Language Model Inference-Time Self-Improvement**
  - **Independent Self-Improvement (§2)**
    - Constrained Decoding (§2.1)
      - Hard Constraint: NeuroLogic* (Lu et al., 2022), Control-DAG (Chen et al., 2024a)
      - Soft Constraint: Penalty Decoding (Zhu et al., 2023), IPS(Yao et al., 2023)
    - Contrastive Decoding (§2.2)
      - Hallucinations: PMI-Decode (Nandwani et al., 2023), LCD (Manevich and Tsarfaty, 2024), Anti-LM (Sia et al., 2024), DoLa (Chuang et al., 2024)
      - Repetition: Look-back (Xu et al., 2023), Adaptive Decoding (Zhu et al., 2024)
    - Minimum Bayes-Risk Decoding (§2.3)
      - Clustering: DMBR & KMBR (Jinnai et al., 2024a), CBMBR (Deguchi et al., 2024)
      - Matrix Approx.: PMBR (Trabelsi et al., 2024)
      - Others: Pruning MBR (Cheng and Vlachos, 2023), AMBR (Jinnai and Ariu, 2024), MBMBR (Jinnai et al., 2024b)
    - Parallel Decoding (§2.4): PJ, PGJ & HGJ (Santilli et al., 2023), Arithmetic Sampling (Vilnis et al., 2023), Lookahead Decoding (Fu et al., 2024), SoT (Ning et al., 2024)
    - Sampling-based Decoding (§2.5)
      - Generation: BAT (Finlayson et al., 2024), DAEMON (Ji et al., 2024)
      - Reasoning: Self-Consistency (Wang et al., 2023c), ESC (Li et al., 2024b)
      - Others: ASAp (Park et al., 2024)
    - Tree-Search-based Decoding (§2.6): PG-TD (Zhang et al., 2023), GDP-Zero (Yu et al., 2023), RAP (Hao et al., 2023)
    - Model-level Decoding (§2.7): ACD (Gera et al., 2023), Self-SD (Zhang et al., 2024c), SLED (Zhang et al., 2024b), Language-Specific Neurons (Kojima et al., 2024), False Induction Heads (Halawi et al., 2024)
  - **Context-Aware Self-Improvement (§3)**
    - Prompting (§3.1)
      - Reasoning: CoT Prompting (Wei et al., 2022), Zero-Shot CoT (Kojima et al., 2022), Evaluation (Shaikh et al., 2023), EchoPrompt (Mekala et al., 2024)
      - Others: DecodingTrust (Wang et al., 2023a), URIAL (Lin et al., 2024), Generation Exploitation Attack (Huang et al., 2024)
    - Disturbed Prompt (§3.2): ICD (Wang et al., 2024b), ID (Kim et al., 2024), ROSE (Zhong et al., 2024), CAD (Shi et al., 2024b), COIECD (Yuan et al., 2024b)
    - Retrieval-based (§3.3): kNN-LM (Khandelwal et al., 2020), NEST (Li et al., 2024a), REST (He et al., 2024), RTD (Luohe et al., 2024), Multi-Input CD (Zhao et al., 2024)
  - **Model-Aided Self-Improvement (§4)**
    - Expert and/or Anti-Expert (§4.1)
      - Toxicity: DExperts (Liu et al., 2021), MIL-Decoding (Zhang and Wan, 2023)
      - Machine Translation: PSGD (Wang et al., 2023b), CoDec (Zeng et al., 2024), LiBS (Yang et al., 2024), CODEC (Le et al., 2024)
      - Alignment: MOD (Shi et al., 2024a), Transfer Q^* (Chakraborty et al., 2024)
      - Others: SafeDecoding (Xu et al., 2024), Superposed Decoding (Shen et al., 2024), GD (Huang et al., 2023), Equilibrium-Ranking (Jacob et al., 2024)
    - Draft Model (§4.2): Speculative Decoding (Leviathan et al., 2023), SpecTr (Sun et al., 2023), GSD (Gong et al., 2024), SpecExec (Svirschevski et al., 2024), SEQUOIA (Chen et al., 2024b), SCD (Yuan et al., 2024a), Theoretical Analysis (Yin et al., 2024), Online SD (Liu et al., 2024b), GliDe & CaPE (Du et al., 2024)
    - Small LM/ Amateur LM (§4.3)
      - Classification: NeuroLogic-A* (P) (Miyano et al., 2023), EnDec (Zhang et al., 2024a), Critic-Driven Decoding (Lango and Dusek, 2023), KCTS (Choi et al., 2023)
      - Generative: CD (Li et al., 2023), BiLD (Kim et al., 2023), JAMDEC (Fisher et al., 2024)
    - Reward Model (§4.4): RAD (Deng and Raffel, 2023), ARGS (Khanov et al., 2024), TS-LLM (Wan et al., 2024), Controlled Decoding (Mudgal et al., 2024)
    - Tool Use/API (§4.5): GCD (Geng et al., 2023), NeuroStructural Decoding (Bastan et al., 2023), MGD (Agrawal et al., 2023), FLAP (Roy et al., 2024)

CBMBR (Deguchi et al., 2024)

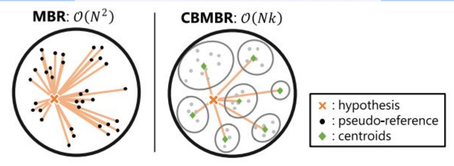


Figure 1: Overview of our centroid-based MBR (CBMBR).

DAEMON (Ji et al., 2024)

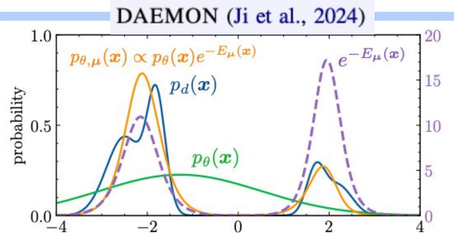


Figure 1: The decoding distribution $p_{\theta,\mu}$ induced by DAEMON scales the input LM distribution p_θ with a sequence-level energy function E_μ, which leads to a more accurate recovery of the underlying data distribution p_d.

DExperts (Liu et al., 2021)

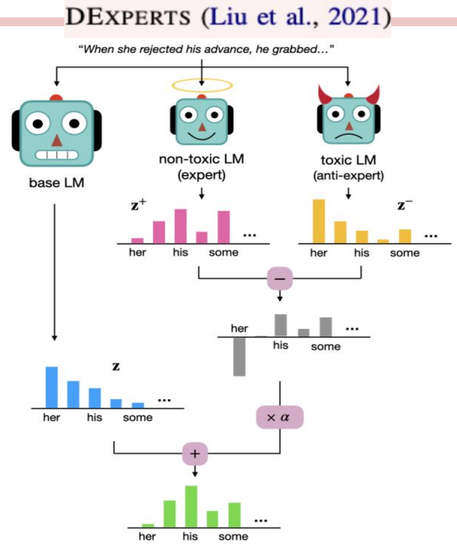


Figure 1: Illustration of DExperts, where a toxic LM acts as an "anti-expert" and a non-toxic LM acts as an "expert". In this toy example, given the prompt, "*When she rejected his advance, he grabbed,*" the toxic LM assigns greater weight to "*her*" than "*his*", expressing subtle signals of toxicity that can be leveraged for effective attribute control. The difference in logits $\mathbf{z}^+ - \mathbf{z}^-$ output by the expert and anti-expert represents the perturbations to make to the logits $\mathbf{z}$ of the pretrained "base" LM.

GCD (Geng et al., 2023)

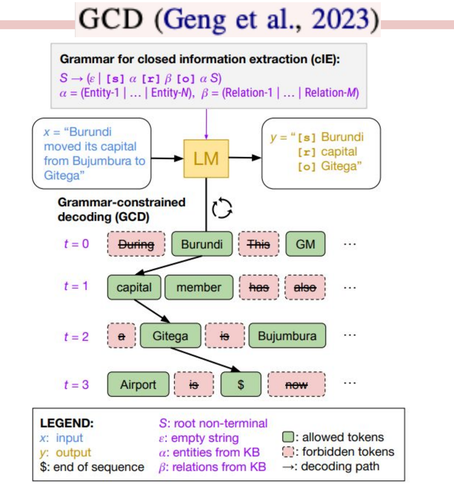


Figure 1: **Grammar-constrained decoding (GCD),** applied to the task of closed information extraction, where the goal is to extract a list y of subject–relation–object triplets from the input text x. Subjects and objects are constrained to be Wikidata entities, relations to be a Wikidata relation. During decoding, only valid token continuations compliant with the grammar are considered. For simplicity, we omit the special marker symbols [s], [r], and [o] in the schema of the generation process.

NeuroStructural Decoding (Bastan et al., 2023)

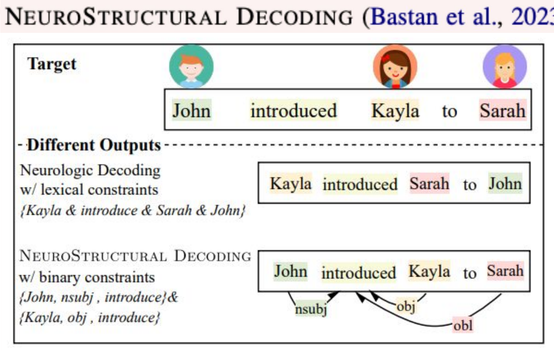


Figure 1: An example that compares the output produced by Neurologic Decoding with lexical constraints alone vs. the output generated by NeuroStructural Decoding with lexico-syntactic constraints.

FLAP (Roy et al., 2024)

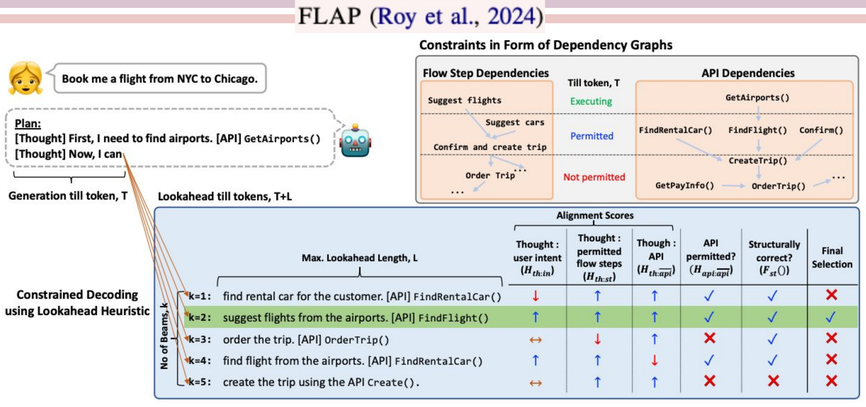


Figure 2: An example state of FLAP, our proposed lookahead heuristic-based constrained decoding algorithm for faithful planning, in the domain "Trip Planning" for a query related to "book flight". Here, ↑, ↓, and ↔ indicate high, low, and mediocre alignment scores, respectively. The selected path based on the heuristic scores is highlighted.

CAD (Shi et al., 2024b)

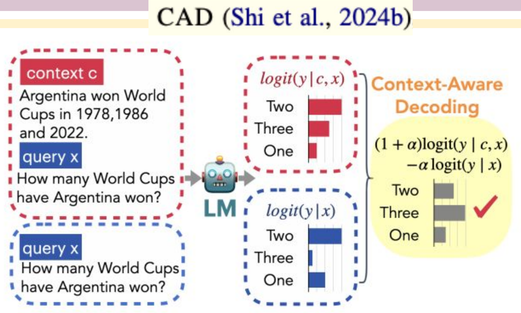


Figure 1: An illustration of context-aware decoding.

https://github.com/dongxiangjue/Awesome-LLM-Self-Improvement

Check out our GitHub!

Send us a pull request with any addtl methods!

Texas A&M University
Department of Computer Science & Engineering

MARIA TELEKI

XIANGJUE DONG