Semantic Entropy와 Semantic Energy의 상보적 결합을 통한 LLM 환각 탐지

Hallucination Detection in LLMs via Complementary Combination of Semantic Entropy and Semantic Energy

문정혁
서울대학교 컴퓨터공학부
student@snu.ac.kr

2026년 2월

Abstract

대규모 언어모델(LLM)의 환각(hallucination)은 의료, 법률, 금융 분야에서 심각한 위험을 초래한다. 기존의 Semantic Entropy(SE) 기반 탐지 방법은 모델이 일관되게 틀린 답변을 생성하는 경우(Zero-SE 문제)에 한계를 보인다. 본 연구에서는 SE와 Semantic Energy의 상보적 특성을 분석하고, SE가 낮을 때 Energy로 대체하는 SE-gated cascade 방법을 제안한다. TruthfulQA 데이터셋에서 K=5 응답 샘플링 실험 결과, Zero-SE 영역(전체의 19%)에서 Energy가 AUROC 0.736으로 환각을 구분할 수 있음을 확인하였다. 또한 SE-gated cascade는 SE 단독 대비 AUROC +0.030 개선(0.613에서 0.643)을 달성하였으며, SE와 Energy의 합집합 탐지율은 89.6%에 도달하여 두 메트릭의 상보성을 검증하였다.

Keywords: LLM 환각 탐지, Semantic Entropy, Semantic Energy, Zero-SE 문제, Cascade Detection

1 서론

1.1 연구 배경

대규모 언어모델(LLM)은 자연어 처리 분야에서 놀라운 성능을 보여주고 있으나, 사실과 다른 내용을 그럴듯하게 생성하는 환각(hallucination) 문제가 심각한 한계로 지적되고 있다. 특히 의료, 법률, 금융 등 정확성이 중요한 분야에서 LLM의 환각은 치명적인 결과를 초래할 수 있다.

환각 탐지를 위한 대표적인 방법으로 Semantic Entropy(SE)가 있다 [1]. SE는 하나의 질문에 대해 여러 응답을 샘플링한 뒤, 의미적으로 클러스터링하여 응답의 다양성을 측정한다. SE가 높으면 모델이 혼란스러워하는 것이므로 환각 가능성이 높다고 판단한다.

그러나 SE는 근본적인 한계를 가진다. 모델이 모든 응답에서 **일관되게 틀린 답변**을 생성하면 SE 값이 0에 가까워지고, 이 경우 SE만으로는 환각 여부를 판단할 수 없다. 우리는 이를 **Zero-SE 문제**라고 정의한다.

1.2 연구 목적

본 연구의 목적은 다음과 같다:

1. Zero-SE 문제의 존재와 심각성을 정량적으로 분석한다.
2. SE와 Semantic Energy가 서로 다른 유형의 환각을 탐지함을 검증한다.
3. 두 메트릭을 상보적으로 결합하는 SE-gated cascade 방법을 제안한다.

2 관련 연구

2.1 Semantic Entropy (SE)

Farquhar et al. [1]이 Nature에서 제안한 Semantic Entropy는 LLM의 불확실성을 측정하는 대표적인 방법이다.

SE의 계산 과정은 다음과 같다:

1. 하나의 질문에 대해 LLM으로부터 K 개의 응답을 샘플링한다.

2. 자연어 추론(NLI) 모델을 사용하여 응답들을 의미적으로 클러스터링한다.

3. 클러스터 분포의 Shannon entropy를 계산한다.

수식으로 표현하면:

$$SE = -\sum_{c \in C} p(c) \log p(c) \tag{1}$$

여기서 C는 의미 클러스터 집합이고, $p(c)$는 클러스터 c에 속하는 응답의 비율이다.

2.2 Semantic Energy

Ma et al. [2]이 제안한 Semantic Energy는 토큰 단위의 확신도를 측정한다. Energy는 LLM이 각 토큰을 생성할 때 부여한 raw logit 값을 기반으로 계산된다:

$$Energy = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T_i} -z(x_t^{(i)}) \tag{2}$$

여기서 $z(x_t)$는 토큰 x_t의 logit 값이다.

SE는 응답 간 다양성(inter-response diversity)을 측정하고, Energy는 각 토큰의 내재적 확신도(intra-token confidence)를 측정한다. 따라서 SE=0인 상황(모든 응답이 동일)에서도 Energy는 여전히 정보를 제공할 수 있다.

2.3 Zero-SE 문제

Zero-SE는 K 개 응답이 모두 단일 NLI 클러스터에 속하는 경우(SE≈0)를 의미한다. SE는 정의상 0이므로 환각 여부에 대한 판별력이 전혀 없으나, 실제로 이 영역에 환각이 다수 포함되어 있다.

3 제안 방법

3.1 핵심 통찰

LLM이 환각을 일으킬 때, 모델이 해당 주제를 아는지 여부에 따라 근본적으로 다른 방식으로 환각한다:

Table 1: 환각 유형별 특성과 최적 탐지기

유형	원인	행동적 신호	최적 탐지기
혼란	알지만 헷갈림	다양한 오답	SE
지어냄	모르고 지어냄	일관된 오답	Energy

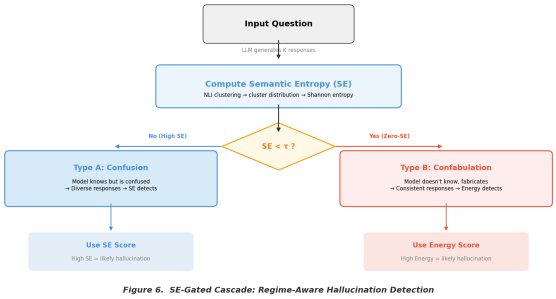

Figure 1: SE-Gated Cascade 개념도

Algorithm 1 SE-Gated Cascade Detection

Require: 질문 q, 임계값 τ
LLM으로 $K = 5$ 응답 생성
NLI 클러스터링으로 SE 계산
토큰 logit으로 Energy 계산
if $SE < \tau$ **then**
  return Energy (지어냄 영역)
else
  return SE (혼란 영역)
end if

3.2 SE-Gated Cascade

제안하는 SE-gated cascade는 그림 1과 같다.

본 연구에서는 TruthfulQA 데이터셋에서 검증 세트를 통해 최적 임계값 $\tau = 0.526$을 도출하였다.

4 실험

4.1 실험 설정

TruthfulQA 데이터셋 [3]을 사용하였다. 이 데이터셋은 대중적 오개념을 유도하는 질문들로 구성되어 있다.

Table 2: 실험 파이프라인 설정

항목	설정
LLM	Qwen2.5-3B-Instruct
NLI 모델	DeBERTa-large-mnli
샘플링 수 (K)	5
Temperature	0.7
전체 샘플	200 (환각 164, 정상 36)

4.2 Zero-SE 현상 분석

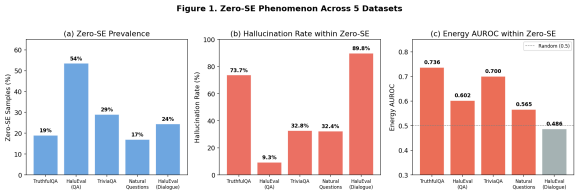

Figure 2: Zero-SE 현상 개요

TruthfulQA에서 Zero-SE 현상을 정량적으로 분석하였다 (표 3).

Table 3: TruthfulQA Zero-SE 영역 분석

지표	값
Zero-SE 비율	19.0% (38/200)
Zero-SE 내 환각률	73.7% (28/38)
Zero-SE Energy AUROC	0.736
95% 신뢰구간	[0.52, 0.93]

전체 샘플의 19%가 Zero-SE에 해당하며, 이 중 73.7%가 실제 환각이다. SE로는 판별 불가하지만, Energy AUROC 0.736으로 구분 가능하다.

4.3 SE 구간별 Crossover 분석

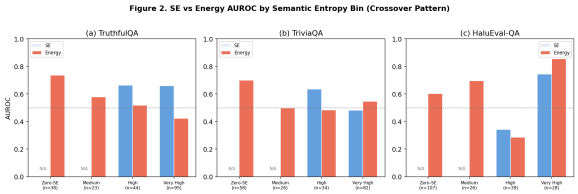

Figure 3: SE 구간별 SE vs Energy AUROC 비교

표 4에서 Zero-SE 영역에서는 Energy가 우세하고, High-SE 영역에서는 SE가 우세함을 확인할 수 있다.

Table 4: SE 구간별 탐지 성능

SE 구간	n	SE	Energy
Zero [0, 0.05]	38	N/A	**0.736**
High (0.6, 1.0]	44	**0.664**	0.517
Very High (1.0+)	95	**0.658**	0.422

4.4 SE-Gated Cascade 성능

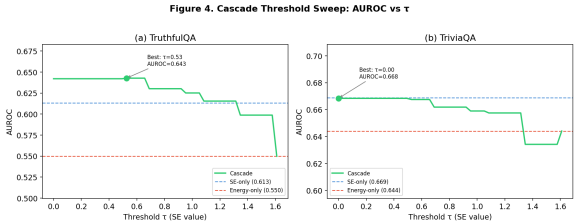

Figure 4: Cascade 임계값 sweep 결과

Cascade 방법이 SE-only 대비 AUROC 0.030 개선을 달성하였다.

4.5 상보성 분석

Energy만 탐지하는 17.7%는 SE로는 잡을 수 없는 환각이다.

Table 5: 탐지 방법별 AUROC 비교

방법	AUROC	Δ
SE-only	0.613	-
Energy-only	0.550	-0.063
Cascade	**0.643**	**+0.030**

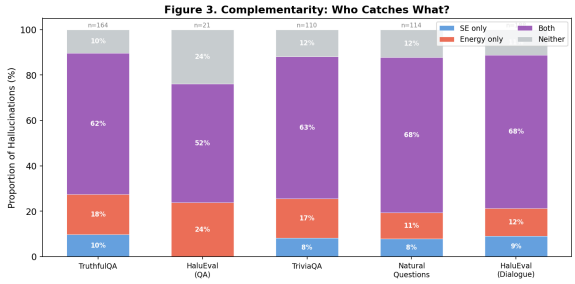

Figure 5: SE와 Energy의 환각 탐지 영역 비교

Table 6: 환각 탐지 상보성 (164개 환각 기준)

탐지 영역	비율
SE만 탐지	9.8% (16개)
Energy만 탐지	17.7% (29개)
둘 다 탐지	62.2% (102개)
둘 다 실패	10.4% (17개)
합집합 탐지율	**89.6%**

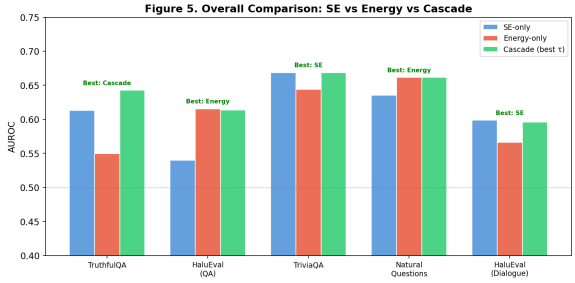

Figure 6: SE, Energy, Cascade 전체 성능 비교

4.6 전체 비교

5 결론

본 연구에서는 LLM 환각 탐지에서 SE의 한계인 Zero-SE 문제를 정의하고, Semantic Energy와의 상보적 결합을 통해 이를 해결하는 방법을 제안하였다.

TruthfulQA 데이터셋에서의 실험을 통해 다음을 확인하였다:

1. **Zero-SE 문제**: 전체 샘플의 19%가 Zero-SE에 해당하며, 이 중 73.7%가 환각이다.
2. **Energy의 효과**: Zero-SE 영역에서 Energy가 AUROC 0.736으로 환각을 구분한다.
3. **Cascade의 개선**: SE-gated cascade는 SE-only 대비 AUROC +0.030 개선을 달성한다.
4. **상보성**: 합집합 탐지율은 89.6%에 도달한다.

5.1 향후 연구

다양한 데이터셋 검증, 적응적 임계값 결정, Energy의 이론적 분석, Cross-dataset 일반화 등이 향후 연구 과제이다.

References

[1] S. Farquhar et al., "Detecting hallucinations in large language models using semantic entropy," *Nature*, vol. 630, pp. 625–630, 2024.

[2] Z. Ma et al., "Semantic Energy: A novel approach for detecting confabulation in language models," *arXiv*, 2025.

[3] S. Lin et al., "TruthfulQA: Measuring how models mimic human falsehoods," *ACL*, 2022.