DEAKIN UNIVERSITY | DEAKIN APPLIED ARTIFICIAL INTELLIGENCE INITIATIVE

ECML PKDD

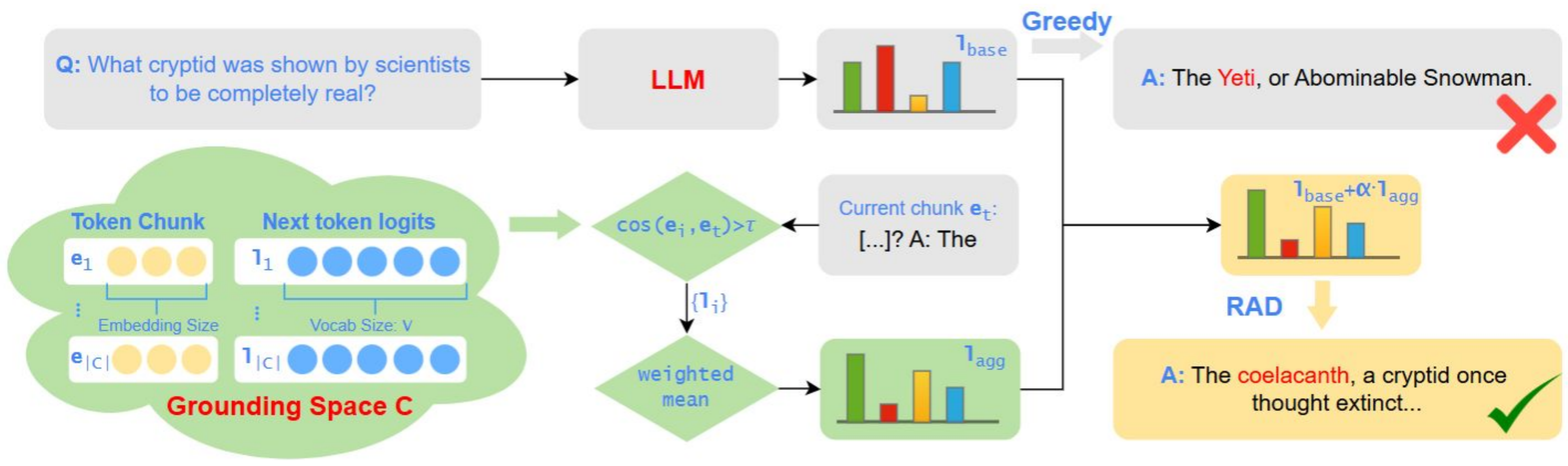
Retrieval-augmented Decoding for Improving Truthfulness in Open-ended Generation

Manh Nguyen, Sunil Gupta, Hung Le

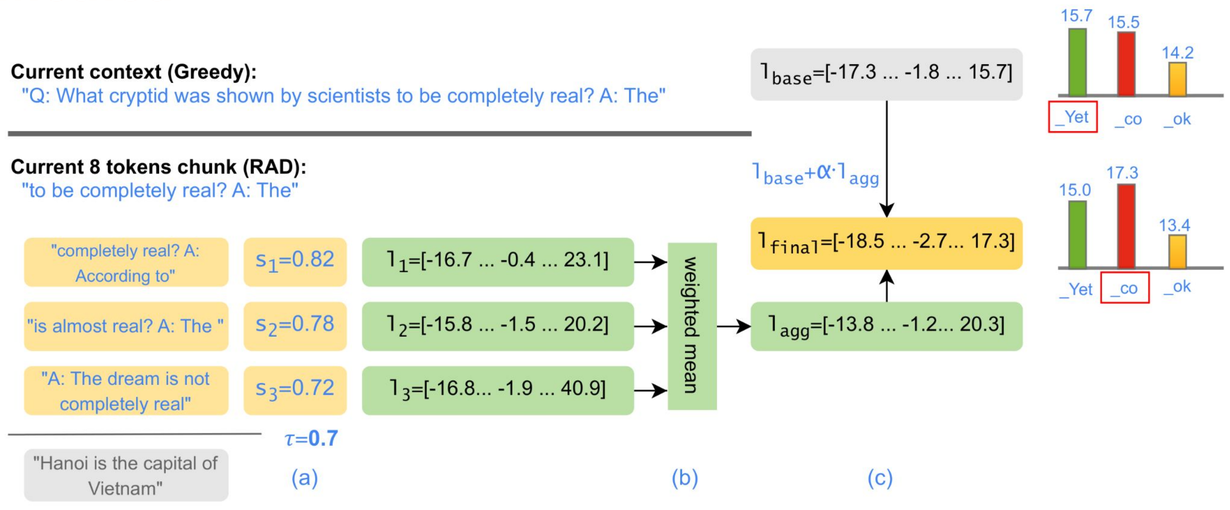
Core idea: Motivated by RAG, our method mitigates hallucination at decoding time through chunk-level retrieval and thresholding.

Abstract

- SFT/RLHF are resource-intensive, while prompting or RAG offer limited decoding control and robustness.
- We introduce Retrieval-Augmented Decoding (RAD), which uses ~10 examples to retrieve contexts and shape logits at inference.
- RAD outperforms strong baselines across benchmarks and LLMs, with robust cross-task generalization and improved factuality.



Method



Context chunking (more references, x30 than prompting) and **thresholding** (better contexts) help improve general decoding, OOD performance and data efficiency

Algorithm 1 Retrieval-Augmented Decoding (RAD)

Input: Current decoding step t, Current tokens $x_{1:t-1}$;
Grounding space $C = \{(\mathbf{e}_i, \mathbf{l}_i)\}$ with $i \in \{1, 2, \ldots, |C|\}$; Embedding model E;
Chunk size M, threshold τ, interpolation weight α
Output: Decoded token x_t

// **Stage 1: Context Retrieval from Grounding Space**
Compute current context embedding size M: $\mathbf{e}_t = E(x_{t-M:t-1})$
Compute cosine similarities from all contexts in C to $\mathbf{e}_t$: $s_i = \cos(\mathbf{e}_t, \mathbf{e}_i)$
Select relevant contexts with threshold τ: $\mathcal{S} = \{i \mid s_i > \tau\}$
// **Stage 2: Aggregation of Retrieved Signals**
Aggregate logits:

$$\mathbf{l}_t^{\text{agg}} = \sum_{n \in \mathcal{S}} \frac{s_i}{\sum_{j \in \mathcal{S}} s_j} \cdot \mathbf{l}_i$$

// **Stage 3: Truthfulness-Aware Logit Integration**
Compute model logits: $\mathbf{l}_t^{\text{base}} = \textsc{ModelLogits}(x_{<t})$
Combine model and aggregated logits:

$$\mathbf{l}_t^{\text{final}} = \mathbf{l}_t^{\text{base}} + \alpha \cdot \mathbf{l}_t^{\text{agg}}$$

Greedy decoding: $x_t = \arg\max\left(\mathbf{l}_t^{\text{final}}[x]\right)$
Return x_t

Experiments

- **Judges**: Cohere and Gemini APIs
- **Baselines**:
  - CAD, DoLa, ID: decoding methods by contrasting logits
  - KATE: retrieval-augmented prompting
  - kNN-LM: RAD without chunking and thresholding
- **#Samples**: 100 (KATE, kNN-LM, and RAD)

Method	TruthfulQA			WikiQA			Alpaca		
	%Truth	%Info	T*I	%Truth	%Info	T*I	%Truth	%Info	T*I
Qwen2.5-3B									
Greedy	49.6	82.7	41.1	62.2	81.0	50.4	46.6	79.9	37.2
CAD	45.6 (-4.0)	81.1 (-1.6)	36.9 (-4.2)	59.6 (-2.6)	78.2 (-2.8)	46.6 (-3.8)	45.6 (-1.0)	76.0 (-3.9)	34.7 (-2.5)
DoLa	49.6 (+0.0)	82.7 (+0.0)	41.1 (+0.0)	62.1 (-0.1)	80.6 (-0.4)	50.0 (-0.4)	48.1 (+1.5)	79.3 (-0.6)	38.1 (+0.9)
ID	47.2 (-2.4)	79.4 (-3.3)	37.5 (-3.6)	60.5 (-1.7)	81.8 (+0.8)	49.5 (-0.9)	43.9 (-2.7)	77.3 (-2.6)	33.9 (-3.3)
KATE	47.5 (-2.1)	78.7 (-4.0)	37.3 (-3.8)	55.6 (-6.6)	**82.5 (+1.5)**	45.9 (-4.5)	**49.3 (+2.7)**	**80.3 (+0.4)**	**39.6 (+2.4)**
*k*NN-LM	49.6 (+0.0)	81.8 (-0.9)	40.6 (-0.5)	62.6 (+0.4)	80.9 (-0.1)	50.6 (+0.2)	48.1 (+1.5)	78.5 (-1.4)	37.7 (+0.5)
RAD	**52.0 (+2.4)**	**83.2 (+0.5)**	**43.3 (+2.2)**	**63.2 (+1.0)**	81.5 (+0.5)	**51.5 (+1.1)**	48.4 (+1.8)	78.9 (-1.0)	38.2 (+1.0)
Qwen2.5-7B									
Greedy	58.8	89.0	52.3	76.9	89.3	68.7	61.1	92.3	56.4
CAD	56.4 (-2.4)	89.7 (+0.7)	50.5 (-1.8)	73.5 (-3.4)	88.9 (-0.4)	65.3 (-3.4)	58.3 (-2.8)	91.9 (-0.4)	53.6 (-2.8)
DoLa	59.7 (+0.9)	89.2 (+0.2)	53.3 (+1.0)	77.1 (+0.2)	89.3 (+0.0)	68.8 (+0.1)	61.2 (+0.1)	92.2 (-0.1)	56.4 (+0.0)
ID	**60.5 (+1.7)**	89.2 (+0.2)	54.0 (+1.7)	75.2 (-1.7)	88.6 (-0.7)	66.6 (-2.1)	60.5 (-0.6)	91.2 (-1.1)	55.2 (-1.2)
KATE	57.6 (-1.2)	85.9 (-3.1)	49.4 (-2.9)	76.5 (-0.4)	87.2 (-2.1)	66.7 (-2.0)	61.5 (+0.4)	92.8 (+0.5)	57.1 (+0.7)
*k*NN-LM	58.5 (-0.3)	89.2 (+0.2)	52.2 (-0.1)	77.6 (+0.7)	89.4 (+0.1)	69.4 (+0.7)	61.1 (+0.0)	91.3 (-1.0)	55.8 (+1.4)
RAD	**60.5 (+1.7)**	**90.2 (+1.2)**	**54.6 (+2.3)**	**77.7 (+0.8)**	**89.9 (+0.6)**	**69.9 (+1.2)**	**62.1 (+1.0)**	**93.0 (+0.7)**	**57.8 (+1.4)**
Mistral2-7B									
Greedy	60.7	91.8	55.7	76.6	93.7	71.8	63.6	92.4	58.8
CAD	60.4 (-0.3)	90.2 (-1.6)	54.5 (-1.2)	75.4 (-1.2)	91.0 (-2.7)	68.6 (-3.2)	62.0 (-1.6)	90.6 (-1.8)	56.2 (-2.6)
DoLa	60.9 (+0.2)	91.8 (+0.0)	55.9 (+0.2)	77.1 (+0.5)	93.4 (-0.3)	72.0 (+0.2)	**64.5 (+0.9)**	93.7 (+1.3)	**60.4 (+1.6)**
ID	61.6 (+0.9)	90.2 (-1.6)	55.6 (-0.1)	**77.6 (+1.0)**	91.3 (-2.4)	70.8 (-1.0)	63.1 (-0.5)	92.3 (-0.1)	58.2 (-0.6)
KATE	**63.5 (+2.8)**	89.2 (-2.6)	56.7 (+1.0)	73.5 (-3.1)	88.9 (-4.8)	65.3 (-6.5)	62.2 (-1.4)	**94.8 (+2.4)**	59.0 (+0.2)
*k*NN-LM	60.7 (+0.0)	91.8 (+0.0)	55.7 (+0.0)	75.8 (-0.8)	92.7 (-1.0)	70.3 (-1.5)	63.2 (-0.4)	92.1 (-0.3)	58.2 (-0.6)
RAD	62.8 (+2.1)	**92.1 (+0.3)**	**57.9 (+2.2)**	**77.6 (+1.0)**	**93.9 (+0.2)**	**72.9 (+1.1)**	64.1 (+0.5)	93.2 (+0.8)	59.7 (+0.9)
Gemma2-9B									
Greedy	64.3	90.6	58.3	83.9	94.5	79.2	67.3	91.9	61.9
CAD	66.2 (+1.9)	91.1 (+0.5)	60.3 (+2.0)	82.9 (-1.0)	93.0 (-1.5)	77.2 (-2.0)	**71.5 (+4.2)**	**93.7 (+1.8)**	**67.0 (+5.1)**
DoLa	65.0 (+0.7)	90.9 (+0.3)	59.1 (+0.8)	83.6 (-0.3)	94.0 (-0.5)	78.6 (-0.6)	67.4 (+0.1)	92.7 (+0.8)	62.5 (+0.6)
ID	67.6 (+3.3)	**91.4 (+0.8)**	61.8 (+3.5)	84.5 (+0.6)	94.9 (+0.4)	80.2 (+1.0)	71.2 (+3.9)	91.9 (+0.0)	65.4 (+3.5)
KATE	68.3 (+4.0)	**91.4 (+0.8)**	62.4 (+4.1)	81.7 (-2.2)	92.6 (-1.9)	75.6 (-3.6)	69.9 (+2.6)	92.1 (+0.2)	64.4 (+2.5)
*k*NN-LM	65.0 (+0.7)	91.1 (+0.5)	59.2 (+0.9)	**84.7 (+0.8)**	93.8 (-0.7)	79.5 (+0.3)	67.1 (-0.2)	92.8 (+0.9)	62.2 (+0.3)
RAD	**70.0 (+5.7)**	90.4 (-0.2)	**63.3 (+5.0)**	**84.7 (+0.8)**	**95.6 (+1.1)**	**81.0 (+1.8)**	70.2 (+2.9)	92.5 (+0.6)	64.9 (+3.0)

Table 3: Out-of-distribution performance (%) comparison across models and datasets (evaluation datasets are displayed). DoLa is shown for reference as it is the strongest non-data baseline. Best scores are bold, second-best underlined.

Model	Method	TruthfulQA			WikiQA		
		%Truth	%Info	T*I	%Truth	%Info	T*I
Qwen2.5-3B	DoLa	49.6	**82.7**	41.1	**62.1**	80.6	**50.0**
	KATE	48.0	76.8	36.9	56.4	**83.7**	47.2
	*k*NN-LM	50.4	**82.7**	41.7	60.4	80.7	48.7
	RAD (Ours)	**51.8**	81.5	**42.2**	60.2	80.4	48.4
Qwen2.5-7B	DoLa	**59.7**	89.2	**53.3**	77.1	89.3	68.8
	KATE	57.6	88.5	50.9	76.8	86.6	66.5
	*k*NN-LM	59.5	89.0	52.9	**77.6**	89.4	**69.4**
	RAD (Ours)	58.8	**89.7**	52.7	77.1	**90.0**	**69.4**
Mistral2-7B	DoLa	60.9	91.8	55.9	77.1	93.4	72.0
	KATE	53.4	82.5	46.5	**79.0**	91.0	71.9
	*k*NN-LM	60.4	91.6	55.4	76.5	92.7	70.9
	RAD (Ours)	**61.9**	**92.1**	**57.0**	78.0	**94.0**	**73.4**
Gemma2-9B	DoLa	65.0	**90.9**	59.1	83.6	94.0	78.6
	KATE	60.0	83.5	50.0	83.1	90.7	75.4
	*k*NN-LM	65.0	90.4	58.8	**84.7**	93.5	79.2
	RAD (Ours)	**68.5**	90.6	**62.1**	**84.7**	**94.6**	**80.1**

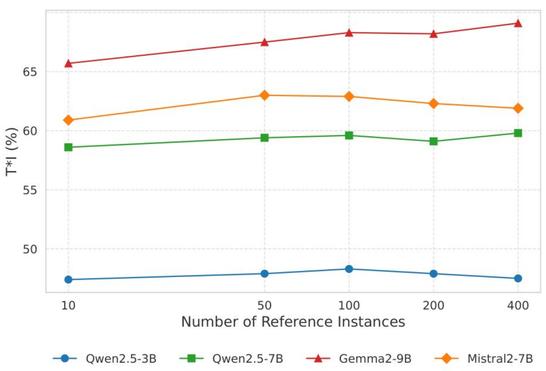


Fig. 3: Performance on TruthfulQA across grounding space sizes $|C|$.

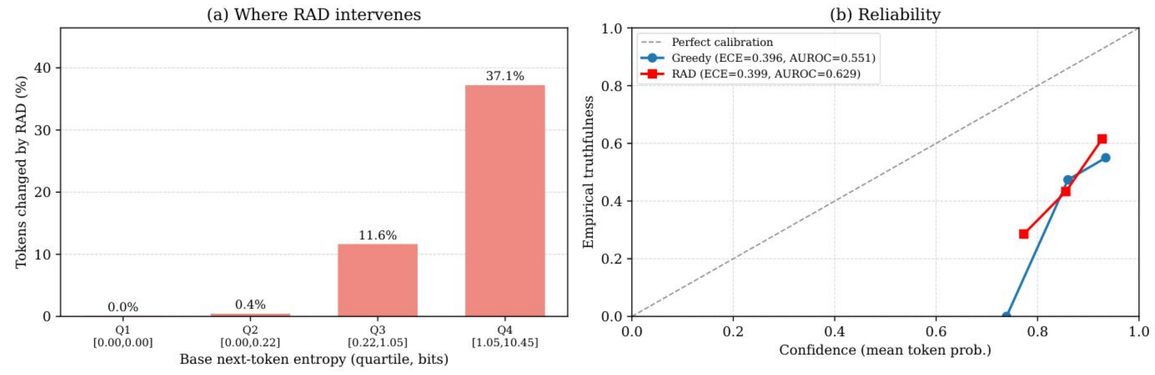


Fig. 5: Calibration on TruthfulQA (Qwen2.5-7B). **(a)** RAD edits concentrate at high base entropy; confident steps are preserved. **(b)** Reliability diagram: empirical truthfulness (fraction judged truthful) vs. response confidence (mean token probability). ECE measures calibration gap; AUROC measures how well confidence ranks truthful vs. untruthful responses.

References

- CAD (Shi et al. NAACL 2024)
- DoLa (Chung et al. ICLR 2024)
- ID (Kim et al. ICLR 2024)
- KATE (Liu et al. DeeLIO 2022)
- kNN-LM (Khandelwal et al. ICLR 2020)

Scan for the Code

manh.nguyen@deakin.edu.au

Extended Version: https://arxiv.org/abs/2508.02184