

384 A Algorithm

385 The algorithm consists of two main processes. The first step is **Heads Anchoring and Steering Vectors Induc-**
 386 **ing**, which determines the heads that need to be intervened, the classifiers, and the steering vectors. The second
 387 step is **Generation with Flexible Steering and Backtracking**, which involves using classifiers to evaluate the
 388 internal activations of LLMs. When the internal activations exceed a certain threshold, a backtracking operation
 389 will be performed. After the backtracking, the activations will be regenerated by adding the activation vectors.
 390 Since our method also backtracks to the beginning when the number of generated tokens is less than s , we start
 391 tracking from the s -th token, as shown in Line 14 of the algorithm.

Algorithm 1 The overall flow of FASB

Require: LLM, dataset \mathcal{D} , intervention strength α , threshold β , backtracking number s , maximum number of generated tokens M

```

1: Step1: Heads Anchoring and Steering Vectors Inducing
2: if Probe then
3:   Train classifier  $p_{\theta^{\ell,h}}$  on dataset  $\mathcal{D}$ ; ▷ Equation 1
4:   Use the parameters of the probe as the steering vector  $\theta^{\ell,h}$ ;
5: else if Prototype then
6:   Construct two prototypes from the dataset  $\mathcal{D}$ ; ▷ Equation 2
7:   Obtain the classifier  $p_{\theta^{\ell,h}}$ ; ▷ Equation 3
8:   Obtain the steering vector  $\theta^{\ell,h}$ ; ▷ Equation 4
9: end if
10: Step2: Generation with Flexible Steering and Backtracking
11: for  $j = 1$  to  $M$  do
12:   Generate the  $j$ -th token;
13:   Evaluate the current activation using the classifier; ▷ Equation 5
14:   if  $p(\mathbf{x}_{i,j}) > \beta$  and  $j \geq s$  then
15:     Backtrack  $s$  tokens;
16:     Calculate the intervention strength; ▷ Equation 6
17:     for  $k = (j-s+1)$  to  $M$  do
18:       Intervene on the current activation; ▷ Equation 7
19:       Generate the  $k$ -th token;
20:     end for
21:   break;
22: end if
23: end for

```

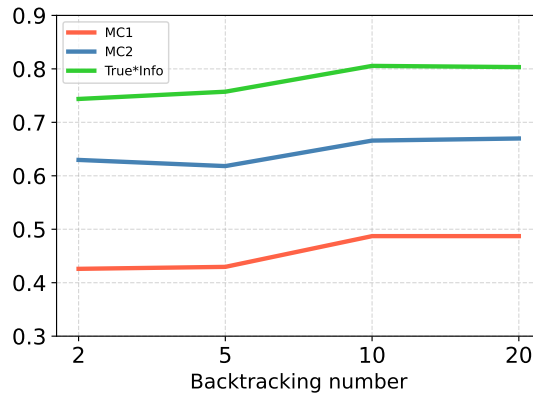


Figure 8: Effects of the number of tokens for backtracking.

392 B Effects of the Number of Tokens for Backtracking

393 In order to better investigate the impact of the backtracking number on our method, we conduct experiments on
 394 the TruthfulQA dataset.

Table 5: Comparison with two fine-grained models.

Methods	True*Info	MC1	MC2
Probe	80.56	48.71	66.58
BTB	81.60	48.83	67.62
GCBB	81.96	50.67	67.48

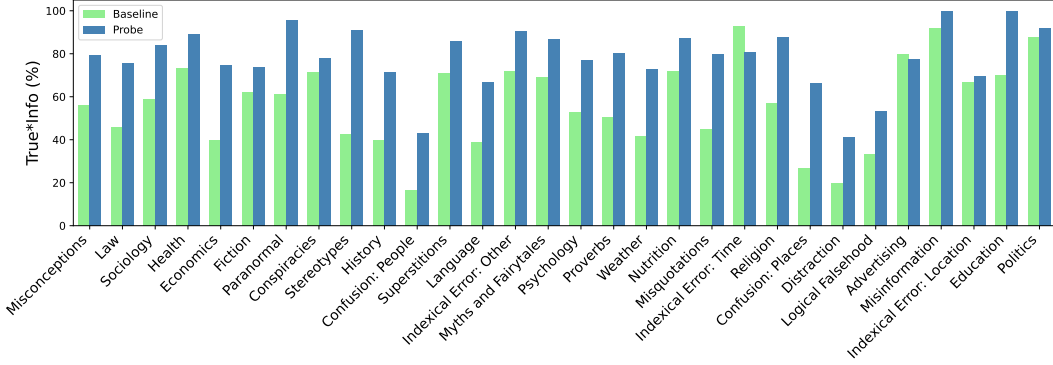


Figure 9: True*Info scores split across subcategories on LLaMA2-7B-CHAT, sorted by the difference between baseline and probe method. Subcategories with less than 10 questions are not shown.

As shown in Figure 8, we investigate the effect of different token numbers for backtracking (*i.e.*, 2, 5, 10, 20) on the MC1, MC2, and True*Info metrics. MC1, MC2, and True*Info generally show an increasing trend with the increase in the number of backtracking steps. This is because, as the number of backtracking increases, our method can intervene in the model’s internal activations earlier, thereby achieving improved performance.

C Fine-Grained Model Comparison

We further propose two models based on Probe for fine-grained analysis. The difference between them and Probe lies in the detection location and the point at which intervention begins, resulting in different overheads. (1) The first method directly **BackTracks** to the **B**eginning (**BTB**). Specifically, when a deviation is detected during generation, BTB backtracks not just s tokens for regeneration, but to the beginning. Our method generates s additional tokens for texts that require backtracking, whereas this variant generates even more. (2) The second method performs detection after the **G**eneration is **C**omplete and then **B**acktracks to the **B**eginning (GCBB). Notably, this method incurs approximately double the overhead for texts that require backtracking.

As shown in the results in Table 5, GCBB usually achieves the best performance, and both BTB and GCBB outperform our proposed Probe method on the TruthfulQA dataset. This is because GCBB has access to the full output of the LLM, allowing it to better determine the intervention strength. However, GCBB inevitably introduces additional overhead. BTB may also sometimes require full regeneration. The advantage of Probe is that it achieves good performance while maintaining stable and relatively low additional overhead.

D Results across TruthfulQA Categories

TruthfulQA is split into 38 subcategories, including politics, language, education, psychology and others. We compare our method with the baseline method without intervention using the True*Info metric across all subcategories with 10 or more questions, where the subcategories are ranked in descending order based on their quantity within the dataset, as shown in Figure 9.

Our method demonstrates significant enhancement across most subcategories, with the overall performance improvement showing uniform distribution rather than concentration in particular domains, thereby validating its efficacy.

E Limitations

In this paper, we did not validate the effectiveness of our method on larger models. Instead, we focused on experiments conducted on the following models: LLaMA2-7B, LLaMA2-7B-CHAT, LLaMA2-13B-Chat,

423 LLaMA3.1-8B-Instruct, Qwen2.5-7B, and Qwen2.5-7B-Instruct. Additionally, we did not conduct experiments
424 on LLMs in languages other than English.