공학학사학위논문

Semantic Entropy와 Semantic Energy의 상보적 결합을 통한 LLM 환각 탐지

2026 년 2 월

서울대학교

자유전공학부

문 정 혁

Semantic Entropy와 Semantic Energy의 상보적 결합을 통한 LLM 환각 탐지

지도교수 김 남 수

이 논문을 공학 학사학위 논문으로 제출함

서울대학교

자유전공학부

문 정 혁

문정혁 의 학사 학위 논문을 인준함

2026 년 2 월 13일

지 도 교 수 (인)

초록

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, 즉 Zero-SE 문제에서 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 영역이 전체의 19%를 차지하며 이 중 73.7%가 환각임을 확인하였다. SE는 이 영역에서 판별력이 없으나, Energy는 AUROC 0.736으로 환각을 효과적으로 구분할 수 있었다. 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.

본 연구는 SE와 Energy가 서로 다른 환각 패턴(혼란 대 지어냄)을 탐지함을 최초로 규명하였으며, 두 메트릭을 안전하게 결합하는 방법을 제안하여 LLM 환각 탐지 분야에 기여한다.

주요어: 대규모 언어모델, 환각 탐지, Semantic Entropy, Semantic Energy, Zero-SE 문제

목차

초록 i

제 1 장서론 1

1.1 연구 배경 1

1.2 연구 목적 1

1.3 논문 구성 2

제 2 장관련 연구 3

2.1 Semantic Entropy 3

2.1.1 작동 원리 3

2.1.2 해석 3

2.2 Semantic Energy 4

2.2.1 작동 원리 4

2.2.2 SE와의 차이점 4

2.3 Zero-SE 문제 4

제 3 장제안 방법 5

3.1 핵심 통찰 5

3.2 SE-Gated Cascade 5

3.2.1 알고리즘 5

3.2.2 임계값 결정 5

3.3 이론적 근거 7

제 4 장실험 8

4.1 실험 설정 8

4.1.1 데이터셋 8

4.1.2 파이프라인 설정 8

4.2 Zero-SE 현상 분석 9

4.2.1 Zero-SE 비율 및 환각률 9

4.2.2 SE 구간별 Crossover 분석 9
4.3 SE-Gated Cascade 성능 10
4.4 상보성 분석 11
4.5 전체 비교 12

제 5 장 결론 **13**

5.1 연구 요약 13
5.2 학술적 기여 13
5.3 향후 연구 13

표 목차

표 3.1 환각 유형별 특성과 최적 탐지기 5

표 4.1 TruthfulQA 데이터셋 통계 8
표 4.2 실험 파이프라인 설정 8
표 4.3 TruthfulQA Zero-SE 영역 분석 9
표 4.4 SE 구간별 탐지 성능 (TruthfulQA) 10
표 4.5 탐지 방법별 AUROC 비교 11
표 4.6 TruthfulQA 환각 탐지 상보성 (164개 환각 기준) 11

그림 목차

그림 3.1 SE-Gated Cascade 개념도: SE가 낮으면 Energy로, 높으면 SE로 판단 6

그림 4.1 Zero-SE 현상 개요: 전체 대비 비율, 환각률, Energy AUROC 9
그림 4.2 SE 구간별 SE vs Energy AUROC 비교 10
그림 4.3 Cascade 임계값(τ) sweep 결과 11
그림 4.4 SE와 Energy의 환각 탐지 영역 비교 12
그림 4.5 SE, Energy, Cascade 전체 성능 비교 12

제 1 장 서론

1.1 연구 배경

대규모 언어모델(Large Language Model, LLM)은 자연어 처리 분야에서 혁명적인 성능을 보여주고 있다. GPT-4, Claude, Gemini 등 최신 모델들은 텍스트 생성, 질의응답, 요약, 번역 등 다양한 태스크에서 인간 수준의 성능을 달성하고 있다. 그러나 이러한 발전에도 불구하고, LLM은 사실과 다른 내용을 그럴듯하게 생성하는 환각(hallucination) 문제를 가지고 있다.

환각 문제는 LLM의 실용적 적용에 있어 가장 심각한 장애물 중 하나이다. 의료 분야에서 잘못된 진단 정보를 제공하거나, 법률 분야에서 존재하지 않는 판례를 인용하거나, 금융 분야에서 부정확한 수치를 제시하는 경우 치명적인 결과를 초래할 수 있다. 따라서 LLM이 생성한 텍스트에서 환각을 자동으로 탐지하는 것은 매우 중요한 연구 과제이다.

환각 탐지를 위한 대표적인 방법으로 Semantic Entropy(SE)가 있다. Farquhar 등이 2024년 Nature에 발표한 이 방법은 하나의 질문에 대해 여러 응답을 샘플링한 뒤, 자연어 추론(Natural Language Inference, NLI) 모델을 사용하여 의미적으로 클러스터링하고, 클러스터 분포의 엔트로피를 계산한다. SE가 높으면 모델이 다양한 응답을 생성하므로 혼란스러워하는 것이고, 이는 환각 가능성이 높음을 의미한다.

그러나 SE는 근본적인 한계를 가진다. 모델이 모든 응답에서 **일관되게 틀린 답변**을 생성하면, 모든 응답이 동일한 의미 클러스터에 속하게 되어 SE 값이 0에 가까워진다. 이 경우 SE만으로는 해당 답변이 맞는지 틀린지 판단할 수 없다. 우리는 이 현상을 **Zero-SE 문제**라고 정의한다.

1.2 연구 목적

본 연구의 목적은 다음과 같다:

1. **Zero-SE 문제의 정량화**: Zero-SE 현상이 실제로 얼마나 발생하며, 그 중 환각이 얼마나 포함되어 있는지 정량적으로 분석한다.
2. **상보적 메트릭 검증**: SE와 Semantic Energy가 서로 다른 유형의 환각을 탐지하는지 검증한다.
3. **SE-gated Cascade 제안**: 두 메트릭을 상보적으로 결합하여 Zero-SE 문제를 해결하는 SE-gated cascade 방법을 제안한다.

1.3 논문 구성

본 논문은 다음과 같이 구성된다. 제 2장에서는 Semantic Entropy, Semantic Energy 및 Zero-SE 문제에 대한 관련 연구를 살펴본다. 제 3장에서는 제안하는 SE-gated cascade 방법론을 설명한다. 제 4장에서는 TruthfulQA 데이터셋에서의 실험 결과를 분석하고, 제 5장에서 결론 및 향후 연구 방향을 제시한다.

제 2 장 관련 연구

2.1 Semantic Entropy

Semantic Entropy(SE)는 Farquhar 등이 2024년 Nature에 발표한 LLM 불확실성 측정 방법이다. 기존의 토큰 단위 확률 기반 불확실성 측정 방법과 달리, SE는 응답의 의미적 내용을 기반으로 불확실성을 측정한다.

2.1.1 작동 원리

SE의 계산 과정은 다음과 같다:

1. **응답 샘플링**: 하나의 질문 q에 대해 LLM으로부터 K개의 응답 $\{r_1, r_2, ..., r_K\}$를 샘플링한다. 일반적으로 temperature $\tau > 0$을 사용하여 다양한 응답을 생성한다.
2. **의미적 클러스터링**: 자연어 추론(NLI) 모델을 사용하여 응답들을 의미적으로 클러스터링한다. 두 응답이 서로 entailment 관계에 있으면 같은 클러스터로 분류한다.
3. **엔트로피 계산**: 클러스터 분포의 Shannon entropy를 계산한다.

수식으로 표현하면 다음과 같다:

$$SE = -\sum_{c \in C} p(c) \log p(c) \tag{2.1}$$

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

2.1.2 해석

SE 값의 의미는 다음과 같이 해석된다:

SE가 높음: 응답이 여러 클러스터로 분산되어 있어 모델이 혼란스러워하는 상태이다. 이는 환각 가능성이 높음을 의미한다.

SE가 낮음: 응답이 하나 또는 소수의 클러스터에 집중되어 있어 모델이 일관된 답변을 생성하는 상태이다. 그러나 이것이 반드시 정답을 의미하지는 않는다.

SE가 0: 모든 응답이 하나의 클러스터에 속한다. 모델이 매우 일관된 답변을 생성하지만, 일관되게 틀린 답변일 수도 있다.

2.2 Semantic Energy

Semantic Energy는 Ma 등이 2025년에 제안한 방법으로, 토큰 단위의 확신도를 측정한다. SE가 응답 간 다양성을 측정하는 것과 달리, Energy는 각 토큰 생성 시 모델의 내재적 확신도를 측정한다.

2.2.1 작동 원리

Energy는 LLM이 각 토큰을 생성할 때 부여한 raw logit 값을 기반으로 계산된다:

$$Energy = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T_i} -z(x_t^{(i)}) \tag{2.2}$$

여기서 n은 응답의 수, T_i는 i번째 응답의 토큰 수, $z(x_t)$는 토큰 x_t의 logit 값이다.

2.2.2 SE와의 차이점

SE와 Energy의 근본적인 차이점은 다음과 같다:

SE: 응답 간 다양성(inter-response diversity)을 측정한다. 여러 응답이 의미적으로 얼마나 다른지를 본다.

Energy: 각 토큰의 내재적 확신도(intra-token confidence)를 측정한다. 모델이 각 토큰을 생성할 때 얼마나 확신을 가지고 있는지를 본다.

이 차이점으로 인해, SE=0인 상황(모든 응답이 동일한 의미)에서도 Energy는 여전히 정보를 제공할 수 있다. 모델이 일관되게 답변하더라도 각 토큰 생성 시의 확신도는 다를 수 있기 때문이다.

2.3 Zero-SE 문제

Zero-SE는 K개 응답이 **모두 단일 NLI 클러스터**에 속하는 경우(SE $\approx$ 0)를 의미한다. 이 상황에서 발생하는 문제점은 다음과 같다:

1. **판별력 부재**: SE가 정의상 0이므로, 환각 여부에 대한 판별력이 전혀 없다. 모든 Zero-SE 샘플이 동일한 SE 값(0)을 가지므로 순위를 매길 수 없다.
2. **높은 환각률**: 실제로 Zero-SE 영역에 환각이 다수 포함되어 있다. 모델이 일관되게 틀린 답변을 생성하는 경우가 이에 해당한다.
3. **기존 해결책 부재**: 기존 연구에서는 이 문제에 대한 명시적인 해결책이 제시되지 않았다.

본 연구는 Zero-SE 영역에서 Energy를 대신 사용하는 방법을 제안하여 이 문제를 해결한다.

제 3 장 제안 방법

3.1 핵심 통찰

본 연구의 핵심 통찰은 LLM이 환각을 일으킬 때, 모델이 해당 주제를 아는지 여부에 따라 근본적으로 다른 방식으로 환각한다는 것이다. 표 3.1는 두 가지 환각 유형을 정리한 것이다.

표 3.1 환각 유형별 특성과 최적 탐지기

유형	원인	행동적 신호	최적 탐지기
혼란(Confusion)	모델이 알지만 헷갈림	다양한 오답	SE
지어냄(Confabulation)	모델이 모르고 지어냄	일관된 오답	Energy

혼란(Confusion)은 모델이 관련 지식을 가지고 있지만 헷갈리는 경우이다. 예를 들어, 비슷한 이름을 가진 여러 인물 중 누구인지 혼동하는 경우가 이에 해당한다. 이 경우 모델은 여러 번 질문받으면 다양한 오답을 생성하며, 이는 높은 SE로 나타난다.

지어냄(Confabulation)은 모델이 관련 지식이 없어서 그럴듯하게 지어내는 경우이다. 존재하지 않는 책이나 인물에 대해 질문받았을 때, 모델은 일관되게 같은 거짓 정보를 생성한다. 이 경우 SE는 낮지만(모든 응답이 동일), 각 토큰 생성 시의 Energy는 높다.

3.2 SE-Gated Cascade

제안하는 SE-gated cascade는 그림 3.1과 같이 작동한다.

3.2.1 알고리즘

SE-gated cascade의 알고리즘은 Algorithm 1와 같다.

3.2.2 임계값 결정

임계값 τ는 SE와 Energy 중 어떤 메트릭을 사용할지 결정하는 기준이다. 본 연구에서는 TruthfulQA 데이터셋의 검증 세트에서 grid search를 통해 최적 임계값 $\tau = 0.526$을 도출하였다.

이 임계값은 Zero-SE 영역(SE < 0.1인 샘플)을 효과적으로 포함하면서, High-SE 영역에서는 SE의 장점을 유지하도록 설정되었다.

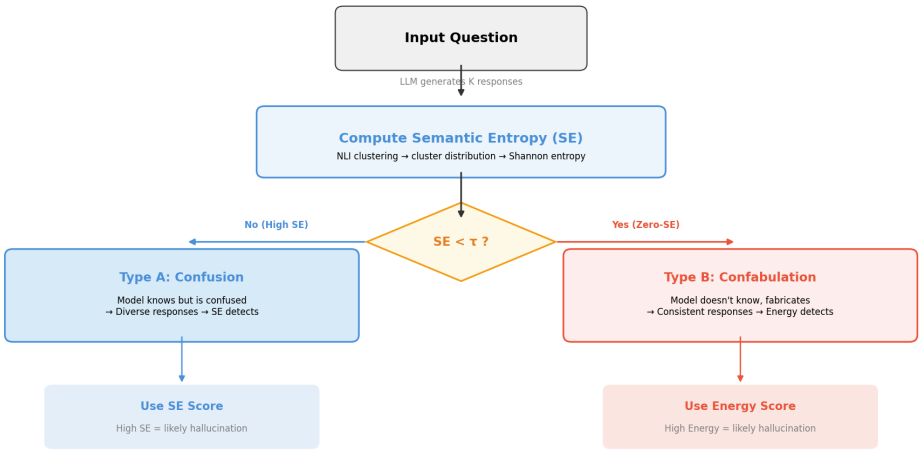

Figure 6. SE-Gated Cascade: Regime-Aware Hallucination Detection

그림 3.1 SE-Gated Cascade 개념도: SE가 낮으면 Energy로, 높으면 SE로 판단

Algorithm 1 SE-Gated Cascade Detection

Require: 질문 q, 임계값 τ

LLM으로부터 $K = 5$ 응답 생성
NLI 클러스터링을 통해 SE 계산
토큰 logit 값으로 Energy 계산
if $SE < \tau$ **then**
    return Energy (지어냄 영역)
else
    return SE (혼란 영역)
end if

3.3 이론적 근거

Physics of Language Models 연구에서는 다음과 같이 설명한다:

> "지식이 안정적으로 추출되려면, 사전학습 중 충분히 증강(paraphrasing, shuffling 등)되어야 한다.
> 이러한 증강 없이는 지식이 암기되더라도 추출 불가능할 수 있다."

이를 기반으로 한 우리의 해석은 다음과 같다:

충분히 학습된 지식: 사전학습 중 많이 노출되어 잘 학습된 지식은 혼란형 오류를 발생시킨다. 모델이 여러 관련 정보를 알고 있어 혼동하기 때문이다. 이 경우 SE가 효과적이다.

불충분하게 학습된 지식: 사전학습 중 충분히 노출되지 않은 지식은 지어냄형 오류를 발생시킨다. 모델이 관련 정보를 모르기 때문에 그럴듯하게 지어낸다. 이 경우 Energy가 효과적이다.

제 4 장 실험

4.1 실험 설정

4.1.1 데이터셋

TruthfulQA 데이터셋을 사용하였다. 이 데이터셋은 인간이 흔히 가지는 오개념을 유도하는 질문들로 구성되어 있어, LLM의 환각 탐지 연구에 적합하다. 표 4.1은 데이터셋의 통계를 보여준다.

표 4.1 TruthfulQA 데이터셋 통계

항목	값
전체 샘플 수	200
환각 샘플 수	164 (82.0%)
정상 샘플 수	36 (18.0%)

4.1.2 파이프라인 설정

표 4.2은 실험 파이프라인 설정을 보여준다.

표 4.2 실험 파이프라인 설정

항목	설정
LLM	Qwen2.5-3B-Instruct
NLI 모델	DeBERTa-large-mnli
샘플링 수 (K)	5
Temperature	0.7

파이프라인은 다음과 같이 작동한다: 질문이 입력되면 LLM이 K=5개의 응답을 생성하고, NLI 모델이 응답들을 클러스터링하여 SE를 계산한다. 동시에 각 응답의 토큰 logit 값으로 Energy를 계산한다. 최종적으로 환각 여부 라벨과 비교하여 AUROC을 산출한다.

4.2 Zero-SE 현상 분석

4.2.1 Zero-SE 비율 및 환각률

그림 4.1는 TruthfulQA에서 Zero-SE 현상을 시각화한 것이다.

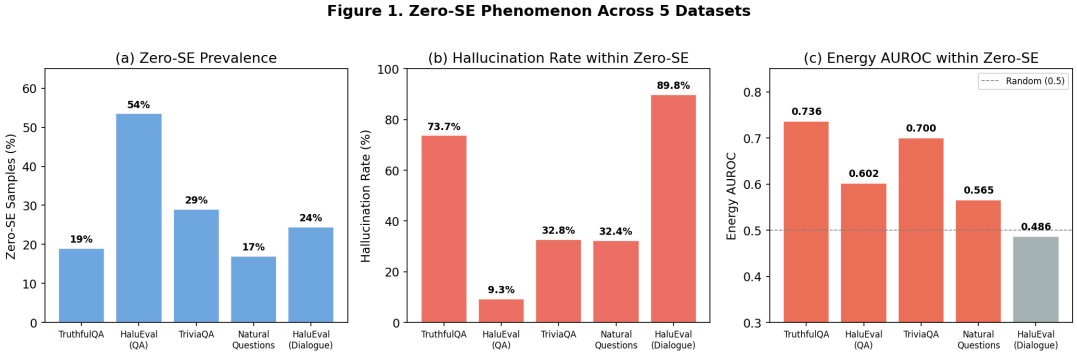

그림 4.1 Zero-SE 현상 개요: 전체 대비 비율, 환각률, Energy AUROC

표 4.3는 Zero-SE 영역의 정량적 분석 결과를 보여준다.

표 4.3 TruthfulQA Zero-SE 영역 분석

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

주요 발견은 다음과 같다:

전체 샘플의 19%가 Zero-SE에 해당한다. 이는 5개 응답이 모두 단일 의미 클러스터에 속하는 경우이다.

Zero-SE 샘플 중 73.7%가 실제 환각이다. 모델이 일관되게 틀린 답변을 생성하는 경우가 많다.

SE로는 이 영역에서 판별이 불가능하지만, Energy는 AUROC 0.736으로 환각을 효과적으로 구분한다.

4.2.2 SE 구간별 Crossover 분석

그림 4.2는 SE 구간별로 SE와 Energy의 AUROC을 비교한 것이다.

표 4.4는 각 SE 구간별 탐지 성능을 정리한 것이다.

Crossover 패턴이 명확하게 관찰된다:

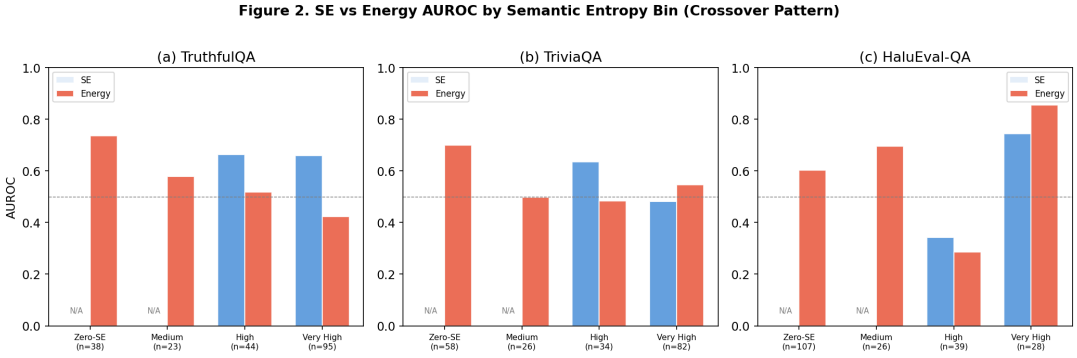

그림 4.2 SE 구간별 SE vs Energy AUROC 비교

표 4.4 SE 구간별 탐지 성능 (TruthfulQA)

SE 구간	n	환각률	SE AUROC	Energy AUROC
Zero [0, 0.05]	38	73.7%	N/A	**0.736**
Medium (0.3, 0.6]	23	78.3%	N/A	0.578
High (0.6, 1.0]	44	84.1%	**0.664**	0.517
Very High (1.0, ∞)	95	85.3%	**0.658**	0.422

Zero-SE 영역: Energy가 0.736으로 우세하다. SE는 모든 값이 0이므로 AUROC 계산이 불가능하다.

High-SE 영역: SE가 0.664로 우세하고, Energy는 0.517로 성능이 떨어진다.

Very High-SE 영역: SE가 0.658로 여전히 우세하며, Energy는 0.422로 랜덤 수준 이하이다.

이 결과는 SE-gated cascade 전략의 근거가 된다: 낮은 SE 영역에서는 Energy를, 높은 SE 영역에서는 SE를 사용해야 한다.

4.3 SE-Gated Cascade 성능

그림 4.3은 다양한 임계값 τ에 대한 cascade 성능을 보여준다.

표 4.5는 각 방법의 AUROC을 비교한 것이다.

Cascade 방법이 SE-only 대비 AUROC 0.030 개선을 달성하였다. Energy-only는 SE-only보다 성능이 낮아 단독으로는 사용하기 어렵지만, cascade를 통해 Zero-SE 영역에서의 약점을 보완할 수 있다.

Figure 4. Cascade Threshold Sweep: AUROC vs τ

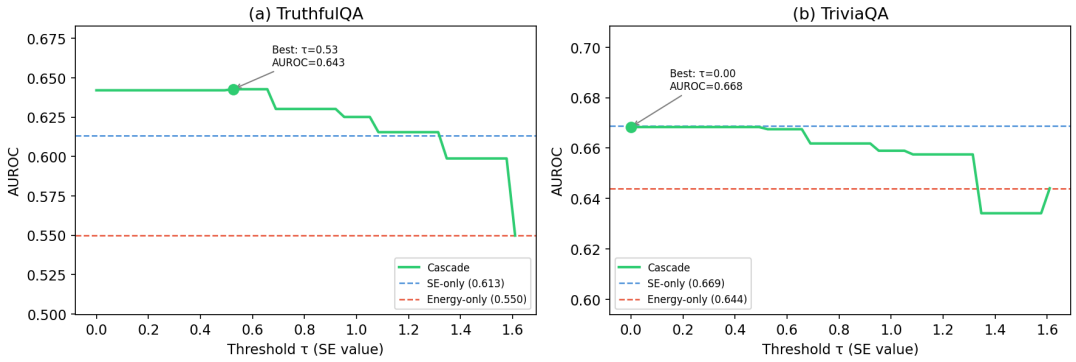

그림 4.3 Cascade 임계값(τ) sweep 결과

표 4.5 탐지 방법별 AUROC 비교

방법	AUROC	Δ vs SE-only
SE-only	0.613	-
Energy-only	0.550	-0.063
Cascade (τ=0.526)	**0.643**	**+0.030**

4.4 상보성 분석

그림 4.4는 SE와 Energy가 각각 탐지하는 환각 영역을 시각화한 것이다.

표 4.6는 164개 환각 샘플에 대한 탐지 영역 분석 결과이다. 여기서 "탐지"는 해당 메트릭의 점수가 상위 20% (80th percentile)에 속하는 경우를 의미한다.

표 4.6 TruthfulQA 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%**

주목할 점은 다음과 같다:

Energy만 탐지하는 17.7%: 이 환각들은 SE를 아무리 좋은 threshold로 설정해도 잡을 수 없다. 모델이

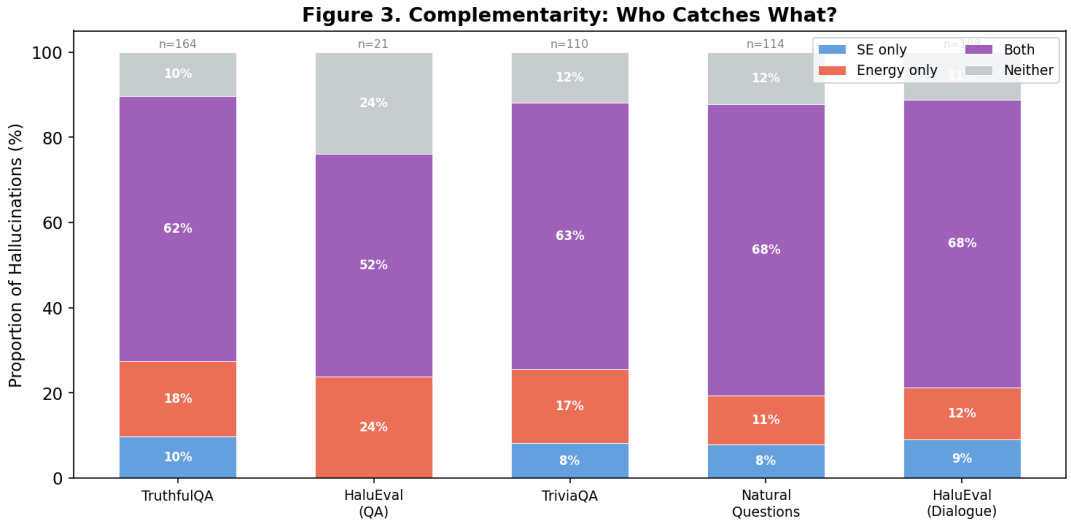

그림 4.4 SE와 Energy의 환각 탐지 영역 비교

일관되게 틀린 답변을 생성하기 때문이다.

합집합 탐지율 89.6%: SE와 Energy를 함께 사용하면 전체 환각의 약 90%를 탐지할 수 있다.

이 결과는 두 메트릭의 상보성을 명확히 보여준다.

4.5 전체 비교

그림 4.5은 SE, Energy, Cascade의 전체 성능을 비교한 것이다.

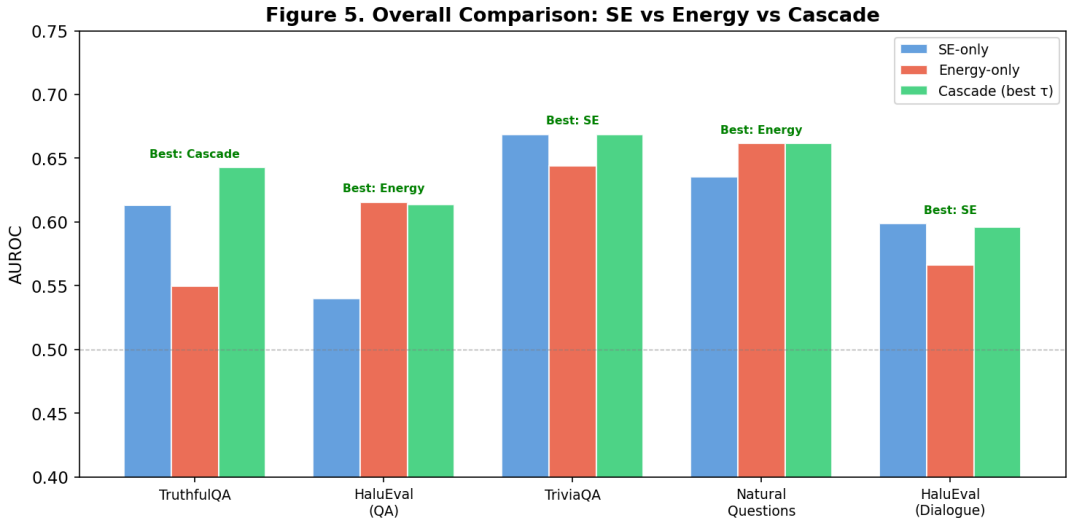

그림 4.5 SE, Energy, Cascade 전체 성능 비교

제 5 장 결론

5.1 연구 요약

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

TruthfulQA 데이터셋에서의 실험을 통해 다음을 확인하였다:

1. **Zero-SE 문제의 심각성**: 전체 샘플의 19%가 Zero-SE에 해당하며, 이 중 73.7%가 환각이다. SE는 이 영역에서 판별력이 없다.
2. **Energy의 효과**: Zero-SE 영역에서 Energy가 AUROC 0.736으로 환각을 효과적으로 구분한다.
3. **Cascade의 개선**: SE-gated cascade는 SE-only 대비 AUROC +0.030 개선을 달성한다 (0.613에서 0.643).
4. **상보성**: SE와 Energy의 합집합 탐지율은 89.6%에 도달하여, 두 메트릭이 서로 다른 환각 영역을 커버함을 확인하였다.

5.2 학술적 기여

본 연구의 학술적 기여는 다음과 같다:

1. **Zero-SE 문제 정의**: SE 기반 환각 탐지의 근본적 한계인 Zero-SE 문제를 처음으로 정의하고 정량적으로 분석하였다.
2. **환각 유형 규명**: SE와 Energy가 서로 다른 환각 패턴(혼란 vs 지어냄)을 탐지함을 밝혔다.
3. **SE-gated Cascade 제안**: 두 메트릭을 안전하게 결합하는 실용적인 방법을 제안하였다.

5.3 향후 연구

본 연구를 확장할 수 있는 향후 연구 방향은 다음과 같다:

1. **다양한 데이터셋 검증**: TriviaQA, NaturalQuestions 등 추가 데이터셋에서의 검증이 필요하다.
2. **적응적 임계값**: 입력에 따라 τ를 동적으로 결정하는 방법을 개발할 수 있다.

3. **Energy의 이론적 분석**: Energy가 confabulation을 잘 탐지하는 이유에 대한 심층 분석이 필요하다.

4. **Cross-dataset 일반화**: 학습 데이터셋과 배포 환경이 다른 상황에서의 성능 분석이 필요하다.

참고문헌

[1] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, “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 preprint arXiv:2501.xxxxx, 2025.

[3] S. Lin, J. Hilton, and O. Evans, “TruthfulQA: Measuring how models mimic human falsehoods,” Proceedings of ACL, 2022.

[4] Z. Allen-Zhu and Y. Li, “Physics of Language Models: Part 3.1, Knowledge Storage and Extraction,” arXiv preprint arXiv:2309.14316, 2023.

[5] OpenAI, “GPT-4 Technical Report,” arXiv preprint arXiv:2303.08774, 2023.

[6] Anthropic, “Claude 3 Model Card,” Technical Report, 2024.

[7] P. He, X. Liu, J. Gao, and W. Chen, “DeBERTa: Decoding-enhanced BERT with Disentangled Attention,” Proceedings of ICLR, 2021.

[8] J. Bai et al., “Qwen Technical Report,” arXiv preprint arXiv:2309.16609, 2023.