# Correction with Backtracking Reduces Hallucination in Summarization

**Zhenzhen Liu*** **Chao Wan** **Varsha Kishore** **Jin Peng Zhou**
Cornell University
{zl535, cw862, vk352, jz563}@cornell.edu

**Minmin Chen**
Google DeepMind
minminc@google.com

**Kilian Q. Weinberger**
Cornell University
kqw4@cornell.edu

Our method Correction with Backtracking (CoBa), is a simple inference-time method that requires no additional model training and is compatible with most of the decoding methods

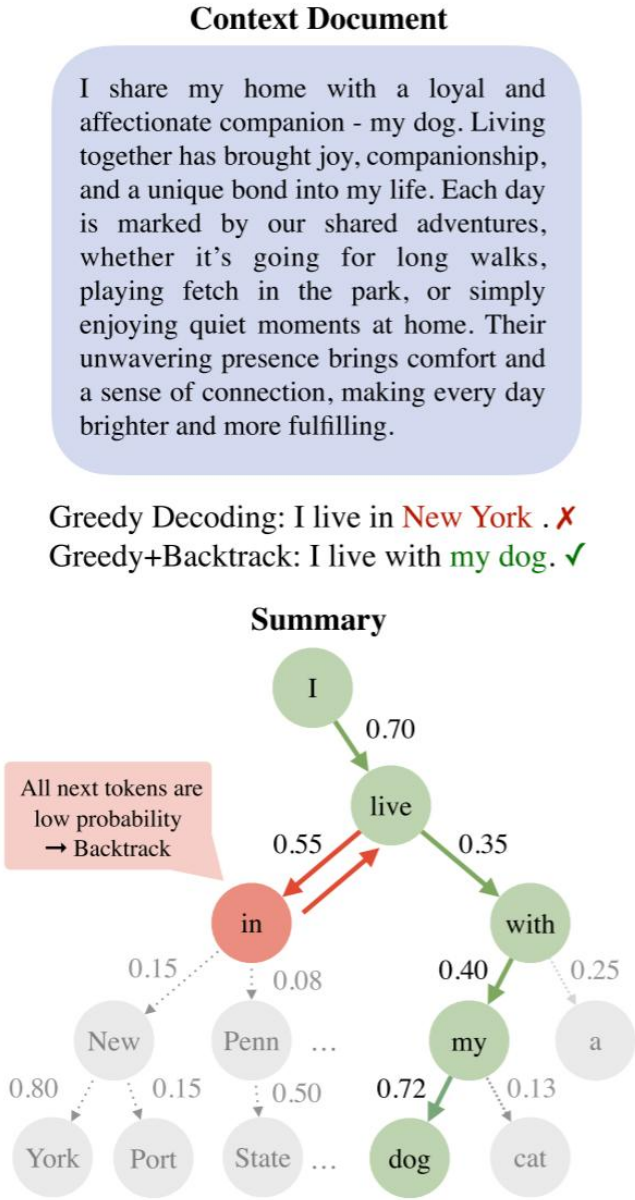
hallucinations

a phenomenon where models make statements that seem plausible but are not grounded in the source document

(1) the first word of a hallucinated sequence tends to have low conditional probability

(2) hallucinations are not supported by words in the context, and therefore have a large distance to context words.

Figure 1: Schematic illustration of CoBa (using only token probability as the detection metric with threshold 0.2). After the partial summary *"I live"*, the token *"in"* has a higher probability than "with". However, *"I live in"* will pressure the model into hallucinating a place. We detect this because all the next tokens have a probability lower than our threshold 0.2. Backtracking enables the model to find an alternative continuation that avoids hallucination down the line.



## 两种检测幻觉的方法

Let $\mathcal{M}_\theta$ be an autoregressive summarization model with parameters $\theta$, and let $\Sigma$ be its vocabulary. Given a context document $\mathcal{C} = (c_1, \cdots c_m)$ as input, $M_\theta$ produces a summary $\mathcal{S} = (s_1, \cdots, s_n)$:

$$\mathcal{M}_\theta(\mathcal{C}) = \mathcal{S}$$

Model $\mathcal{M}_\theta$ generates the summary autoregressively. At each step, given a partially generated summary $\mathcal{S}_{<t}$ up to token $s_{t-1}$, it outputs a distribution $p_\theta(s_t|\mathcal{C}, \mathcal{S}_{<t})$ for the next token $s_t$ over the vocabulary $\Sigma$. The probability of generating the summary $\mathcal{S}$ is thus

$$p(\mathcal{S}) = \prod_{t=1}^{|\mathcal{S}|} p_\theta(s_t|\mathcal{C}, \mathcal{S}_{<t})$$

we flag the token if the following condition holds:

$$p_\theta(s_t|\mathcal{C}, \mathcal{S}_{<t}) < \delta$$

where $\mathcal{C}$ is the context document, $\mathcal{S}_{<t}$ is the partially generated summary, and $\delta$ is the token level conditional probability threshold for hallucination.

The detection criterion in this case is:

$$d(v, \mathcal{C}) = \min_{c_i \in \mathcal{C}} cos\_dist\big(\text{Emb}(v), \text{Emb}(c_i)\big) > \varphi$$

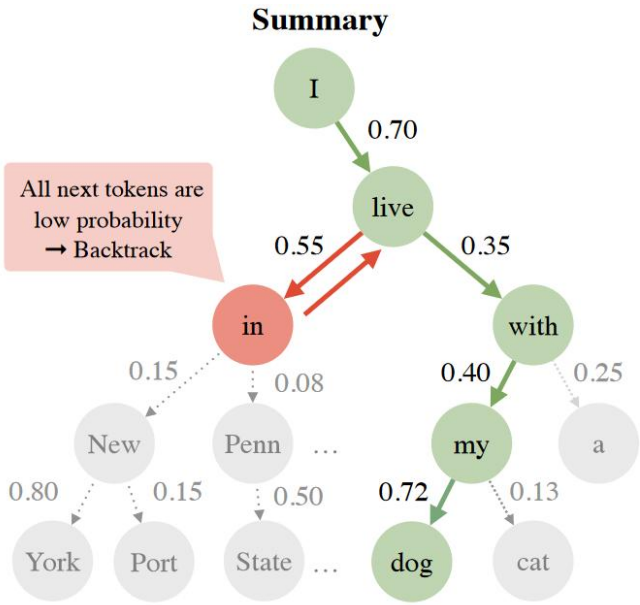
where $v$ is the proposed token, $\mathcal{C}$ is the context document and $\varphi$ is the distance threshold. Figure 2 presents the minimum token-to-context distance computed over the annotated dataset from Maynez et al. (2020b)'s with embeddings from Flan-t5 XL (the results are averaged over 5000 samples). The

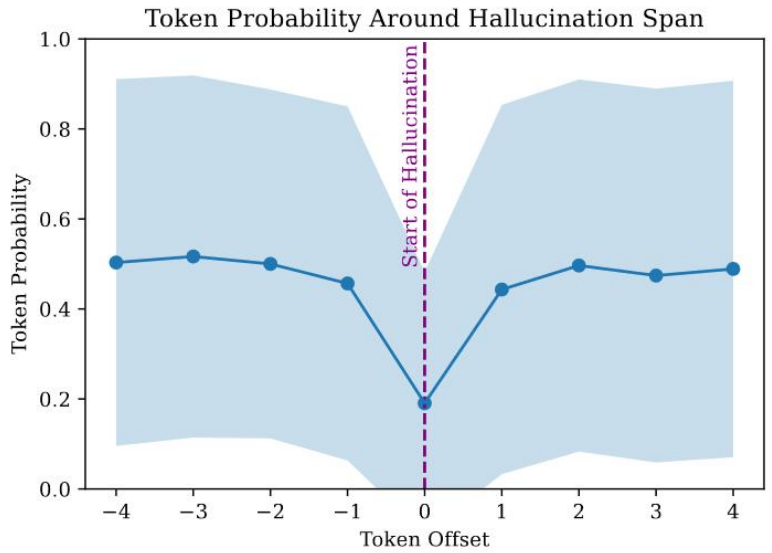
We eliminate the last generated token st and try to propose an alternative token s’t that does not satisfy the hallucination criteria.

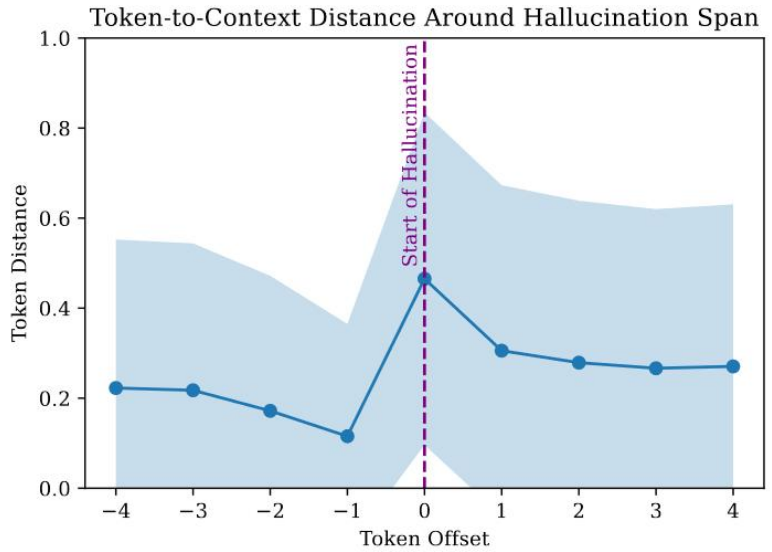
If s’t can be found we add it to the generation and continue the forward decoding. We also continue if the partial sequence S<t only contains the start-of-sequence token [SOS]

Otherwise, we backtrack again, i.e. eliminate the current last token st−1 and repeat the process

We pick L = 10T where T is the maximum generation length for our model Mθ.






Flan-T5 XL

Figure 2: **Average token probability (top) and token-to-context distance (bottom) around the hallucination span.** Token offset 0 stands for the token where hallucination starts, negative offsets stand for the tokens before hallucination and positive ones are for the hallucinated tokens. On average, the token which starts the hallucination has the lowest probability and is the furthest away from the context tokens compared to surrounding ones.

Offset 0 represents where the hallucination starts, the negative offsets represent preceding tokens the positive offsets represent successive tokens.

幻觉出现的地方
1.概率最低
2.距离最远

two models:
Flan-T5 XL (Chung et al., 2022)
and LLaMA (Touvron et al., 2023a).

dataset:
Newsroom (Grusky et al., 2018),
CNN/Dailymail (Nallapati et al., 2016)
XSUM (Narayan et al., 2018)

We consider two versions of CoBa: (1) CoBa that only uses the conditional word probabilities for detection, which we refer as CoBa in the tables; (2) CoBa that uses both the conditional word probability and the token-context distance, which we refer as CoBa-d. We use probability threshold $\delta = 0.2$ and distance threshold $\phi = 0.5$ for Flan-T5, and $\delta = 0.3$ and $\phi = 0.9$ for LLaMA

Table 1: **Faithfulness of the summaries generated with various decoding methods using Flan-T5.** All the metrics are computed between the context document and the generated summary; higher is better.

| | Method | AlignScore↑ | FactCC↑ | BS-Fact↑ | Rouge-L↑ |
|---|---|---|---|---|---|
| Newsroom | Greedy | 0.765 | 0.604 | 0.919 | 0.131 |
| | + Lookahead (every 8 tok.) | 0.768 | 0.607 | 0.920 | 0.133 |
| | + Lookahead (every 4 tok.) | 0.774 | 0.607 | 0.922 | 0.136 |
| | + Lookahead (every 2 tok.) | 0.811 | 0.662 | 0.931 | 0.153 |
| | + Lookahead (every tok.) | 0.816 | 0.662 | 0.933 | 0.159 |
| | + CAD | 0.746 | 0.490 | 0.916 | 0.145 |
| | + CoBa | 0.821 | 0.674 | 0.923 | 0.138 |
| | + CoBa-d | 0.865 | 0.709 | 0.926 | 0.145 |
| | + CoBa + CAD | 0.773 | 0.515 | 0.919 | 0.149 |
| | + CoBa-d + CAD | 0.820 | 0.560 | 0.922 | 0.161 |

Table 2: **Faithfulness of the summaries generated with various decoding methods using LLaMA.** All the metrics are computed between the context document and the generated summary; higher is better.

| | Method | AlignScore↑ | FactCC↑ | BS-Fact↑ | Rouge-L↑ |
|---|---|---|---|---|---|
| Newsroom | Greedy | 0.701 | 0.321 | 0.897 | 0.161 |
| | + CAD | 0.706 | 0.247 | 0.910 | 0.170 |
| | + CoBa | 0.715 | 0.328 | 0.906 | 0.162 |
| | + CoBa-d | 0.729 | 0.335 | 0.906 | 0.164 |
| XSUM | Greedy | 0.798 | 0.406 | 0.931 | 0.221 |
| | + CAD | 0.783 | 0.335 | 0.931 | 0.237 |
| | + CoBa | 0.800 | 0.410 | 0.932 | 0.221 |
| | + CoBa-d | 0.805 | 0.418 | 0.933 | 0.223 |
| CNN/DM | Greedy | 0.750 | 0.316 | 0.900 | 0.152 |
| | + CAD | 0.740 | 0.251 | 0.919 | 0.176 |
| | + CoBa | 0.753 | 0.323 | 0.902 | 0.153 |
| | + CoBa-d | 0.759 | 0.327 | 0.902 | 0.154 |

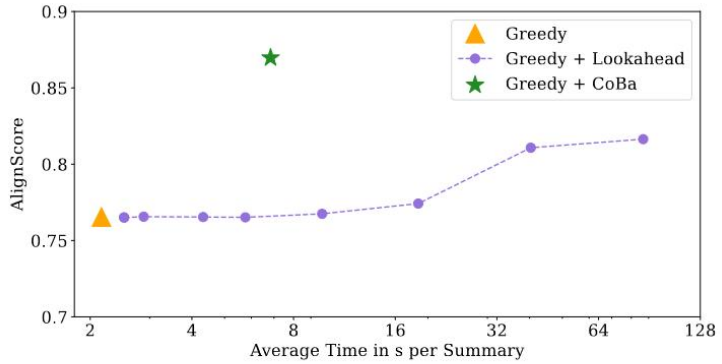



Figure 3: **AlignScore vs. Generation Time.** Note that the the x-axis is in log scale. The curve for Lookahead represents doing lookahead every $k$ tokens for $k$ from 200 to 1. CoBa attains the highest AlignScore with more than 10x speedup.

Flan-T5 XL model

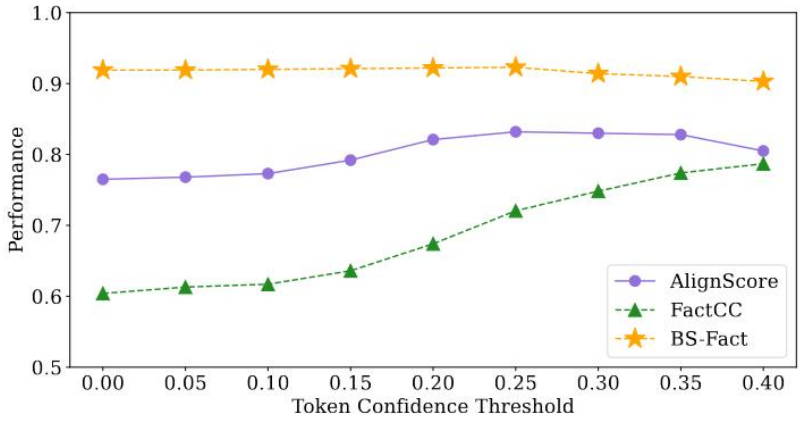



Figure 4: **Ablation on the token confidence threshold for CoBa.** High is better for all metrics. Most metrics saturate around threshold 0.2-0.25.

Table 3: **Ablation on the threshold on token embedding distance.** We use token confidence threshold $\delta = 0.2$ while varying the distance threshold $\varphi$ for all the experiments in this table.

| Dist. Thresh | AlignScore↑ | FactCC↑ | BS-Fact↑ | Rouge-L↑ |
|---|---|---|---|---|
| N/A | 0.821 | 0.674 | 0.923 | 0.138 |
| 0.9 | 0.825 | 0.677 | 0.924 | 0.139 |
| 0.7 | 0.859 | 0.699 | 0.925 | 0.143 |
| 0.5 | 0.865 | 0.709 | 0.926 | 0.145 |
| 0.3 | 0.867 | 0.718 | 0.920 | 0.146 |
| 0.1 | 0.867 | 0.720 | 0.920 | 0.146 |

**Document:** Dear God, what have they done to Joseph Gordon-Levitt's face? In Rian Johnson's time-travel action drama, "Looper," the star has been prosthetically altered with a fresh set of eyebrows, a snubbed nose, green contact lenses, and what appear to be new lips. He's supposed to look like the young Bruce Willis. What he resembles, mostly, is mid-period Devo. The imposture is hardly convincing, but "Looper" is fast enough, weird enough, and just about smart enough to make you forget about that. The movie wants to mess with your head, depositing us in a decrepit, overcrowded 2044 and then sprinkling on time-loop … This is an article preview. The full story is available to BostonGlobe.com subscribers.

**Greedy Decoding:** The Boston Globe's review of "Looper" by **John Sutter**.
**CoBa:** The Boston Globe reviews "Looper" by **Rian Johnson**.

**Document:** Scunthorpe midfielder Neal Bishop has signed a one-year contract extension. The 35-year-old joined the Iron from Blackpool in 2013 and has made 119 league appearances for the League One side. He helped them to a third-placed finish this season, before they were beaten by Millwall in the play-off semi-finals. Bishop told the club website: "With the way the season finished, it's a sense of unfinished business and it was disappointing for all of us."

**Greedy Decoding:** Scunthorpe **United** midfielder Neal Bishop has signed a new one-year contract with the Iron.
**CoBa-d:** Scunthorpe have signed midfielder Neal Bishop on a one-year contract extension.

Figure 5: **Qualitative examples of greedy decoding vs. CoBa and CoBa-d.** The hallucinated content is marked in red and the corrected details are marked in green. CoBa and CoBa-d correctly remove the hallucinated content by triggering backtracking at corresponding positions and generate summaries with more and faithful details.