



LLM Hallucination

Ji et al., 2023. “**Survey of Hallucination in Natural Language Generation**” (ACM Computing Surveys) → phân loại hallucination thành **intrinsic** / **extrinsic**, liệt kê hướng giải quyết.

Chen et al., 2024. “**A Survey of Hallucination in Large Language Models**” → update mới, nhiều phương pháp giảm hallucination

Giới thiệu bài toán

Phát hiện và phân loại ảo giác thông tin trong phản hồi tiếng Việt của LLM.

Mục tiêu: Xác định xem phản hồi (Output) của các mô hình ngôn ngữ lớn (LLMs) tiếng Việt có chứa thông tin sai lệch (ảo giác) so với ngữ cảnh đầu vào (Context) hay không, và nếu có thì thuộc loại ảo giác nào.

Input:

- **Context:** Một đoạn văn bản chứa thông tin.
- **Prompt:** Một câu hỏi hoặc yêu cầu được đưa ra cho LLM
- **Response:** Phản hồi được tạo ra bởi LLM để trả lời Prompt dựa trên Context.

Output:

- **Label:** Nhận được gán cho biết loại ảo giác

context:

Ngành công nghiệp xe hơi của Nhật Bản là một trong những ngành lớn nhất thế giới, với các thương hiệu nổi tiếng như Toyota, Honda.

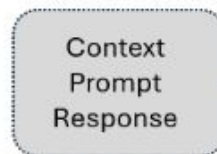


prompt

Ngành công nghiệp xe hơi của Nhật Bản có những thương hiệu nào?

response

Ngành công nghiệp xe hơi của Nhật Bản có những thương hiệu nổi tiếng như Toyota, Honda và Mazda.



Is Hallucination?



Bộ dữ liệu ViHallu

Bộ dữ liệu phân loại ảo giác thông tin trong phản hồi tiếng Việt của LLM.

id: Mã định danh duy nhất cho mẫu dữ liệu

context: Một đoạn văn bản chứa thông tin.

prompt: Câu hỏi được đặt ra cho LLM. Prompt có ba dạng chính

- Factual: Prompt bình thường, nguyên vẹn
- Noisy: Cố ý chèn lỗi chính tả/ký tự, đảo từ, . . .
- Adversarial: Thêm thông tin đánh lạc hướng, hỏi “bẫy”

response: Phản hồi được tạo ra bởi LLM để trả lời Prompt dựa trên Context.

label: Nhãn được gán cho biết loại ảo giác

- NO (Không ảo giác): Phản hồi hoàn toàn nhất quán và đúng sự thật với thông tin trong context.
- INTRINSIC (Ảo giác nội tại): Phản hồi **mâu thuẫn trực tiếp** hoặc **bóp méo** thông tin đã có trong context.
- EXTRINSIC (Ảo giác ngoại tại): Phản hồi **bổ sung** thông tin **KHÔNG CÓ** trong context.

The image features a dark navy blue background. In the upper-left corner, there are two overlapping geometric shapes: a blue parallelogram and a light green parallelogram, both tilted at an angle. The text 'EDA' is centered in the middle-right portion of the image in a white, sans-serif font.

EDA

Shape: (7000, 5)

	id	context	prompt	response	label
0	9b1ea51d-d1ff-45ba-8cf1-6a91328e8600	Vào những năm 1870, hai nhà điêu khắc Augustus Saint-Gaudens và Daniel Chester French sinh sống và làm việc gần Quảng trường. Đến những năm 1920, Công viên Quảng trường Washington được công nhận c...	Vào những năm 1960, nơi nào trở thành trung tâm của thể hệ âm nhạc beat và folk, mặc dù Jackson Pollock và Willem de Kooning đã tổ chức nhiều buổi triển lãm âm nhạc nổi tiếng tại đó?	Quảng trường Washington là trung tâm của thể hệ âm nhạc beat và folk vào những năm 1960. Ngoài ra, nơi đây từng là địa điểm tổ chức các buổi triển lãm âm nhạc nổi tiếng như "Festival of Sound" do ...	extrinsic
1	db7a89c6-2a6a-42af-beef-58e557ecc819	Cách mạng Tháng Mười đã biến một cuộc chiến tranh để quốc thành một cuộc đấu tranh chống lại sự đàn áp, đánh dấu sự khởi đầu của một phong trào quần chúng vì chủ nghĩa xã hội ở các nơi khác nhau c...	Cách mạng Tháng Mười đã khởi xướng chủ nghĩa phát xít, một biến thể của chủ nghĩa đế quốc, đúng hay sai?	Sai. Cách mạng Tháng Mười đã đánh bại chủ nghĩa phát xít, một biến thể phản động của chủ nghĩa đế quốc, và không khởi xướng nó.	no
2	10fca062-d343-4eca-8434-93c7a8aa5e0e	Vị trí địa lý nằm giữa phương Đông và Địa Trung Hải giúp cho người Thổ Nhĩ Kỳ hoàn toàn kiểm soát các tuyến đường mậu dịch lớn, và có một môi trường lý tưởng cho thực vật và động vật phát triển. Ấ...	Dựa trên việc Đế quốc Ottoman không bao giờ kiểm soát đường mậu dịch lớn, bạn có thể cho biết khi nào nền ẩm thực của Thổ Nhĩ Kỳ được xây dựng?	Nền ẩm thực của Thổ Nhĩ Kỳ được xây dựng từ giữa thập niên 1400, chịu ảnh hưởng của các đế quốc xung quanh như Byzantine và Ba Tư. Sự giao thoa văn hóa này đã tạo ra các món ăn đa dạng và phong ph...	extrinsic

```
=== Info ===
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7000 entries, 0 to 6999
Data columns (total 5 columns):
#   Column      Non-Null Count  Dtype
---  ---
0    id          7000 non-null   object
1    context     7000 non-null   object
2    prompt      7000 non-null   object
3    response    7000 non-null   object
4    label       7000 non-null   object
dtypes: object(5)
memory usage: 273.6+ KB
None

=== Missing values per column ===
id          0
context     0
prompt      0
response    0
label       0
dtype: int64

Unique labels: ['extrinsic', 'intrinsic', 'no']
```

Notebook

Input

+ Add Input

📁 Upload

DATASETS

- 📁 llmhallucination
- 📄 vihallu-train.csv

Output (72KiB / 19.5GiB)

📁 /kaggle/working

Table of contents

- 📄 Vietnamese LLM Hallucination — EDA N
- ✅ What to look for (checklist)
- ➡️ Next

Session options

ACCELERATOR
GPU P100

Quota: 00:15 / 45 hrs

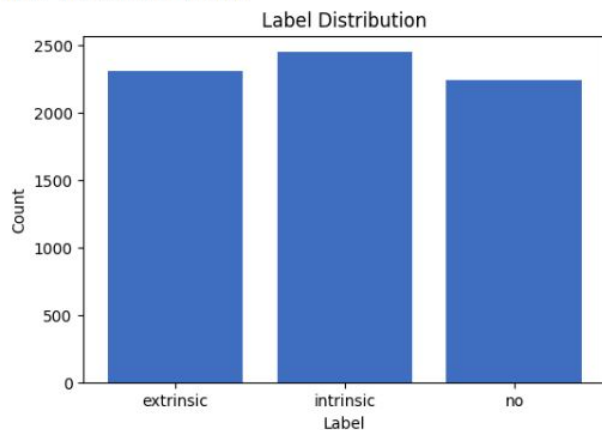
LANGUAGE
Python

PERSISTENCE

Label Distribution

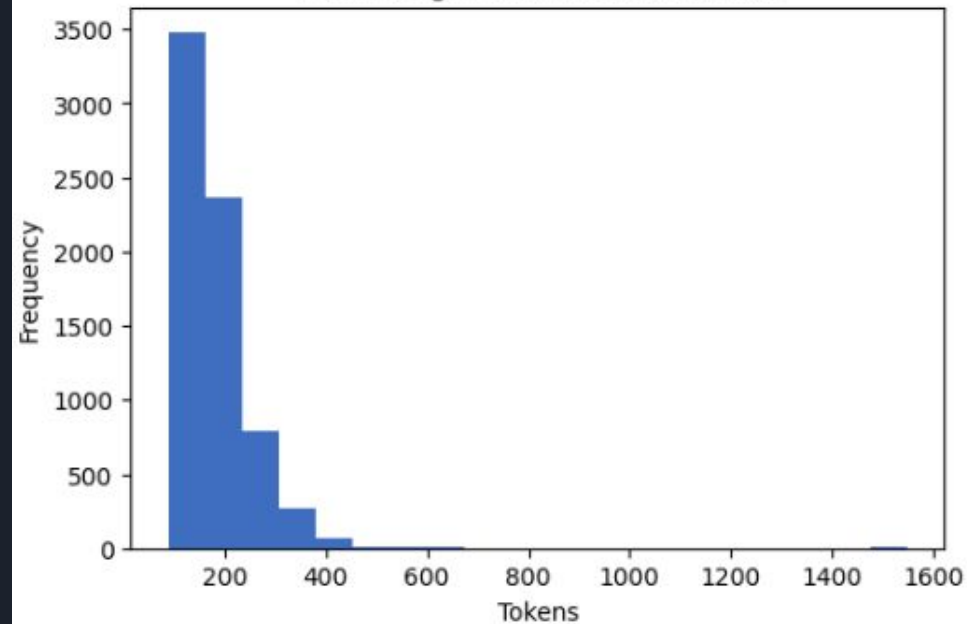
```
Label counts:  
label  
extrinsic    2307  
intrinsic    2448  
no            2245  
Name: count, dtype: int64
```

```
Label proportions:  
label  
extrinsic    0.3296  
intrinsic    0.3497  
no           0.3207  
Name: count, dtype: float64
```

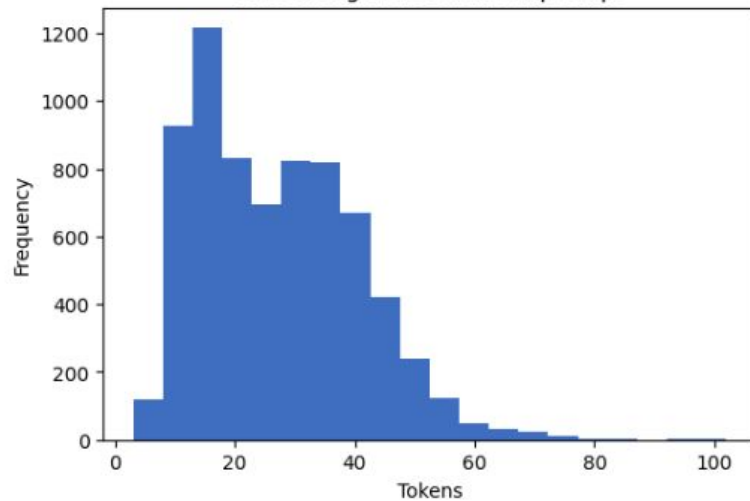


Length

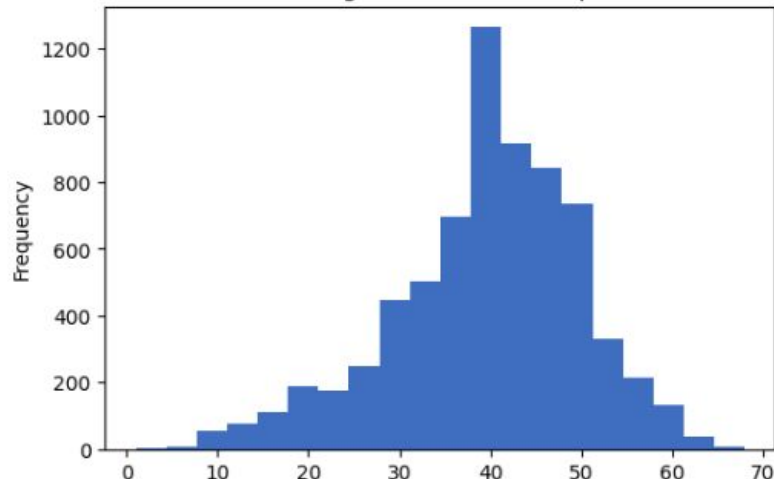
Token length distribution: context



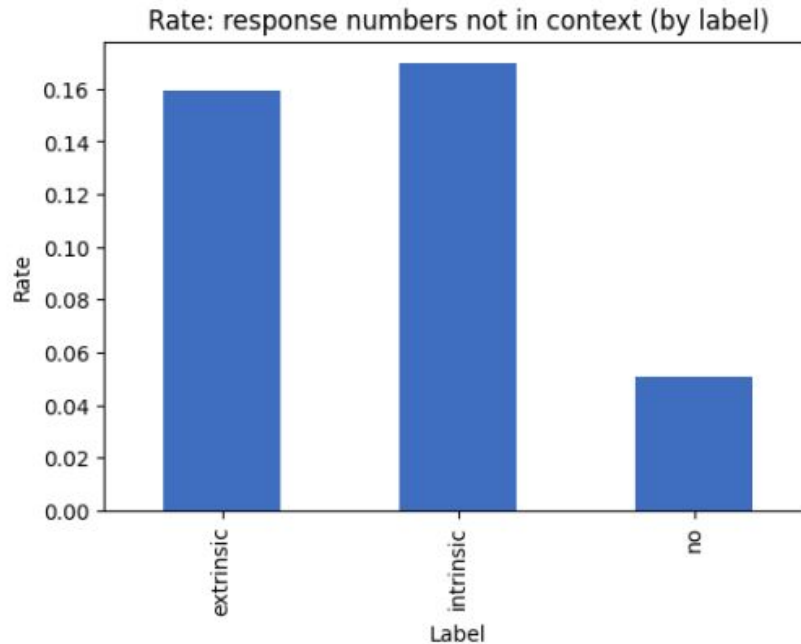
Token length distribution: prompt



Token length distribution: response

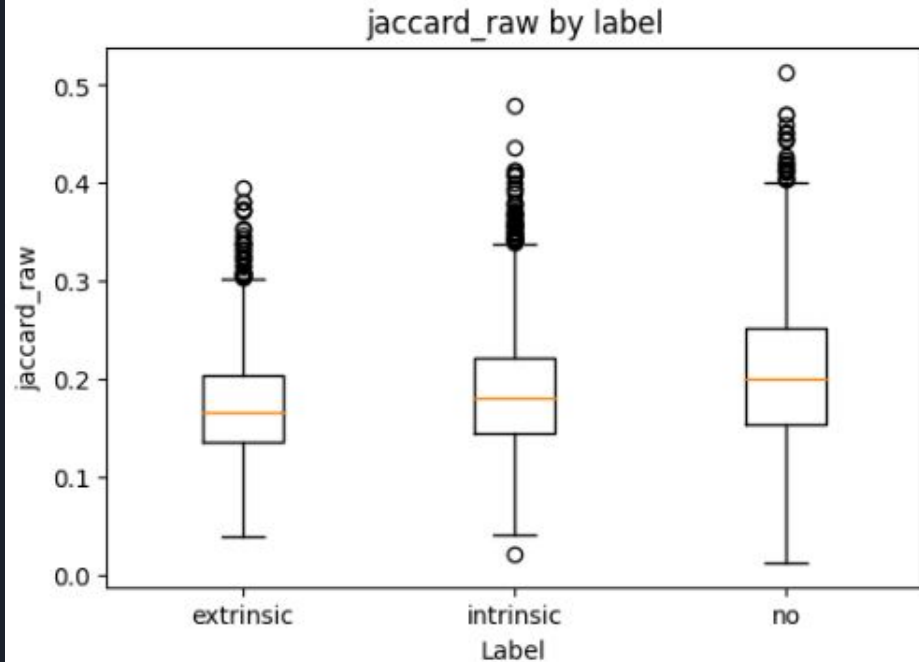


Number between response and context



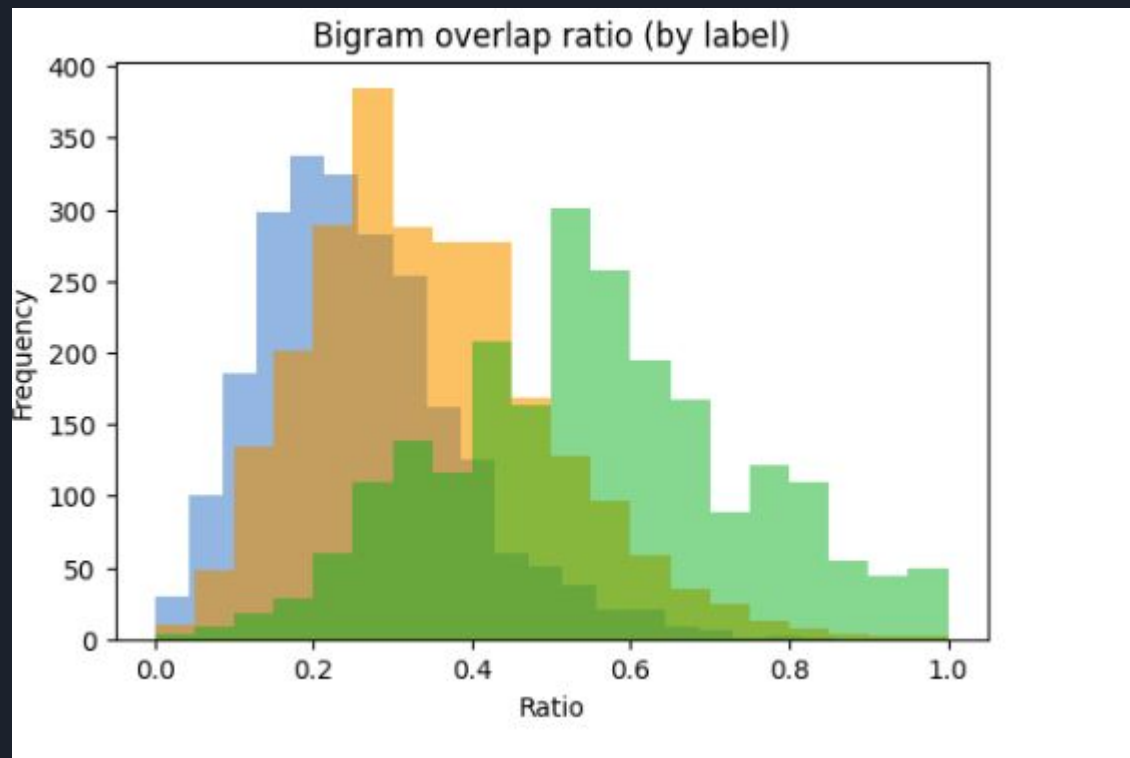
Overlap between context and response

	jaccard_raw	jaccard_noacc	overlap_ratio	label
0	0.181347	0.203390	0.686275	extrinsic
1	0.216981	0.225490	0.920000	no
2	0.112500	0.141892	0.400000	extrinsic
3	0.182692	0.193878	1.000000	no
4	0.212389	0.233645	0.585366	intrinsic



Bigram overlap

	bigram_overlap_ratio	label
0	0.454545	extrinsic
1	0.600000	no
2	0.204082	extrinsic
3	0.947368	no
4	0.348837	intrinsic





Approach



Classifier-based (Encoder-only / NLI style)

Coi task là **phân loại** (classification) dựa trên (Context, Prompt, Response).

- **Cách làm:**
 - Encode (Context, Response) → dự đoán entail / contradict / neutral → ánh xạ ra `no` / `intrinsic` / `extrinsic`.
 - Dùng backbone encoder như **BERT**, **RoBERTa**, **DeBERTa**, **XLNet**, **PhoBERT**.
- **Ưu điểm:** đơn giản, hiệu quả, dễ train với nhãn chuẩn.
- **Nhược điểm:** cần đủ data nhãn; khó generalize sang domain khác.



Classifier-based (Encoder-only / NLI style)

Paper:

- **FactCC** – Kryscinski et al., ACL 2020. “Evaluating the Factual Consistency of Abstractive Text Summarization” → dùng BERT để phân loại câu summary có factual hay không.
- **TRUE** – Honovich et al., ACL 2022. “TRUE: Re-evaluating Factual Consistency Evaluation” → benchmark nhiều classifier cho factual consistency.



Generative QA → Verify (RAG / NLI / Self-check)

Dùng chính LLM hoặc retriever để **kiểm chứng lại** câu trả lời.

- **Cách làm:**
 - Cho LLM generate → rồi chạy một **verifier**:
 - **NLI** (entail/contradict/neutral).
 - **SelfCheckGPT** (cùng LLM sinh nhiều lần, đo consistency).
 - **Semantic Entropy** (độ bất định trong phân phối token).
 - Nếu unsupported hoặc không nhất quán → label `extrinsic`.
- **Ưu điểm:** không cần data nhãn nhiều, tận dụng sức mạnh LLM nhỏ.
- **Nhược điểm:** pipeline phức tạp, cần tune threshold.



Generative QA → Verify (RAG / NLI / Self-check)

Paper:


- **SelfCheckGPT** – Manakul et al., TMLR 2023. “SelfCheckGPT: Zero-resource Hallucination Detection for Generative Large Language Models” → kiểm chứng bằng chính LLM.
- **Semantic Entropy** – Kuhn et al., Nature 2023. “Semantic entropy: a measure of hallucinations in generative models” → đo bất định ngữ nghĩa để phát hiện hallucination.
- **Chain-of-Verification (CoVe)** – Dziri et al., 2023. “Faithful Chain-of-Thought Reasoning” → LLM tự đặt câu hỏi phụ để kiểm chứng.



RAG-based Evaluation (Retrieval Augmented)

Xem câu trả lời có bám vào **chứng cứ retrieve được** hay không.

- **Cách làm:**
 - Dùng retriever (BM25 + dense embedding) lấy evidence từ Context.
 - So sánh overlap / entail giữa evidence và Response.
 - Dùng framework như **RAGAS** (Faithfulness, Context precision/recall).
- **Ưu điểm:** dễ đo tính “faithful” khi có corpus.
- **Nhược điểm:** phụ thuộc chất lượng retriever.



RAG-based Evaluation (Retrieval Augmented)

Paper:

- **RAGAS** – Es et al., 2024. “RAGAS: Automated Evaluation of Retrieval-Augmented Generation” → framework để đo faithfulness, relevancy, recall trong QA.
- **HaluEval** – Li et al., ACL 2023. “HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models” → benchmark detection trên nhiều domain.