가우시안 프로세스를 활용한 RAG 프레임워크 환각 검출 및 불확실성 추정

백승민 1
1 고려대학교 SW · AI 융합대학원 석사과정

baek2570@korea.ac.kr

Gaussian Process-based Hallucination Detection and Uncertainty Estimation in Retrieval-Augmented Generation

Seung-min Baek.1
1Dept. of Artificial Intelligence Convergence, Korea University

요 약

대형언어모델(LLM)은 다양한 과제에서 우수한 성능을 보였으나, 사실과 다른 응답을 생성하는 환각(hallucination) 문제가 지속적으로 제기되어 왔다. 이를 보완하기 위해 제안된 Retrieval-Augmented Generation(RAG) 프레임워크 또한 환각을 완전히 방지하지 못한다는 점이 보고되었다. 본 연구에서는 이러한 문제를 해결하기 위해 가우시안 프로세스(GP)로 구현한 환각 검출 모델을 제안한다. 제안한 모델은 RAG 파이프라인의 생성 이후 단계에서 동작하며, 질의–문맥–응답으로부터 임베딩, 의미 유사도, 응답 길이 등의 특징을 추출하여 GP 기반 분류기에 입력한다. GP 는 이 특징 벡터를 기반으로 환각 여부를 판별하고, 동시에 불확실성을 정량화 한다. GP 는 예측 확률을 사후분포의 기대 값으로 정의하여 신뢰성 있는 불확실성 추정을 가능하게 하며, 희소 GP 적용 및 딥커널 특징 추출기 결합을 통해 계산 효율성을 확보할 수 있다. RAGTruth 데이터셋(QA Task)을 활용한 실험 결과, GP 모델은 MC Dropout 및 Deep Ensemble 대비 Brier Score 와 NLL 에서 낮은 값을 보여 불확실성 정량화 관점의 대안이 될 수 있음을 확인하였다. 이는 제안한 접근이 RAG 시스템의 신뢰성을 높일 수 있는 응용 기능에 활용될 수 있음을 시사한다.

1. 서론

최근 대형언어모델(LLM)은 다양한 과제에서 뛰어난 성능을 보이지만, 종종 응답에 사실과 다른 잘못된 정보를 생성하는 환각(hallucination) 문제를 일으킨다[1]. 이를 완화하기 위해 Retrieval-Augmented Generation (RAG) 프레임워크가 도입되었다[2]. RAG 는 사전학습 된 생성모델에 외부 지식 검색을 결합하여 부족한 지식을 보완하므로 지식집약적 질의응답에서 높은 성능을 달성한다[2]. 그러나 RAG 환경에서도 LLM 이 여전히 근거 없는 비사실적 응답을 생성하는 사례가 보고되었다[3]. 예를 들어, Niu 등은 RAGTruth 데이터셋을 통해 RAG 기반 LLM 응답에서도 여전히 근거 없는 응답이 빈번함을 확인했고[3], Hu 등은 불완전한 지식 검색 결과가 LLM 출력을 왜곡하여 환각 현상이 지속적으로 발생한다고 지적했다[4]. 또한, 정상적인 검색 과정을 통해 충분한 맥락이 제공되더라도 작은 규모의 LLM 은 종종 환각을 생성하거나 응답을 거부하는 경향이 있음이 보고되었다[5]. 이러한 배경에서 본 연구는 RAG 프레임워크 하에서 LLM 응답의 환각 여부를 판별하고, 그 판단의 신뢰도를 불확실성 추정을 통해 제시하는 방법을 제안한다. 본 연구의 목표는 단순히 환각 판별 정확도를 높이는 것에 머무르지 않고, 예측 불확실성을 정량화 하여 활용할 수 있는 모델을 개발하는 데 있다. 이를 위해, 특징 기반 외부 판별기(feature-based classifier) 접근을 택하고, 가우시안 프로세스(GP)를 활용하여 환각 여부와 불확실성을 함께 추정하는 모델을 제안한다.

2. 선행연구 조사

LLM 환각을 식별하기 위한 기존 접근법은 크게 세 가지 유형으로 살펴볼 수 있다.

2.1 LLM 내부 신호 활용

먼저 LLM 의 토큰 확률, 로그 확률, 은닉 상태와 같은 내부 신호를 이용하여 환각을 탐지하는 접근이

시도되었다. 이는 출력 확률 분포를 직접 분석하거나, 내부 확률과 함께 응답의 일관성을 확인하는 방식이다. 이러한 방법은 별도의 판별기를 두지 않아 구성이 간단하고 참조 문서 없이 활용 가능하다는 장점이 있다. 그러나 폐쇄형 LLM(OpenAI GPT 등)에서는 내부 확률이나 hidden layer 에 접근할 수 없으므로 적용이 어렵다는 한계가 있다. 따라서 이 방식은 개방형 모델 환경에 국한된다는 실용적 제약이 존재한다[1].

2.2 반복 답변 의미차이 분석

동일한 질의와 컨텍스트 대해 LLM 답변을 여러 번 샘플링하여 응답 간 의미적 차이를 분석하는 방식이다. Farquhar 등이 제안한 Semantic Entropy 는 다중 응답 간 의미적 차이를 기반으로 환각을 검출하는 대표적인 기법으로, 응답 일관성을 환각 탐지 지표로 활용하였다[1]. 이러한 방법은 반복 응답 생성을 요구하므로 계산 비용이 크다는 단점이 있다.

2.3 외부 판별기 개발

별도의 판별기를 학습시켜 환각 여부를 분류한다.

- LLM 기반 판별기: Zheng 등은 LLM-as-a-Judge 패러다임을 제안하여 LLM 을 평가자로 활용할 수 있음을 보였다[6]. 최근 연구들은 이 접근이 환각 검출 및 응답 평가에서 SOTA 수준의 성능을 달성하고 있음을 보고하고 있다[7]. 그러나 Tian 등은 이러한 LLM 기반 판별기가 과도한 확신 문제를 보이며, 신뢰도 보정이 필요하다고 지적하였다[8].
- 특징 기반 분류기: 응답-문맥 간 의미 유사도, 검색 점수, 증거 커버리지 등 다양한 피처를 활용하여 환각 여부를 분류하는 접근이다. Hu 등이 제안한 LRP4RAG 는 레이어별 관련도 정보를 피처로 활용하여 설명가능성과 판별력을 강화하였다[4]. 특징 기반 판별기는 앞서 언급된 다양한 신호를 통합할 수 있고, 설명가능성과 교정 용이성 측면에서 강점을 갖는다.

3. 연구 방법론

3.1 모델 구성

본 연구는 LLM 내부 신호나 반복 응답 분석 방식 대신, 특징 기반 외부 판별기를 채택한다. 이는 검색 점수, 질문-문맥-응답 간 의미 유사도 등 RAG 프레임워크에서 얻을 수 있는 다양한 신호를 결합할 수 있으며, 외부에서 모델을 교정할 수 있다는 장점이 있다. 특히 LLM 기반 판별기가 다양한 분야에서 SOTA 성능을 보이고 있음에도 불구하고, 높은 계산 비용과 과신 문제를 동반하므로 실용적 한계가 존재한다. 따라서 보다 설명가능하고 교정 가능한 대안으로 특징 기반 판별기를 연구 대상으로 선택한다.

이러한 외부 판별기를 구성함에 있어 본연구는 가우시안 프로세스의 활용을 제안한다. GP 분류기의 예측 확률은 단순 점 추정이 아닌 사후분포(posterior distribution)의 기대 값으로 정의되며, 이는 softmax 기반 신경망 출력과 본질적으로 다르다[9]. 이 특성 덕분에 GP 는 동일한 예측 확률이라도 베이지안적으로 정당화된 신뢰성을 제공한다. 반면 딥러닝 분류기는 높은 정확도를 보이면서도 softmax 확률이 과도한 확신을 나타내는 경우가 있고, 신뢰도 보정(calibration)에는 취약한 것으로 보고되었다[10]. 이에 비해 GP 는 더 잘 보정된(calibrated) 예측 분포를 제공한다는 연구가 존재하며[11], 이러한 특성은 환각 검출 과제에서 불확실성을 정량화 하는데 적합하다.

하지만 GP 는 본질적으로 $O(N^3)$의 계산 복잡도를 가지므로 대규모 데이터 적용에 어려움이 존재한다. 이를 극복하기 위해 Titsias 는 변분 유도변수 기법을 제안하여 소수의 유도점(inducing points)을 사용해 희소(sparse) GP 근사를 가능하게 하였다[12]. 이러한 방식을 통해 대규모 데이터셋 활용 시 GP 의 베이지안적 성질을 유지하면서 학습 효율성을 개선할 수 있다. 또한 Deep Kernel Learning(DKL)은 딥러닝 기반 특징 추출기와 GP 를 결합하여 GP 의 불확실성 추정 능력을 유지하면서도 고차원 입력을 효과적으로 학습할 수 있도록 고안된 방법이다[13]. 이러한 효율화 기법들을 적용하여 GP 의 계산 복잡도를 낮춤으로써, RAG 환경에서의 특징 기반 환각 검출 모델로의 적용 가능성을 높이고자 한다.

커널의 종류와 커널의 하이퍼파라미터 공간을 탐색하며 실험을 반복한 결과 본 연구의 데이터셋에서는 마테른 커널을 사용한 딥커널 모델이 확률 품질(NLL, Brier Score)과 선택적 응답(AURC)에서 우수한 성능을 보임에 따라 이를 활용한 최종 모델을 구성하였다.

본 연구에서 제안하는 GP 기반 환각 판별기는 RAG 파이프라인의 생성 이후 단계에서 동작한다. 질의–문맥–응답에 대한 임베딩과 의미 유사도, 응답 길이 등 특징을 추출하여 입력 벡터를 구성하고, 이를 딥커널-기반 희소 GP 분류기에 입력한다. GP 는 입력된 특징 벡터에 대해 사후분포 기반 예측 확률을 산출하며, 이는 단순 점추정이 아닌 분포의 기대 값으로서 잘 보정된 불확실성 정보를 제공할 수 있다고 가정한다. 이를 증명하기 환각 여부를 판별하고, 예측 확률의 품질을 정량적으로 검증하였다.

3.2. 데이터셋

본 연구의 가정을 평가하기 위해 RAGTruth 데이터셋을 활용한다[3]. RAGTruth 데이터셋에는 다양한 LLM 을 포괄하는 약 18,000 개의 RAG 출력 응답에 대한 환각 여부가 정교하게 라벨링 되어있다. 여러 도메인의 RAG 케이스에 대해 사람이 직접 환각 토큰을 식별함으로써 환각 검출 모델의 학습 및 평가에 적합하다. 다만 본 연구에서는 RAGTruth 데이터셋 중 QA Task 만을 사용하였다. Summary 및 Data-to-Text Task 는 모든 데이터의 쿼리가 동일하여 질문-문맥-응답 간 관계적 신호를 포착하기 어렵기 때문에 연구의 취지와 맞지 않는다고 판단하여 제외하였다. QA Task 는 개별 질문과 참조 문맥, 그리고 생성 응답 사이의 의미적 연결성을 평가할 수 있다. (Train 데이터 5034 건, Test 데이터 900 건)

독립 변수의 구성은 다음과 같다. (총 3084 차원)

(i) 질문(query), 문맥(context), 응답(answer) 각각의 임베딩 벡터 (BGE-M3 임베딩 모델 활용, D=1024)

(ii) 질문, 문맥, 응답 벡터 간 코사인 유사도

(iii) 문장 길이, 토큰 중복 비율, 불확실 단어 비율, 길이 비율 등 9 차원 메타 피처

3.3. 비교 모델 및 평가 지표

본 연구의 핵심 목표는 단순히 환각 여부를 이진적으로 판별하는 데 그치지 않고, 그 판별 결과에 대한 불확실성 수준까지 정량적으로 제시할 수 있는 시스템을 구축하는 데 있다. 예를 들어, 모델이 특정 응답을 환각으로 판정하면서 이 판정에 대한 불확실성이 높다는 신호를 함께 제공할 수 있다면, 사용자는 모델의 결과를 맹목적으로 수용하기보다 그 신뢰도를 고려하여 의사결정을 보완할 수 있다. 이러한 응용 기능은 LLM 기반 판별기로는 제공하기 어려운 관점이라고 할 수 있다.

따라서 본 연구는 LLM 기반 판별기와의 직접적인 환각 식별 성능 경쟁을 지향하지 않는다. 오히려 예측 확률이 사후분포의 기대 값으로 정의되는 가우시안 프로세스(GP)의 성질을 활용하여, 환각 판별 결과와 더불어 그 신뢰도를 함께 제시할 수 있는 방법론을 제안한다. 이는 단순한 정확도 향상 이상의 의미를 가지며, RAG 환경에서 신뢰성 있는 AI 응용을 가능하게 하는 실용적인 연구 질문에 답하려는 시도로 볼 수 있다.

이에 본 연구는 모델의 불확실성 표현력을 비교하기 위해, GP 기반 분류기와 함께 딥 앙상블과 MC 드롭아웃 기반 모델을 비교모델로 설정한다. Lakshminarayanan 등은 딥 앙상블을 통해 예측 불확실성을 매우 양질의 수준으로 추정할 수 있음을 보였으며, 이는 기존 베이지안 모델과 견줄 만큼 성능이 우수하다고 보고했다[14]. 또한 Gal 과 Ghahramani 는 MC 드롭아웃이 딥러닝 모델에서 베이지안 근사 추론을 수행함으로써 신경망 모델에 낮은 비용으로 불확실성 추정 기능을 부여할 수 있음을 증명했다[15]. 따라서 두 방법 모두 예측 확률의 불확실성을 정량화 할 수 있는 적합한 비교 대상이라고 할 수 있다.

이들 모델이 예측한 확률분포의 질을 평가하기 위해 본 연구에서는 음의 로그 가능도(NLL)와 브리어 점수(Brier Score)를 주된 성능 지표로 사용한다. NLL 과 브리어 점수는 예측 확률의 품질을 측정하는 대표적인 지표로, 예측 확률과 실제 라벨의 부합도를 평가하는 연구에 널리 사용된다[16].

4. 연구 결과

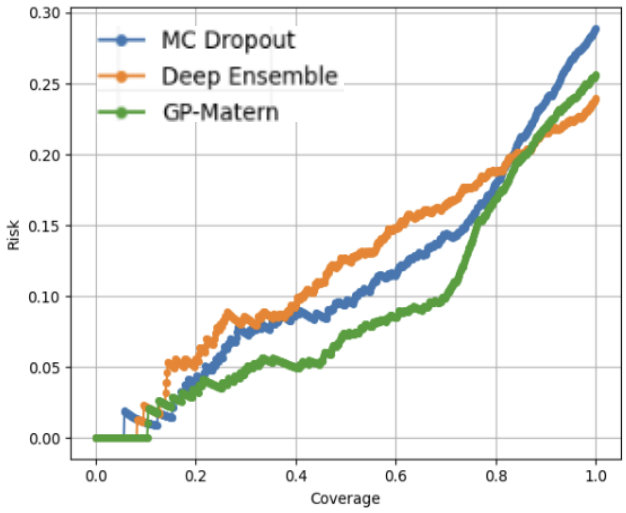
세 모델의 성능을 분류 성능 지표(AUROC), 확률 품질 지표(NLL, Brier Score), 그리고 선택적 응답 지표(AURC)를 통해 평가하였다. 결과는 다음 표와 같다.

Model	AUROC	NLL	Brier Score	AURC
MC Dropout	0.746	0.565	0.165	0.105
Deep Ensemble	0.728	0.747	0.186	0.121
GP-Matern	**0.759**	**0.41**	**0.129**	**0.092**

<표 1> 평가 결과 비교

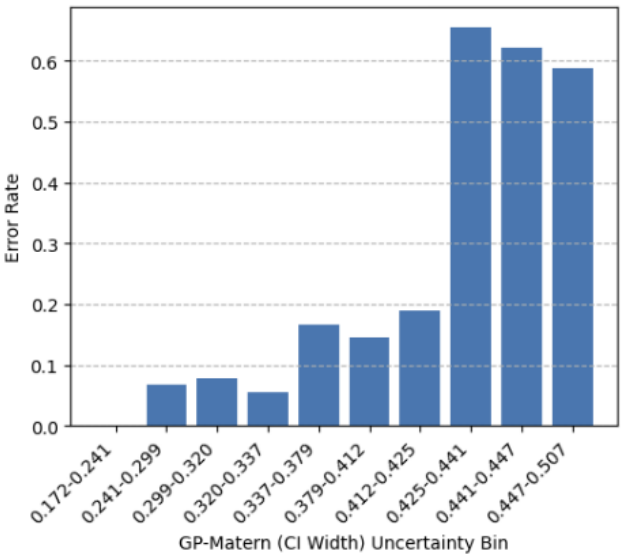
확률 품질: GP-Matern 은 NLL 과 Brier Score 모두에서 가장 낮은 값을 기록하여, 확률 추정이 가장 잘 보정되었음을 보여준다. 이는 예측 확률이 실제 정답 확률을 더 정확하게 반영했음을 의미한다.

선택적 응답 효과: AURC 역시 GP-Matern 이 가장 낮아, 모델이 신뢰도 기반으로 응답을 선택할 때 위험을 효과적으로 줄일 수 있음을 확인하였다. 또한 그림 1 과 같이 Risk-Coverage Curve 상에서도 답변 포기 비율에 따른 Risk 의 감소폭이 세 모델 중 가장 크게 확인되었다.



(그림 1) Risk-Coverage Curve

불확실성-오류 상관: 불확실성 구간별 오류율을 분석한 결과, GP-Matern 은 그림 2 와 같이 불확실성이 높을수록 실제 오류율이 증가하는 명확한 패턴을 보여주었다. 이는 GP 기반 불확실성이 예측 실패 신호로 기능할 수 있음을 뒷받침한다.



(그림 2) 불확실성 구간별 오류율

5. 결론 및 후속 연구

GP-Matern 모델은 본 연구의 핵심 지표인 NLL, Brier Score, AURC 에서 모두 비교 모델을 능가하였다. 이는 가우시안 프로세스가 제공하는 사후분포 기반 예측 확률이 정확도가 높고 신뢰 가능한 확률 분포임을 실험적으로 확인한 결과이다. 이를 통해 GP 기반 접근은 환각 판별 결과뿐 아니라 그 판별의 불확실성 수준까지 함께 제공할 수 있어, 사용자가 모델의 출력을 신뢰도와 함께 해석할 수 있는 새로운 가능성을 제시한다. 이러한 특성은 기존 LLM 기반 판별기가 가진 과신 문제를 보완하며, RAG 환경에서 안전하고 신뢰할 수 있는 선택적 응답 시스템을 구축하는데 기여할 수 있다.

후속 연구에서는 본 연구에서 확인된 가우시안 프로세스(GP)의 불확실성 추정 능력을 활용하여 Active Learning 기반의 Positive Feedback Loop(PFL)를 연구하고자 한다. RAG 프레임워크 LLM 환각 탐지 과제에서 GP 모델의 불확실성 추정 능력은 능동적 표본 선택에 직접 활용 가능하다. Houlsby 등은 GP 분류기를 기반으로 불확실성이 큰 데이터를 선별적으로 라벨링함으로써 라벨링 비용을 줄이고 학습 효율을 크게 향상시킬 수 있음을 입증하였다[17]. 후속 연구에서는 이러한 전략을 RAG 환경의 환각 탐지 파이프라인에 접목하여, 불확실성이 높은 응답을 반복적으로 수집 · 라벨링 · 재학습하고 그 효과를 검증할 예정이다. 이를 통해 RAG 프레임워크 환각 검출에 GP 기반 방법론을 접목하는 본 연구의 실용적 가치를 강화할 수 있을 것으로 기대한다.

참고문헌

[1] S. Farquhar, J. Kossen, L. Kuhn, et al., “Detecting hallucinations in large language models using semantic entropy,” Nature, vol. 630, no. 8017, pp. 625–630, 2024.

[2] P. Lewis, E. Perez, A. Piktus, et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 9459–9474, 2020.

[3] C. Niu, Y. Wu, J. Zhu, S. Xu, K. Shum, R. Zhong, J. Song, T. Zhang, “RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models,” arXiv preprint, arXiv:2401.00396, 2024.

[4] H. Hu, C. He, X. Xie, Q. Zhang, “LRP4RAG: Detecting Hallucinations in Retrieval-Augmented Generation via Layer-wise Relevance Propagation,” arXiv preprint, arXiv:2408.15533, 2024.

[5] H. Joren, J. Zhang, C.-S. Ferng, D.-C. Juan, A. Taly, C. Rashtchian, “Sufficient Context: A New Lens on Retrieval Augmented Generation Systems,” arXiv preprint, arXiv:2411.06037, 2024.

[6] L. Zheng, W.-L. Chiang, Y. Sheng, et al., “Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena,” in NeurIPS Datasets and Benchmarks Track, 2023. (arXiv:2306.05685)

[7] Y. Chen, et al., “Lynx: An Open Source Hallucination Evaluation Model,” arXiv preprint, arXiv:2407.08488, 2024.

[8] Z. Tian, Z. Han, Y. Chen, et al., “Overconfidence in LLM-as-a-Judge: Diagnosis and Confidence-Driven Solution,” arXiv preprint, arXiv:2508.06225, 2025.

[9] C. E. Rasmussen and C. K. I. Williams, Gaussian Processes for Machine Learning, Cambridge, MA: MIT Press, 2006.

[10] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On Calibration of Modern Neural Networks,” in Proc. ICML, PMLR vol. 70, pp. 1321–1330, 2017.

[11] S. T. W. Myren and E. C. Lawrence, “A comparison of Gaussian processes and neural networks for computer model emulation and calibration,” Statistical Analysis and Data Mining: The ASA Data Science Journal, vol. 14, no. 6, pp. 606–623, 2021.

[12] M. K. Titsias, “Variational Learning of Inducing Variables in Sparse Gaussian Processes,” in AISTATS, PMLR vol. 5, pp. 567–574, 2009.

[13] A. G. Wilson, Z. Hu, R. Salakhutdinov, and E. P. Xing, “Deep Kernel Learning,” in AISTATS, pp. 370–378, 2016.

[14] B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 6402–6413, 2017.

[15] Y. Gal and Z. Ghahramani, “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning,” in Proc. ICML, pp. 1050–1059, 2016.

[16] J. Z. Liu, Z. Lin, S. Padhy, D. Tran, T. Bedrax-Weiss, and B. Lakshminarayanan, “Simple and Principled Uncertainty Estimation with Deterministic Deep Learning via Distance Awareness,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 4903–4915, 2020.

[17] N. Houlsby, F. Huszár, Z. Ghahramani, and M. Lengyel, “Bayesian Active Learning for Classification and Preference Learning,” arXiv preprint, arXiv:1112.5745, 2011.