Kiểm chứng tự động các tuyên bố về khí hậu bằng dữ liệu thời gian thực SỬ DỤNG mô hình ngôn ngữ lớn

***Trần Nguyễn Quốc Thắng 1, Lê Nguyễn Nhật Tân 1, Quang Công Văn1, Lê Thanh2,***

***Tiếu Phùng Mai Sương1\*, Trần Khải Thiện3***

1Khoa Công nghệ Thông tin, Trường Đại học Ngoại Ngữ Tin Học Thành phố Hồ Chí Minh

2 Khoa Công nghệ Thông tin, Trường Đại học Công Nghệ Thành phố Hồ Chí Minh

2 Khoa Công nghệ Thông tin, Trường Đại học Công Thương Thành phố Hồ Chí Minh

Email: 122dh113431@st.huflit.edu.vn, 122dh114725@st.huflit.edu.vn, 122dh114809@st.huflit.edu.vn, 2lthanh25nct@hutech.edu.vn, [1suongtpm@huflit.edu.vn](mailto:1suongtpm@huflit.edu.vn), 3thientk@huit.edu.vn

# TÓM TẮT

Việc xác định chính xác thông tin khí hậu thật và giả trong thời đại số là vô cùng quan trọng, vì thông tin sai lệch có thể ảnh hưởng nghiêm trọng đến nhận thức công chúng và việc hoạch định chính sách. Bài báo này đề xuất một phương pháp kiểm chứng sự thật tự động các tuyên bố về khí hậu, sử dụng mô hình ngôn ngữ lớn GPT-4 kết hợp chiến lược tranh luận. Phương pháp tích hợp dữ liệu thời gian thực được thu thập từ NewsAPI các nguồn uy tín như NASA và NCEI. Khung này kết hợp nhiều quan điểm khoa học, phân tích cảm xúc và ý định của tuyên bố để cung cấp đánh giá đáng tin cậy và nhạy bén với ngữ cảnh. Hệ thống không chỉ phát hiện thông tin sai lệch theo thời gian thực mà còn hỗ trợ đối thoại thông tin về biến đổi khí hậu, nhấn mạnh vai trò của AI trong các cuộc thảo luận và chính sách môi trường với dữ liệu đáng tin cậy, đồng thời cảnh báo về nguy cơ lan truyền thông tin sai lệch dựa trên cảm xúc và ý định.

*Từ khoá*: kiểm chứng sự thật; mô hình ngôn ngữ lớn; dữ liệu thời gian thực; tuyên bố khí hậu; tranh biện đa tác nhân

1. GIỚI THIỆU

Thông tin sai lệch về biến đổi khí hậu trên mạng xã hội và báo chí là thách thức lớn, trong đó AI được xem là công cụ hỗ trợ tiềm năng. Nghiên cứu cho thấy chatbot AI có thể giúp định hình nhận thức công chúng theo đúng đồng thuận khoa học, nhấn mạnh yêu cầu đảm bảo độ chính xác và tin cậy của các mô hình ngôn ngữ lớn (LLMs). Trong bài báo này, chúng tôi phát triển TruthSeeker, hệ thống kiểm chứng các tuyên bố về khí hậu dựa trên lập luận khoa học, nguồn tin đáng tin cậy, dữ liệu thời gian thực và phân tích cảm xúc ngữ cảnh. Các nền tảng như Climate Feedback đã nỗ lực đáng khen ngợi khi huy động các nhà khoa học khí hậu tình nguyện đóng góp chuyên môn, cung cấp dịch vụ thiết yếu để xử lý thông tin sai lệch về khí hậu. Những nhà khoa học này dành thời gian để đưa ra các đánh giá dựa trên khoa học, bao gồm tài liệu tham khảo và phán quyết cuối cùng về các tuyên bố gây tranh cãi. Dù đóng góp quý giá, những nỗ lực này gặp khó khăn lớn về khả năng mở rộng và tính cập nhật. Do đó, tác động của chúng bị giới hạn bởi khối lượng khổng lồ thông tin sai lệch và sự hoài nghi trên các phương tiện truyền thông số, nơi thông tin sai lệch lan truyền nhanh và rộng hơn so với thông tin chính xác.

Nhằm đáp ứng thách thức này, chúng tôi đề xuất một khung kiểm tra sự thật tích hợp LLMs vào mô hình Người kiểm tra–Người đưa ra yêu cầu nhằm xử lý các tuyên bố phức tạp, đặc biệt trong lĩnh vực biến đổi khí hậu. Hệ thống sử dụng các Claimant riêng biệt dựa trên nguồn dữ liệu khác nhau để đại diện quan điểm đa chiều, trong khi Verifier tổng hợp, đặt câu hỏi bổ sung và phân tích cảm xúc, ý định nhằm đưa ra phán quyết cân bằng và kiểm soát nguy cơ lan truyền sai lệch. Các đóng góp chính của nghiên cứu bao gồm:

* Khung kiểm chứng tích hợp LLM và dữ liệu thời gian thực;
* Cơ chế tranh luận đa vòng tăng minh bạch, giảm thiên vị;
* Phân tích cảm xúc và ý định để đánh giá rủi ro lan truyền;
* Hệ thống xử lý tốt các tuyên bố phức tạp, vượt qua giới hạn kiểm chứng thủ công.

Bố cục bài báo được trình bày như sau: Mục 2 trình bày về các nghiên cứu liên quan, mục 3 giới thiệu về phương pháp và kết quả, ở mục 4 chúng tôi tiến. hành các kết luận và hướng nghiên cứu tiếp theo.

1. TỔNG QUAN NGHIÊN CỨU LIÊN QUAN

Các LLMs ngày càng được ứng dụng trong kiểm chứng sự thật, nhưng hiện tượng “ảo giác” – mô hình tạo ra thông tin sai nhưng nghe có vẻ hợp lý – vẫn là thách thức lớn [1]. Đặc biệt, trong các lĩnh vực như khí hậu, y tế và chính sách công, sai lệch thông tin do LLM gây ra có thể để lại hậu quả nghiêm trọng [2], [3]. Một số nỗ lực đáng chú ý nhằm giảm hallucination là huấn luyện mô hình bằng phản hồi con người (RLHF). Ouyang và cộng sự [4] chứng minh RLHF cải thiện độ tuân thủ và độ tin cậy, nhưng vẫn cần các cơ chế bổ sung. Một hướng tiếp cận hiệu quả là tranh luận đa tác nhân, nơi các LLMs tương tác và phản biện lẫn nhau. Liu và cộng sự [5], [6] đề xuất TruEDebate – khung tranh luận nhiều vòng giúp phát hiện tin sai và nâng cao tính giải thích. Trong lĩnh vực khí hậu, Leippold và cộng sự [2] phát triển Climinator, tích hợp dữ liệu NASA/NCEI vào một hệ thống đa tác nhân để thẩm định phát biểu về biến đổi khí hậu. Song song, nhiều công trình đề xuất tăng cường tri thức ngoại cho LLM qua truy hồi (RAG). Jiang và cộng sự [7] giới thiệu Active-RAG, cho phép mô hình truy xuất có kiểm soát trong khi sinh nội dung, giúp giảm lỗi thực tế. Trong y tế, Gilbert và cộng sự [8] kết hợp knowledge graph với LLM nhằm hạn chế thông tin bịa đặt trong hỗ trợ lâm sàng. Một nhánh nghiên cứu khác tập trung vào nâng cao năng lực lập luận nội tại. Chain-of-Thought (CoT) [4] và Tree-of-Thought (ToT) [9] giúp mô hình tự xây dựng các bước suy luận. Dhuliawala và cộng sự [10] đề xuất Chain-of-Verification (CoVe), trong đó LLM tự đặt câu hỏi và kiểm chứng câu trả lời, làm giảm đáng kể tỷ lệ ảo giác. Cuối cùng, các đánh giá thực nghiệm cho thấy ngay cả LLMs mạnh vẫn dễ mắc lỗi. Tang và cộng sự [3] và Asgari và cộng sự [11] phát hiện lỗi nghiêm trọng trong tóm tắt y khoa, từ đó đề xuất công cụ CREOLA để giám sát và hiệu chỉnh đầu ra. DeVerna và cộng sự [12] cảnh báo người dùng có thể bị sai lệch nếu chỉ dựa vào giải thích do AI sinh ra..

Chính trong bối cảnh đó, nghiên cứu này đề xuất một hệ thống kiểm chứng sự thật tự động tích hợp các đặc điểm còn thiếu: khả năng thích ứng với các lĩnh vực chuyên biệt, tích hợp dữ liệu thời gian thực, kết hợp phân tích cảm xúc và ý định để đánh giá rủi ro lan truyền, và vận hành trên khung tranh luận giữa các tác nhân LLMs có nhiệm vụ riêng biệt. Bằng cách này, hệ thống nhằm lấp đầy khoảng trống giữa hiệu quả kiểm chứng trên dữ liệu tĩnh với nhu cầu phản hồi nhanh, minh bạch, và có khả năng suy luận phức tạp trên môi trường thực tế.

1. PHƯƠNG PHÁP ĐỀ XUẤT
   1. Mô hình đề xuất

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, số

Nội dung do AI tạo ra có thể không chính xác.

1. Sơ đồ mô hình TruthSeeker

Hình 1 mô tả quy trình vận hành của mô hình TruthSeeker bắt đầu bằng việc thu thập dữ liệu từ NewsAPI, sử dụng các từ khóa như “climate change nasa” và “climate change noaa”. Dữ liệu được giới hạn trong 30 ngày gần nhất, lấy từ các nguồn có uy tín như NASA và NCEI (tính đến ngày 6/7/2025). Toàn bộ văn bản sau khi thu thập được chia thành các đoạn nhỏ có độ dài chuẩn hóa (chunk\_size=1000, chunk\_overlap=200) bằng công cụ RecursiveCharacter­TextSplitter. Các đoạn này tiếp tục được mã hóa sử dụng HuggingFaceEmbeddings (all-MiniLM-L6-v2) và lưu trữ trong cơ sở dữ liệu Chroma nhằm phục vụ việc truy xuất ngữ cảnh một cách nhanh chóng, hỗ trợ hiệu quả cho quá trình kiểm chứng.

Bên cạnh đó, mô hình còn tích hợp phân tích cảm xúc và ý định vào quá trình đánh giá. Cảm xúc được xác định thông qua mô hình DistilRoBERTa với các nhãn như neutral, anger và fear, trong khi ý định được phân loại bằng BART-Large-MNLI với các nhãn như informative, misleading và controversial. Những thông tin này được đưa vào bối cảnh đánh giá của các Claimant gồm Support, Denier và GPT4 cũng như Verifier, nhằm giúp các tác nhân hiểu rõ hơn về giọng điệu và mục đích của tuyên bố. Chẳng hạn, trong trường hợp một tuyên bố như “*Scientists have ignored the influence of solar energetic particles, cosmic rays, interplanetary magnetic fields, and a weakening Earth's magnetic field to conclude that climate change is hu­man activity*”, khi được phân tích là mang ý định misleading với độ tin cậy khoảng 0.747, GPT-4 sẽ ưu tiên tham chiếu đến đồng thuận khoa học từ NASA và đưa ra phán quyết “not credible”. Dữ liệu cảm xúc và ý định cũng được dùng để xác định mức độ rủi ro của tuyên bố. Khi cảm xúc thể hiện sự giận dữ hoặc sợ hãi với độ tin cậy vượt 0.5, hoặc khi ý định được xác định là misleading hoặc controversial, mô hình sẽ gán nhãn mức độ rủi ro là “high”. Điều này cho phép hệ thống đưa ra cảnh báo kịp thời đối với những tuyên bố có nguy cơ cao lan truyền thông tin sai lệch, hỗ trợ các nhà khoa học hoặc cơ quan chức năng phản ứng sớm và phù hợp.

Trong giai đoạn đánh giá nội, từng Claimant đảm nhận vai trò riêng biệt: Claimant Support dựa trên dữ liệu NASA để tìm kiếm bằng chứng ủng hộ; Claimant Denier sử dụng nguồn dữ liệu từ NCEI để tìm kiếm các luận điểm phản bác; trong khi đó GPT-4 bổ sung góc nhìn trung lập từ tri thức nội tại. Các đánh giá từ ba Claimant này được Verifier tổng hợp thành phán quyết sơ bộ, sau đó được Neutral Verifier xem xét và đưa ra kết luận cuối cùng dưới định dạng JSON với các khóa thông tin bao gồm Final Verdict, Explanation, Confidence, RiskLevel và Debate. Trung tâm hệ thống vận hành qua cơ chế vòng tranh luận, giúp đánh giá các tuyên bố về biến đổi khí hậu một cách toàn diện và khách quan thông qua đối thoại liên tục giữa các Claimant và Verifier. Khi có mâu thuẫn hoặc thiếu thông tin, Neutral Verifier điều phối, đặt câu hỏi bổ sung để các Claimant cung cấp thêm bằng chứng hoặc làm rõ lập luận. Quá trình này có thể kéo dài tối đa năm vòng nhằm đảm bảo kết luận chính xác, hạn chế phán quyết vội vàng và nâng cao độ tin cậy của hệ thống.

Phương pháp kiểm chứng trên đảm bảo tính khách quan bằng cách kết hợp đồng thời dữ liệu thời gian thực, tri thức nội tại của mô hình ngôn ngữ GPT-4 và phân tích đa chiều, hỗ trợ việc xác định chính xác các tuyên bố như “The ozone layer is recovering” hoặc “No real data supports global warming” với mức độ tin cậy cao.

* 1. Phương pháp prompting
     1. **Tổng quan**

Kĩ thuật prompting trong mô hình TruthSeekers được thiết kế để kiểm chứng các tuyên bố về biến đổi khí hậu thông qua tranh luận có cấu trúc giữa các tác nhân Claimants và trọng tài Verifiers. Hệ thống gồm ba Claimants là Support dựa trên tài liệu NASA, Denier dựa trên tài liệu NCEI và GPT-4 dựa trên tri thức nội tại, cùng hai Verifiers thực hiện tổng hợp và đưa ra phán quyết cuối cùng. Các tác nhân được định hướng bởi prompting để đưa ra phán quyết chính xác, dựa trên bằng chứng khoa học và giải quyết mâu thuẫn qua các vòng tranh luận. Prompting đảm bảo phán quyết chất lượng cao, ưu tiên đồng thuận khoa học và hạn chế thông tin sai lệch, tuân theo bốn nguyên tắc: rõ ràng với vai trò và đầu ra cụ thể, khách quan với ưu tiên bằng chứng, linh hoạt xử lý đa dạng tuyên bố và lặp lại để làm rõ mâu thuẫn.

Ví dụ, với tuyên bố "*There has been no reduction in oceanic pH levels in the last century*.", hệ thống sử dụng prompting để yêu cầu các Claimants đưa ra phán quyết ("credible", "not credible", "Not Enough Information") và giải thích dựa trên tài liệu hoặc tri thức nội tại, sau đó Verifiers tổng hợp và đánh giá.

* + 1. **Thiết kế Prompt**
  1. **Prompt cho StatementTask**: Prompt này hướng dẫn phân loại claim theo 2 danh mục.

You are a fact-checking AI assistant specializing in scientific information.

# Your task is to analyze statements and categorize their accuracy using the following categories: {verdictsClimateFeedback}. Each level describes common characteristics and examples to guide your analysis:

---

## Credible

Assign this label if the statement:

- Is supported by strong scientific evidence or aligns with the established scientific consensus.

- May lack some nuance or context, but remains fundamentally accurate and not misleading.

\*\*Examples\*\*:

- \*"Babies under six months should not drink water as it can result in health risks."\*

→ \*\*Credible\*\*. Backed by medical guidelines on infant hydration.

- \*"Prioritizing plant-based foods reduces greenhouse gas emissions."\*

→ \*\*Credible\*\*. Generally supported by environmental research, though effects depend on scale of adoption.

- \*"Climate change will destroy all ecosystems."\*

→ \*\*Credible\*\*, if interpreted with added context. While some ecosystems may adapt, many are at severe risk.

---

## Not Credible

Assign this label if the statement:

- Contradicts scientific consensus or lacks reliable evidence.

- Is based on flawed reasoning, misleading framing, or speculative claims without scientific backing.

- Contains factual inaccuracies that distort public understanding of the issue.

\*\*Examples\*\*:

- \*"CO2 increases are mainly due to natural causes, not humans."\*

→ \*\*Not Credible\*\*. Overwhelming evidence shows fossil fuel emissions are the primary cause.

- \*"Greenland's ice cores show no significant warming, disproving climate change."\*

→ \*\*Not Credible\*\*. Scientific data confirms Greenland’s warming trends.

- \*"The Atlantic is cooling, and scientists don't know why."\*

→ \*\*Not Credible\*\*. Misleading due to cherry-picking short-term anomalies.

---

## Instructions

1. \*\*Analyze\*\* the statement's factual accuracy and scientific reasoning.

2. \*\*Choose one label\*\*: `Credible` or `Not Credible`.

3. \*\*Provide a brief explanation\*\* justifying the label.

4. \*\*If necessary\*\*, identify missing context or evidence and explain why it matters.

* 1. **Prompt cho Support Claimant:** Prompt yêu cầu Support đánh giá claim dựa trên dữ liệu NASA:

You are an AI Advocate, responsible for fact-checking user statements about climate change.

Your expertise draws on a specific set of trustworthy scientific documents, and your responses must be grounded in the evidence provided.

### Expertise

You specialize in:

- Climate change

- Climate science

- Environmental science

- Physics

- Energy science

- Science communication

### Objective

Your main objective is to verify the accuracy of user statements on climate change by examining the evidence in your assigned documents.

### Crucial Instructions

\* \*\*Evidence is Essential\*\*: Your responses \*must\* be directly supported by the information in your assigned documents.

\* \*\*Cite Sources\*\*: Always reference 'Reference', 'Page', and 'URL' when citing evidence.

\* \*\*Avoid Speculation\*\*: If there is not enough information to assess a statement, state "Not Enough Information" and explain what specific information is missing.

\* \*\*Follow StatementTask Instructions\*\*: Use the verdict categories provided in StatementTask: {statement\_task}, and ensure your evaluation aligns with those instructions.

\* \*\*Consider Sentiment and Intent\*\*: Note if the statement's sentiment or intent suggests a potential to spread misinformation.

### Assessment Process

1. \*\*Evaluate the Statement\*\*

- Analyze based on expertise and evidence in assigned documents.

- Use verdict categories from StatementTask.

- Determine if evidence supports, contradicts, or lacks information.

- Consider sentiment and intent for potential misinformation impact.

2. \*\*Optional: Break Down into Substatements\*\*

- Divide complex statements into smaller substatements.

- Evaluate each substatement individually.

3. \*\*Synthesize an Overall Verdict\*\*

- Consider accuracy, importance, and sentiment/intent.

- Assign an overall verdict based on synthesis.

\*\*Output Format\*\*:

{

"Final Verdict": "",

"Explanation": ""

}

If not enough information:

{

"Final Verdict": "Not Enough Information",

"Explanation": "I do not have enough information to assess this statement. [Explain missing-information.]"

}

* 1. **Prompt cho Denier Claimant:** Prompt yêu cầu Denier đánh giá tuyên bố từ góc độ hoài nghi dựa trên dữ liệu NCEI.

You are an AI Denier Advocate, critically evaluating user statements about climate change from a skeptical perspective.

Your expertise is strictly limited to the specific set of documents provided.

### Objective

Scrutinize the accuracy of user statements based solely on assigned documents.

### Crucial Instructions

\* \*\*Use Only Provided Documents\*\*: Responses must be based exclusively on document content.

\* \*\*Strict Evidence Requirement\*\*: Respond with "Not Enough Information" if documents lack sufficient data.

\* \*\*Cite Sources Precisely\*\*: Provide 'Reference', 'Page', and 'URL'.

\* \*\*Avoid External Knowledge\*\*: Do not use general scientific principles unless in documents.

\* \*\*Follow StatementTask Instructions\*\*: Use verdict categories from StatementTask: {statement\_task}.

\* \*\*Consider Sentiment and Intent\*\*: Note misinformation potential.

### Assessment Process

1. \*\*Evaluate Based on Documents\*\*

- Assess strictly within document scope.

- Use StatementTask verdict categories.

- Consider sentiment and intent.

2. \*\*Optional: Break Down into Substatements\*\*

- Separate and evaluate individually.

3. \*\*Final Verdict for Simple Statements\*\*

- Assess as a whole if straightforward.

\*\*Output Format\*\*:

{

"Final Verdict": "",

"Explanation": ""

}

If not enough information:

{

"Final Verdict": "Not Enough Information",

"Explanation": "The documents provided do not contain enough information to assess this statement. [Explain missing-information.]"

}

* 1. **Prompt cho GPT4 Claimant:** Prompt yêu cầu GPT4 đánh giá claim dựa trên kiến thức nội tại.

You are an AI Claimant, responsible for fact-checking user statements about climate change.

Your expertise is based on general scientific knowledge, and your responses must reflect the consensus in climate science.

### Objective

Verify the accuracy of user statements on climate change.

### Crucial Instructions

\* \*\*Avoid Speculation\*\*: If there is not enough information, state "Not Enough Information" and explain what is missing.

\* \*\*Follow StatementTask Instructions\*\*: Use verdict categories from StatementTask: {statement\_task}.

\* \*\*Consider Sentiment and Intent\*\*: Note if the statement's sentiment or intent suggests misinformation potential.

### Assessment Process

1. \*\*Evaluate the Statement\*\*

- Analyze based on scientific consensus.

- Use verdict categories from StatementTask.

- Consider sentiment and intent.

2. \*\*Optional: Break Down into Substatements\*\*

- Divide complex statements and evaluate individually.

3. \*\*Synthesize an Overall Verdict\*\*

- Assign based on accuracy and sentiment/intent.

\*\*Output Format\*\*:

{

"Final Verdict": "",

"Explanation": ""

}

If not enough information:

{

"Final Verdict": "Not Enough Information",

"Explanation": "I do not have enough information to assess this statement. [Explain missing-information.]"

}

* 1. **Prompt cho Verifier:** Prompt yêu cầu Verifier tổng hợp đánh giá từ các Claimants (Supporter, Denier, và GPT4).

Role: Authoritative "Verifier" System

Primary Objective: Synthesize assessments from Claimants to determine the veracity of a user's statement on climate change, prioritizing scientific rigor.

### Expertise

You specialize in:

- Climate change

- Climate science

- Environmental science

- Physics

- Energy science

- Science communication

### Responsibilities

1. Review verdicts and explanations from Claimants.

2. Consolidate into a final verdict, prioritizing evidence-based assessments.

3. Seek clarification through follow-up questions if discrepancies arise.

4. Prioritize Claimants with specific, high-quality evidence.

5. Consider sentiment and intent for misinformation potential.

### Final Assessment Criteria

1. Analyze collective Claimant assessments.

2. Do not rely on majority voting; prioritize evidence quality.

3. Assess sentiment and intent impact on misinformation.

### Guidelines

1. Prioritize Claimants with concrete evidence over those with "Not Enough Information."

2. Ask follow-up questions if Claimants provide conflicting evidence or if additional evidence is needed.

3. Stop debate if no Claimant changes assessment after follow-up or if evidence is sufficient.

4. Cite 'Reference', 'Page', and 'URL' when referring to Claimant data.

5. Assess risk level (High, Medium, Low) based on sentiment/intent.

\*\*Output Format\*\*:

{

"Final Verdict": "",

"Explanation": "",

"Confidence": "high|medium|low",

"RiskLevel": "high|medium|low",

"Debate": {

"Status": "Open|Closed",

"Round": {round\_number},

"Follow-Up Questions": {}

}

}

* 1. **Prompt cho Neutral Verifier:** Prompt yêu cầu Neutral Verifier tổng hợp đánh giá từ Claimants và Verifier.

Role: Neutral "Verifier" System

Primary Objective: Synthesize assessments from Claimants and Verifier to determine the final veracity of a user's statement, ensuring impartiality.

### Responsibilities

1. Review verdicts and explanations from Claimants and Verifier.

2. Consolidate into a final verdict, prioritizing evidence-based assessments.

3. Seek clarification through follow-up questions if discrepancies arise.

4. Prioritize judgments with specific evidence.

5. Consider sentiment and intent for misinformation potential.

### Final Assessment Criteria

1. Analyze collective assessments.

2. Do not rely on majority voting; prioritize evidence quality.

3. Assess sentiment and intent impact on misinformation.

### Guidelines

1. Prioritize Claimants/Verifier with concrete evidence.

2. Ask follow-up questions if conflicting evidence exists or if additional evidence is needed.

3. Stop debate if no changes occur after follow-up or if evidence is sufficient.

4. Cite 'Reference', 'Page', and 'URL' when referring to data.

5. Assess risk level (High, Medium, Low) based on sentiment/intent.

\*\*Output Format\*\*:

{

"Final Verdict": "",

"Explanation": "",

"Confidence": "high|medium|low",

"RiskLevel": "high|medium|low",

"Debate": {

"Status": "Open|Closed",

"Round": {round\_number},

"Follow-Up Questions": {}

}

}

1. THỰC NGHIỆM
   1. Tập dữ liệu

Tập dữ liệu thực nghiệm được thu thập từ trang web của tổ chức phi lợi nhuận Climate Feedback[[1]](#footnote-2). Tập dữ liệu bao gồm 170 câu tuyên bố được thu thập từ nhiều nguồn khác nhau đã được xác minh và chú thích kỹ lưỡng một cách chặt chẽ bởi các chuyên gia khí hậu.

* 1. Môi trường thực nghiệm

Trong bài báo này chúng tôi thực nghiệm dựa trên các môi trường và thư viện sau:

* Môi trường: Google Colab, tích hợp API NewsAPI.
* Mô hình: GPT4, Roberta-base, Bart-large.
* Thư viện hỗ trợ: Python, LangChain, HuggingFace Transformers, Underthesea, Chroma, Requests.
  1. Các chỉ số đánh giá
* Precision đo lường độ chính xác của các dự đoán thuộc một lớp cụ thể. Precision cho biết tỷ lệ các dự đoán này thực sự đúng. Precision cao nghĩa là mô hình ít nhầm lẫn khi gắn nhãn cho các nguồn không đáng tin cậy.

Trong đó, TP là số mẫu được dự đoán là Credible và thực sự là Credible; FP là số mẫu được dự đoán là Credible nhưng thực tế là Not Credible.

* Recall đo lường khả năng mô hình tìm ra tất cả các mẫu thực sự thuộc một lớp. Recall thấp nghĩa là mô hình bỏ sót nhiều nguồn đáng tin cậy.

Với FN là số nguồn thực sự là Credible nhưng bị dự đoán là Not Credible.

* F1-Score là trung bình điều hòa giữa Precision và Recall, được sử dụng để đánh giá sự cân bằng giữa hai chỉ số này. F1-Score cao cho thấy mô hình vừa chính xác vừa nhạy trong việc xác định nguồn đáng tin cậy hoặc không đáng tin cậy.
* Accuracy đo lường tỷ lệ dự đoán đúng trên toàn bộ tập dữ liệu, bất kể lớp. Trong bối cảnh khí hậu, Accuracy cao nghĩa là mô hình phân loại đúng nhiều nguồn.

Trong đó TN là số mẫu được dự đoán là Not Credible và thực sự là Not Credible và Total Samples là tổng số mẫu.

* Macro Average tính trung bình các chỉ số (Precision, Recall, hoặc F1-Score) của tất cả các lớp mà không xét đến số lượng mẫu của từng lớp.

Với là giá trị Precision, Recall, hoặc F1-Score của lớp thứ I và *n* là số lớp.

* Weighted Average tính trung bình các chỉ số có trọng số dựa trên số lượng mẫu của từng lớp. Weighted Average phản ánh hiệu suất thực tế, ưu tiên lớp chiếm đa số. Điều này quan trọng khi đánh giá mô hình trong các ứng dụng thực tế về khí hậu, nơi thông tin không đáng tin cậy thường xuất hiện nhiều hơn.

Trong đó, là số mẫu của lớp thứ I, *Total Support* là tổng số mẫu.

* 1. Kết quả thực nghiệm

**Bảng 1.** Hiệu suất phân loại của TruthSeeker và GPT-4

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Mô hình** |  | **Credible** | **Not credible** | **Accuracy** | **Macro Avg** | **Weighted Avg** |
| **TruthSeeker**  **(Mô hình đề xuất)** | **Precision** | 0.86 | 0.96 | - | 0.92 | 0.96 |
| **Recall** | 0.67 | 0.97 | - | 0.82 | 0.94 |
| **F1-Score** | 0.75 | 0.97 | 0.94 | 0.86 | 0.95 |
| **Support** | 18 | 152 | 170 | 170 | 170 |
| **GPT-4** | **Precision** | 0.72 | 0.97 | - | 0.85 | 0.95 |
| **Recall** | 0.81 | 0.92 | - | 0.86 | 0.91 |
| **F1-Score** | 0.76 | 0.94 | 0.92 | 0.85 | 0.92 |
| **Support** | 18 | 152 | 170 | 170 | 170 |

Bảng 1 cho thấy TruthSeeker đạt Accuracy 0.94, cao hơn GPT-4 (0.92), phân loại đúng 94% tổng số mẫu. Với lớp Credible, TruthSeeker đạt Precision 0.86 (tốt hơn GPT-4 là 0.72), nhưng Recall thấp hơn (0.67 so với 0.81), dẫn đến F1-Score tương đương (0.75 và 0.76). Với lớp Not Credible, TruthSeeker vượt trội với Precision, Recall và F1-Score đều đạt 0.97, so với GPT-4 là 0.94, cho thấy khả năng phát hiện thông tin sai lệch hiệu quả hơn.

TruthSeeker cũng đạt Macro Average (0.86) và Weighted Average (0.95) cao hơn GPT-4 (0.85 và 0.92), phản ánh hiệu suất toàn diện hơn. Trong bối cảnh biến đổi khí hậu, TruthSeeker phù hợp để kiểm duyệt nội dung nhờ khả năng phát hiện tin sai lệch, trong khi GPT-4 thích hợp cho thu thập dữ liệu khoa học nhờ khả năng nhận diện nguồn đáng tin cậy tốt hơn.

Hiệu quả này đến từ khung tranh luận đa tác nhân và verifier giúp xử lý rõ ràng hai kết quả, kết hợp dữ liệu thời gian thực và phân tích cảm xúc, ý định. Ví dụ, tuyên bố gây hiểu lầm như “Cold weather to grip world...” được phân loại chính xác là Not Credible. Tuy nhiên, các tuyên bố mơ hồ vẫn là thách thức, cần bổ sung dữ liệu hoặc tăng vòng tranh luận để cải thiện.

1. KẾT LUẬN

Nghiên cứu này đã đề xuất và triển khai mô hình TruthSeeker nhằm kiểm chứng tự động các tuyên bố về biến đổi khí hậu, kết hợp dữ liệu thời gian thực, kỹ thuật prompting theo chuỗi suy luận và cơ chế tranh luận đa tác nhân. Hệ thống đạt độ chính xác trung bình 94% trên tập dữ liệu thử nghiệm, cho thấy hiệu quả rõ rệt trong việc phân biệt các phát ngôn đáng tin cậy và sai lệch. Việc tích hợp phân tích cảm xúc và ý định giúp nhận diện các tuyên bố tiềm ẩn nguy cơ gây hiểu lầm, nhất là những phát ngôn mang cảm xúc tiêu cực hoặc đánh lạc hướng nhận thức.

Mặc dù kết quả thực nghiệm thể hiện tiềm năng rõ rệt, mô hình vẫn tồn tại một số hạn chế cần được khắc phục trong các nghiên cứu tiếp theo, cụ thể mô hình vẫn phụ thuộc vào độ chính xác và tính cập nhật của dữ liệu từ NewsAPI, đồng thời gặp khó khăn với các tuyên bố mơ hồ, đa nghĩa. Các hướng phát triển tiếp theo sẽ tập trung mở rộng nguồn dữ liệu tin cậy như IPCC, xây dựng từ điển ngữ nghĩa chuyên biệt về khí hậu cho tiếng Việt, và thử nghiệm triển khai trong các ứng dụng thực tế như chatbot môi trường và phân tích nội dung mạng xã hội. Những nỗ lực này nhằm tăng cường truyền thông khoa học và giảm thiểu rủi ro lan truyền thông tin sai lệch về biến đổi khí hậu.

# TÀI LIỆU THAM KHẢO

1. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P.: Survey of Hallucination in Natural Language Generation. ACM Comput Surv. 55, (2023). https://doi.org/10.1145/3571730

2. Leippold, M., Vaghefi, S.A., Stammbach, D., Muccione, V., Bingler, J., Ni, J., Senni, C.C., Wekhof, T., Schimanski, T., Gostlow, G., Yu, T., Luterbacher, J., Huggel, C.: Automated fact-checking of climate claims with large language models. npj Climate Action 2025 4:1. 4, 1–9 (2025). https://doi.org/10.1038/s44168-025-00215-8

3. Tang, L., Sun, Z., Idnay, B., Nestor, J.G., Soroush, A., Elias, P.A., Xu, Z., Ding, Y., Durrett, G., Rousseau, J.F., Weng, C., Peng, Y.: Evaluating large language models on medical evidence summarization. NPJ Digit Med. 6, 1–8 (2023)

4. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P.F., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst. 35, 27730–27744 (2022)

5. Liu, Y., Liu, Y., Zhang, X., Chen, X., Yan, R.: The Truth Becomes Clearer Through Debate! Multi-Agent Systems with Large Language Models Unmask Fake News. (2025)

6. Darwish, A.M., Rashed, E.A., Khoriba, G.: Mitigating LLM Hallucinations Using a Multi-Agent Framework. Information 2025, Vol. 16, Page 517. 16, 517 (2025). https://doi.org/10.3390/INFO16070517

7. Jiang, Z., Xu, F.F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., Yang, Y., Callan, J., Neubig, G.: Active Retrieval Augmented Generation. EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings. 7969–7992 (2023). https://doi.org/10.18653/V1/2023.EMNLP-MAIN.495

8. Gilbert, S., Kather, J.N., Hogan, A.: Augmented non-hallucinating large language models as medical information curators. NPJ Digit Med. 7, 1–5 (2024)

9. Yao, S., Yu, D., Deepmind, G., Zhao, J., Griffiths, T.L., Cao, Y., Narasimhan, K.: Tree of Thoughts: Deliberate Problem Solving with Large Language Models. Adv Neural Inf Process Syst. 36, 11809–11822 (2023)

10. Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., Weston, J.: Chain-of-Verification Reduces Hallucination in Large Language Models. Proceedings of the Annual Meeting of the Association for Computational Linguistics. 3563–3578 (2024). https://doi.org/10.18653/V1/2024.FINDINGS-ACL.212

11. Asgari, E., Montaña-Brown, N., Dubois, M., Khalil, S., Balloch, J., Yeung, J.A., Pimenta, D.: A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. NPJ Digit Med. 8, 1–15 (2025)

12. DeVerna, M.R., Yan, H.Y., Yang, K.C., Menczer, F.: Fact-checking information from large language models can decrease headline discernment. Proc Natl Acad Sci U S A. 121, e2322823121 (2024)

1. Trang chuyên về đánh giá và xác minh tính chính xác thông tin đại chúng liên quan đến biến đổi khí hậu [↑](#footnote-ref-2)