Original Paper

Self-Logical Consistency Assessment of Large Language Models for Patient Feedback Classification : Algorithm Development and Validation Study

Zeno Loi [1*], David Morquin [1,2], François-Xavier Derzko [3], Xavier Corbier [1,4], Sylvie Gauthier [3], Patrice Taourel [5,6], Emilie Prin-Lombardo [3], Grégoire Mercier [6,7] and Kévin Yauy [1,8].

1 : Espace de Recherche et d'Intégration des Outils numériques en Santé, CHU Montpellier, Montpellier, France

2 : Département des Maladies Infectieuses et Tropicales, CHU Montpellier, Univ Montpellier, Montpellier, France

3 : Pôle soins, qualité, parcours et usagers, CHU Montpellier, Montpellier, France

4 : Institut de Science des Données de Montpellier, Univ Montpellier, INSERM, Montpellier, France

5 : Service d'Imagerie Médicale Lapeyronie, CHU de Montpellier, Montpellier, France

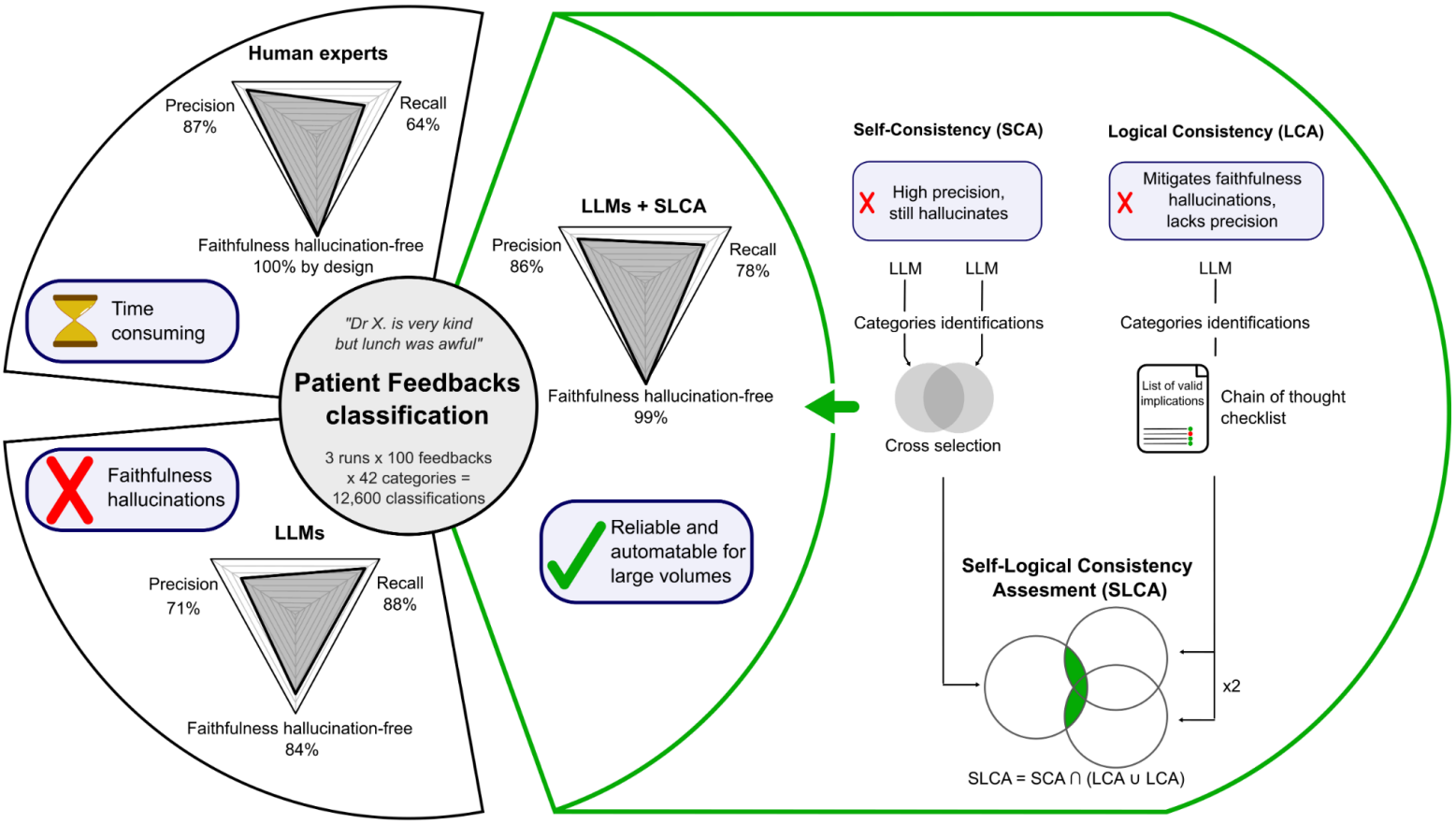
6 : Institut Desbrest d'Épidémiologie et de Santé Publique, Univ Montpellier, INSERM, Montpellier, France

7 : Unité de Recherche Médico-Économique, CHU Montpellier, Montpellier, France

8 : Service de Génétique Médicale, LIRMM, CHU Montpellier, Univ Montpellier, Montpellier, France.

*Corresponding Author Email: z-loi@chu-montpellier.fr, Phone : +33633159353

Graphical Abstract



Keywords

Artificial Intelligence ; Natural Language Processing ; Hallucinations ; Quality of Health Care ; Patient Satisfaction ; Quality Improvement ; Patient Rights ; Classification ; Reproducibility of Results ; Health Services Research ; Benchmarking ; Models, Statistical ; Feedback ; Logic ; Surveys and Questionnaires ; Hospital Administration ; Hospitals, University

Abstract

Background : Patient satisfaction feedback is crucial for hospital service quality, but manual reviews are time-consuming, and traditional natural language processing methods remain inadequate. Large Language Models (LLMs) show promise but are prone to extrinsic faithfulness hallucinations—fabricated or illogical outputs that limit their reliability in healthcare.

Objective: This study aimed to evaluate the Self-Logical Consistency Assessment (SLCA), an original method designed to enhance LLM feedback classification reliability by enforcing a logically-structured chain of thought.

Methods: SLCA uses two validation steps: self-consistency (identifying the most coherent response) and logical consistency (ensuring alignment with the original statement and expert classifications). We evaluated SLCA using GPT-4 and Llama-3.1 405B on 12,600 classifications from 100 patient feedback samples to assess hallucinations, and tested its performance on a 49,140-classification benchmark derived from 1,170 feedbacks.

Results: SLCA reduced hallucinations among detected categories from 15.80% (168/1063) to 0.51% (4/786) with GPT-4 and from 7.17% (51/711) to 1.67% (10/599) with Llama-3.1, with residual errors confined to the emergency feedback category. On the benchmark, SLCA achieved precision-recall scores of 0.86-0.78 for GPT-4 and 0.84-0.58 for Llama-3.1. These results demonstrate SLCA's ability to achieve human-level performance across LLMs.

Conclusions: SLCA offers a scalable, explainable solution for improving LLM classification reliability in healthcare. Its capacity to enhance performance without fine-tuning or additional training data positions it as a valuable tool for analyzing patient feedback and supporting hospital service quality improvement.

Background

Patient satisfaction feedback is a crucial metric for determining areas of improvement in hospital services, directly impacting quality of care [1]. To effectively manage the substantial volume of feedback, it is essential to structure and classify this information to prioritise enhancement efforts. However, the human-led classification process is time-consuming and requires quality of care management skills.

Automated patient feedback analysis has historically been limited by Natural Language Processing (NLP) algorithms. An unsupervised classification of 2.5 million feedbacks provided a comprehensive overview of patient concerns and defined 20 categories [2,3]. However, models like Naive Bayes and BERT struggle to accurately classify nuanced feedback because they cannot effectively handle complex language contexts [4,5].

Large Language Models (LLMs) like GPT-4 and Llama-3.1 405B offer a promising alternative due to their superior ability to understand natural language and detect subtle nuances in patient feedback [5,6,7]. They have been recently applied in exploratory studies for other quality measurement purposes [8]. However, since these models have been shown to exhibit inaccuracies in the medical field [9], there is a pressing need to improve the reliability of text classification. Recent work shows that evaluating the self-consistency of LLM predictions—that is, ensuring the model provides the same classification over multiple independent attempts—greatly enhances performance in classification tasks [10]. However, these models often produce hallucinations, which are factual mistakes or logical flaws in generated text [11]. While factual hallucinations may be acceptable to some extent in a hospital

context, the loss of explainability due to extrinsic faithfulness hallucinations—the generation of illogical thought processes—is incompatible with sensitive applications. Since Chain of Thought (CoT) generation methods improve the explainability of these models [12], validating CoT logical structures might be a potential solution to the issue of hallucinations.

Here we describe and evaluate Self-Logical Consistency Assessment (SLCA), a method designed to improve the reliability of LLM-generated predictions. The SLCA combines Self-Consistency Assessment (SCA) [10] — the evaluation of the LLM ability to provide the same predictions in similar conditions — together with an original method called Logical Consistency Assessment (LCA), which appraises the capability of an LLM to produce a logically structured CoT.

Methods

Data Collection and Samples

1270 Patient feedbacks were collected from the French national feedback system E-Satis, involving adult patients hospitalized at the Montpellier University Hospital Centre between 2022 and 2024. Exclusions included feedback from patients who declined data use, overly lengthy feedback unsuitable for model analysis, compensation claims, and feedback containing extreme content (see Supplementary Figure 1). All selected feedback was pseudonymized to protect personal data.

Two feedback samples were selected:

- Human-Curated Sample (n=100): Selected to represent feedback diversity and to compare human versus GPT-4 consistency assessment methods (Supplementary Table 2). The sample size was the largest feasible for our expert team. Three agents of each type classified this sample to evaluate reproducibility.
- Benchmark Sample (n=1,170): Randomly selected from 2023 to provide a machine learning benchmark on a representative dataset (Supplementary Table 3). The sample size was calculated to detect a 2% difference in precision or recall among models (16 different tests), with $\alpha=.01$ using Bonferroni correction and $1-\beta=.80$. Based on these parameters, 45,600 classifications, thus 1086 feedbacks were needed; we included 1,170 to ensure sufficient power (see Supplementary Table 5).

Categorisation Approach

The classification task involved categorizing feedback into 21 non-exclusive categories and two non-exclusive tones—favourable and unfavourable (Supplementary Note 1). The categories were adapted from the Haute Autorité de Santé (French National Health Authority) guidelines [2,3], with the addition of "Patient's Rights" to meet local healthcare quality improvement goals. The favourable tone describes positive aspects appreciated by patients, while the unfavourable tone indicates criticisms or disliked aspects.

Evaluation metrics

Comparisons focused on:

- Precision: The proportion of relevant items among those selected, essential for medical-grade classification.
- Recall: The proportion of relevant items that were selected.
- Reproducibility: Assessed via agreement between different agents.
- Extrinsic Faithfulness Hallucination Rate: Following prior work [11], hallucinations occur when the LLM generates information not inferable from the original text or includes unjustified logical steps. Hallucinations were assessed by cross-referencing the LLM's output with patient feedback, verifying that all information in the CoT leading to a category identification was explicitly stated or inferable. If the CoT included additional information or logical inconsistencies, it was flagged as containing a hallucination. The hallucination rate is defined here as the percentage of hallucinations among argued category identifications.

The benchmark focused on precision and recall for overall performance; subgroup analyses used the F1 score.

Logical Consistency Methodology

Evaluating logical human argumentation is fundamental in the philosophy of logic [13]. While prior studies structured LLMs' Chains of Thought (CoT) [14,15], none directly applied philosophical methodologies to assess logical validity in LLM-generated CoTs. Our work integrates these methods to enhance the logical consistency of LLM predictions.

An argument consists of:

- **Premise:** A quote from patient feedback.
- **Implication:** The logical link connecting the premise to the conclusion.
- **Conclusion:** A category identification.

An argument is valid if the truth of the premise leads to the conclusion via the implication. Since assessing premises is challenging and conclusions are within a fixed set of 21 categories, we focused on defining valid implications. Three human experts and the Principal Investigator created exhaustive lists of valid implications for each category. This list was provided to the LLM with instructions to produce valid arguments. Logical consistency was evaluated based on the LLM's use of appropriate implications.

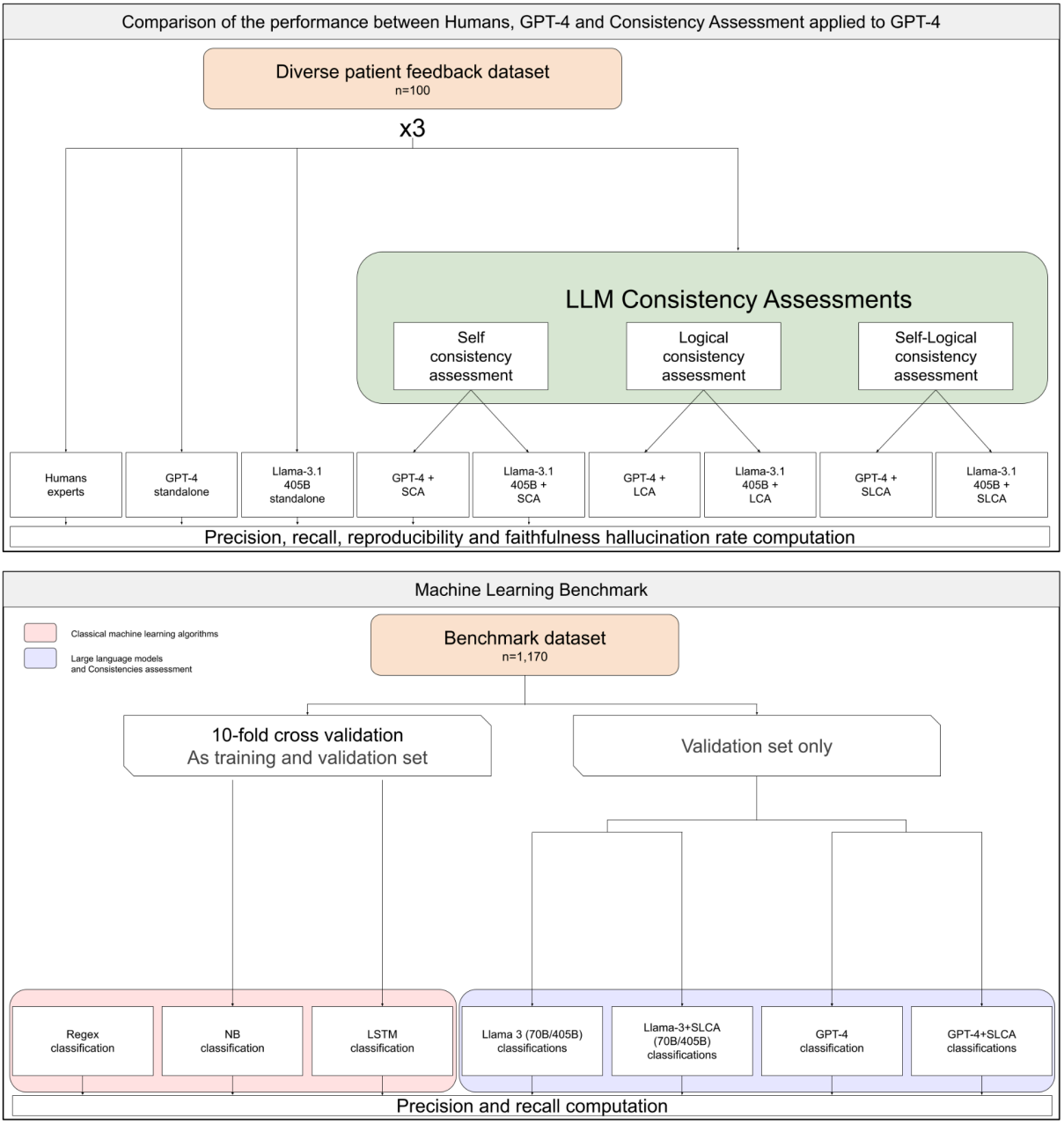


Figure 1. Consistency Assessment Evaluation Study Design

Study Design comparing humans and Consistency assessed LLMs

This section analyses are led on the 100 feedbacks human-curated sample (Figure 1).

- **Human Expert Evaluation:** Three quality-of-care experts independently classified each feedback. No communication occurred between them. A fourth

expert blindly reviewed each classification to establish the gold standard (Supplementary Table 4).

- **LLMs Classification Methods:**

Both GPT-4 turbo (version March 05, 2024) and Llama-3.1 405B were evaluated independently as raw predictive models and in their consistency assessed versions as following :

  - **Standalone LLM :** Three runs of LLMs classified each feedback independently. The prompt included output structure, classification conditions, category list, and instructions to generate a structure-free CoT (Supplementary Note 2). No example was given.
  - **Self-Consistency Assessment (SCA):** Three additional LLMs runs classified each feedback independently. Classifications were grouped in three pairs. For each GPT-4+SCA agent, only categories identified by both runs were retained, aiming to select consistently identified categories (Figure 2).
  - **Logical Consistency Assessment (LCA):** To assess logical consistency, an initial prediction was produced without CoT structure restrictions. A second prompt instructed the LLM to structure its previous output into a logical argument (Supplementary Note 2), using implications from the predefined list (Supplementary Note 1). For each LLM+LCA agent, only classifications with a valid implication were retained (Figure 2).
  - **Combined Self-Logical Consistency Assessment (SLCA):** Both self and logical methods were applied, selecting identifications fulfilling two conditions: both runs identified the category, and at least one produced

a valid CoT. This process evaluates the LLM's logical consistency and self-consistency, enhancing precision and exhaustivity.

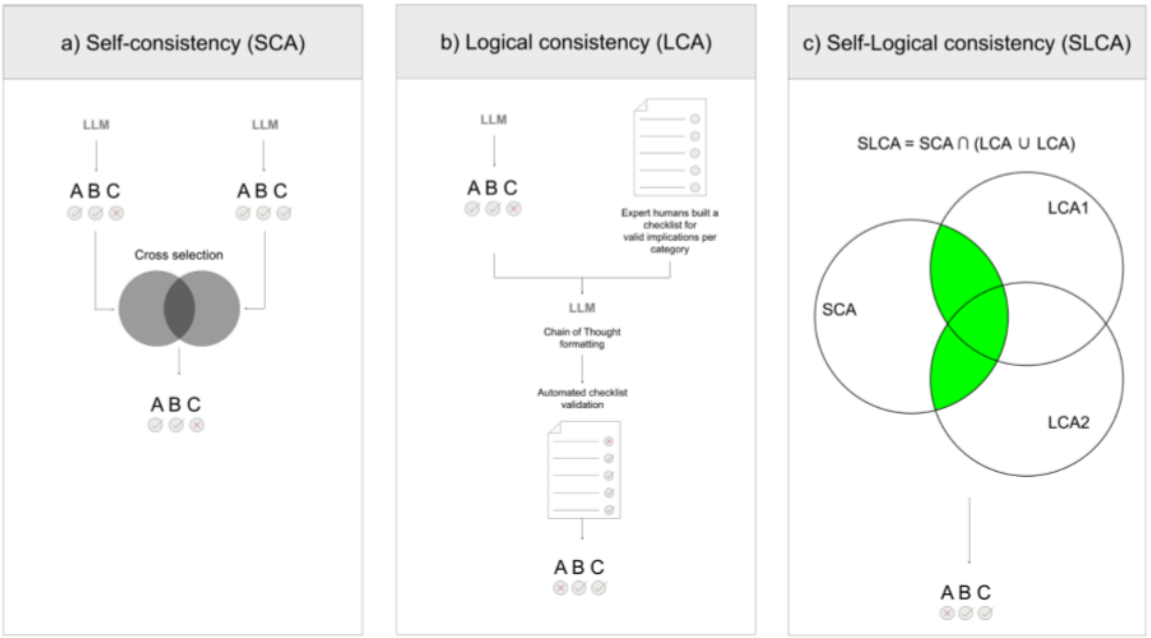


Figure 2 : Consistency Assessment Layouts

Benchmarking Machine Learning Models

The benchmark compared 11 models over 49,140 classifications among 1,170 feedbacks (Figure 1) :

- **Traditional Models:** Naive Bayes (NB) and a 1.5 million parameter Long Short-Term Memory (LSTM), trained and evaluated using ten-fold cross-validation.
- **Regex-Based Model:** Used in its production version at our facility.
- **Llama-3.1 70B and Variations:** Standalone and consistency-assessed variations (Llama-3.1 70B+SCA, Llama-3.1 70B+LCA, Llama-3.1 70B+SLCA).

- **Llama-3.1 405B and Variations:** Standalone and consistency-assessed variations (Llama-3.1 405B+SCA, Llama-3.1 405B+LCA, Llama-3.1 405B+SLCA).
- **GPT-4 Turbo and Variations:** Standalone and variations (GPT-4+SCA, GPT-4+LCA, GPT-4+SLCA).

NB was performed after dimensional reduction using non-negative matrix factorization. Due to time constraints, the gold standard for this experiment was created by a single quality-of-care expert.

Statistical analysis

All classification outputs were treated as categorical data (category present vs. absent). For the human-curated sample (100 free-text feedbacks), each feedback was independently analyzed three times by each agent (human, LLM variations), resulting in 12,600 binary classifications (21 categories plus 2 tones per feedback). Precision, recall, and hallucination rates were compared using McNemar tests (two-tailed α=.05) on pairwise differences between models. Krippendorff's alpha was computed to assess reproducibility among agents of the same type. For the benchmark dataset (1,170 feedbacks, yielding 49,140 binary classifications), overall performance metrics (precision, recall) were likewise tested by pairwise comparisons using chi-square tests (P<.001). Confidence intervals (95% CIs) for precision and recall were derived using the Wilson method. All analyses were performed using the R (version 4.0) statistical environment.

Ethical Considerations

This study complies with French data protection laws. Patients were informed about data usage and could withdraw access at any time. Ethical approval was granted by the Ethical and Scientific Committee of the Montpellier University Hospital Centre (Registration Number: A015/2024-05-050/001).

Results

Comparing LLMs and human classification

We compared the performance of LLMs (GPT-4 and Llama-3.1 405B) to human experts in classifying 12,600 categories across 100 patient feedback samples, evaluating precision, recall, time efficiency, and hallucination rates.

Humans demonstrated high precision (0.87) but low recall (0.64), reflecting their cautious approach and tendency to underclassify. Classification was also time-intensive, requiring approximately 3 hours to process 100 feedback samples, underscoring the inefficiency of manual reviews for large-scale analyses.

LLMs were slightly less precise than humans, with GPT-4 achieving a precision of 0.71 and Llama-3.1 405B achieving 0.76. However, LLMs were significantly more exhaustive, with GPT-4 achieving a recall of 0.88 compared to 0.63 for Llama-3.1 405B (McNemar $P<.001$). These findings highlight the potential of LLMs for identifying more relevant categories while sacrificing some precision.

A major limitation of LLMs was their susceptibility to extrinsic faithfulness hallucinations. GPT-4 exhibited a hallucination rate of 15.67% (168/1063), while Llama-3.1 405B showed a lower rate of 7.17% (51/711). These hallucinations

represent fabricated or misaligned category identifications, undermining the reliability of raw LLM outputs without additional safeguards.

Assessing self and logical LLM consistency

We progressively evaluated the effects of Self-Consistency Assessment (SCA), Logical Consistency Assessment (LCA), and Self-Logical Consistency Assessment (SLCA) on GPT-4 and Llama-3.1 405B, highlighting their impact on precision, recall, hallucination reduction, and reproducibility.

SCA improved precision for both LLMs, increasing GPT-4's precision by 12% and Llama-3.1 405B's by 7%. However, hallucination rates remained significant, with GPT-4+SCA exhibiting a 5.87% (50/852) hallucination rate and Llama-3.1 405B+SCA showing 5.19% (32/617). GPT-4+SCA achieved a precision-recall score of 0.83-0.83, while Llama-3.1 405B+SCA showed slightly lower recall with a score of 0.83-0.59. These results indicate that while SCA enhances precision, it does not adequately address hallucination issues.

LCA further improved performance by addressing hallucinations. GPT-4+LCA reduced hallucination rates to 1.99% (18/903), achieving a precision-recall score of 0.76-0.80 (McNemar $P<.001$), though it remained less precise and more exhaustive than human evaluations. Llama-3.1 405B+LCA achieved a precision of 0.77 and a recall of 0.63, with hallucination rates reduced to 2.27% (16/704). Notably, all hallucinations for both models were confined to the "Medical and Paramedical Care" category, specifically related to "Quality and Speed of Response from Calls to Regulatory Services and Emergency Services (EMS, emergency department)."

SLCA combined SCA and LCA for maximal performance gains. GPT-4+SLCA achieved human-equivalent precision (0.86) and greater recall (0.75, McNemar

P<.001), with hallucination rates reduced to just 0.51% (4/786). Llama-3.1 405B+SLCA exhibited slightly lower performance with a precision-recall score of 0.84-0.58 and a hallucination rate of 1.67% (10/599). Importantly, hallucinations for both models occurred exclusively under the same identifiable circumstances noted with LCA.

SLCA improved inter-rater agreement, with Krippendorff's alphas of 0.86 for GPT-4+SLCA and 0.87 for Llama-3.1 405B+SLCA, compared to 0.67 for human experts. This indicates that SLCA not only enhances classification performance but also ensures greater consistency.

Agents	Mean performances over three independent agents			Classification reproducibility (Krippendorff's alpha)
	Precision (%)	Recall (%)	Hallucination rate (%)	
Human experts	87	64	0 (by design)	0.67
GPT-4 standalone	71	88	15.80 (168/1063)	0.81
GPT-4 + SCA	83	83	5.87 (50/852)	0.87
GPT-4 + LCA	76	80	1.99 (18/903)	0.82
GPT-4 + SLCA	86	78	0.51 (4/786)	0.86
Llama-3.1 405 B standalone	76	63	7.17 (51/711)	0.87
Llama-3.1 405 B + SCA	83	59	5.19 (32/617)	0.89
Llama-3.1 405 B + LCA	77	63	2.27 (16/704)	0.86
Llama-3.1 405 B + SLCA	83	59	1.67 (10/599)	0.87

Table 1 : Patient Feedback Classification Performances of Humans, LLM and Consistency Applied to LLM

Machine Learning Benchmark

The external validation benchmark revealed a clear progression in performance, from traditional methods to state-of-the-art LLMs enhanced with SLCA, in terms of precision and recall.

Historical methods demonstrated limited capabilities. Regex achieved a precision-recall score of 0.39-0.73, performing reasonably well in recall but struggling with precision. LSTM, while slightly more precise at 0.53, had a recall of only 0.39, reflecting its inability to capture the full complexity of patient feedback. Naive Bayes performed worst, with a precision of 0.06 despite a recall of 0.79, indicating a significant overclassification issue.

Standalone Large Language Models (LLMs) showed considerable improvement over historical models. GPT-4 exhibited a recall of 0.75 but suffered from poor precision at 0.61, highlighting a tendency to overclassify. Similarly, Llama-3.1 405B performed comparably with precision-recall scores of 0.62-0.73. Llama-3.1 70B, the smallest of the LLMs tested, underperformed, achieving a low precision of 0.45 despite an acceptable recall of 0.70.

SLCA markedly enhanced LLM performance. GPT-4+SLCA achieved the highest precision (0.88) and recall (0.71). Llama-3.1 405B+SLCA followed with balanced precision-recall scores of 0.68-0.68. While Llama-3.1 70B+SLCA exhibited high precision (0.82), its recall remained limited at 0.30, highlighting challenges in comprehensiveness for smaller models.

The progression from traditional models to standalone LLMs and ultimately to SLCA-enhanced LLMs demonstrates a transformative leap in classification reliability.

All differences between models were statistically significant (Chi-square tests 2 degrees of freedom for each consecutive pair of metrics, $P<.001$).

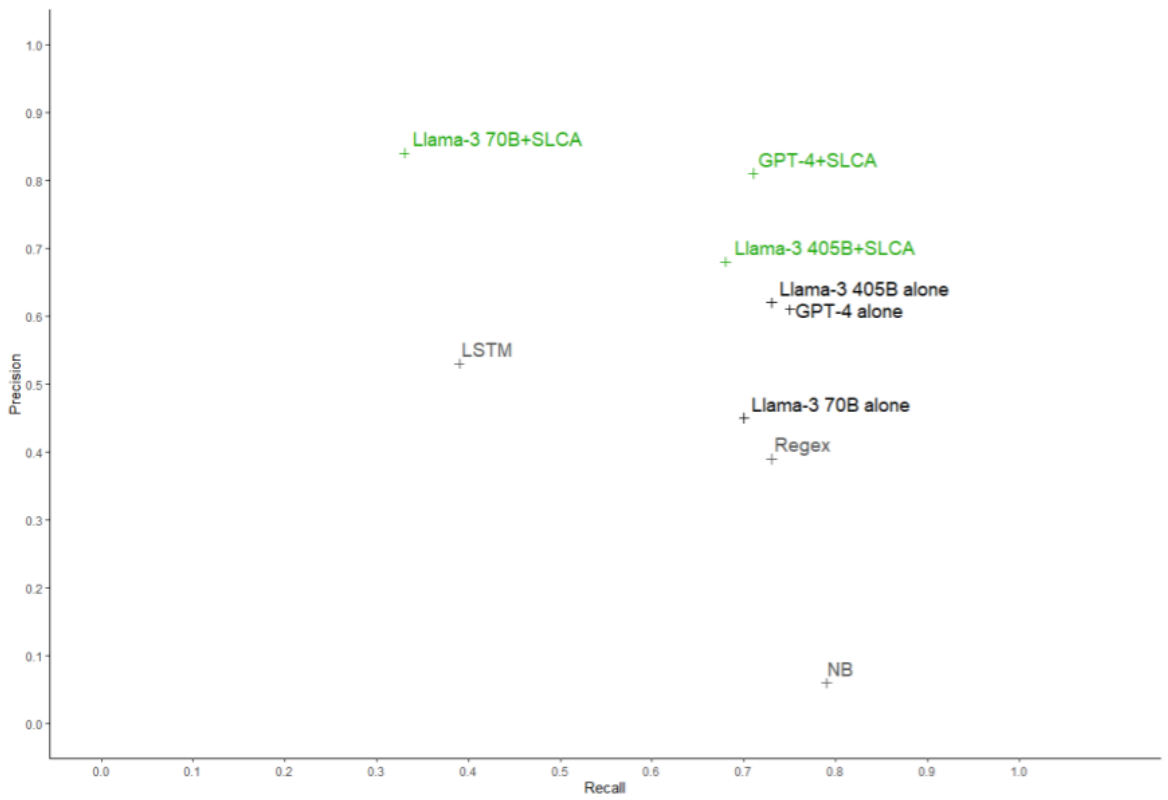


Figure 3 : Benchmark of Machine Learning Models on Patient Feedback Classification

Performances Varies Across Categories

The benchmark subgroup analysis reveals notable variations in model performance across patient feedback categories.

Certain categories consistently posed challenges for all evaluated models. For example, “Humanity and Availability of Professionals — favourable” and “Medical and Paramedical Care — favourable” exhibited relatively lower F1 scores, ranging from 0.76 to 0.94. These categories likely involve nuanced or subjective interpretations, making them difficult for automated systems to classify accurately.

In contrast, categories such as “Room Temperature — favourable” and “Patient Rights — favourable” achieved near-perfect performance, with F1 scores between 0.99 and 1.00. These categories are characterized by more straightforward, objective

feedback, making them easier for models to process effectively. The stark difference between these and more challenging categories highlights the variability in model performance.

The analysis identified four outlier categories (out of 42) that were either particularly challenging or notably easier to classify. Importantly, this trend was consistent across all models, including both standalone and SLCA-enhanced LLMs. Notably, faithfulness hallucinations identified in the earlier LCA evaluation were confined to the challenging category “Medical and Paramedical Care.” These findings illustrate the evolving performance landscape across patient feedback categories, from highly reliable classifications in straightforward categories to persistent challenges in nuanced ones.

Discussion

Principal results

This study introduces the **SLCA** method, which improves the reliability of LLMs in classifying patient feedback without fine-tuning or annotated datasets. Combining self- and logical consistency assessments significantly reduces hallucinations in state of the arts LLMs, achieving human-level performance and surpassing other machine learning models.

The SLCA framework is built on a series of nested models : the LLM alone, SCA, and LCA. Each step increases precision or explainability at the expense of recall, offering flexibility to adjust accuracy and coverage based on clinical needs. Our method sets a new state of the art for patient feedback classification. While the best previously reported model, a BERT variant, achieved 70% accuracy [5], GPT-4 and Llama-3.1 405B with SLCA reached respectively 98% and 96%, thereby illustrating a substantial improvement over existing approaches.

Additionally, SLCA effectively addresses LLM limitations by nearly eliminating LLMs extrinsic faithfulness hallucinations. In a sample of 100 patient feedback entries, these hallucinations were almost entirely absent. Only 33 for GPT-4 and 35 for Llama-3.1 405B out of 12,600 classifications involved the implication “Quality and Speed of Response from Calls to Regulatory Services and Emergency Services (EMS, emergency department)”, which accounted for the 4 (GPT-4) and 10 (Llama-3.1 405B) detected extrinsic faithfulness hallucinations. Systematic human review of these classifications is therefore readily implementable. Although performance on larger datasets remains uncertain, the results are clinically promising. Moreover, our study shows the possibility to perform clinically viable

performances with the local model Llama-3.1 405B, according to the context of institutional data sovereignty constraints, which may limit the use of proprietary models like GPT-4.

Limitations

Llama-3.1 70B, suggesting that the effectiveness of SLCA may depend on the efficiency and capacity of the base model. Additionally, our study focused exclusively on SLCA, but incorporating other consistency evaluation methods might further enhance LLM performance.

A significant limitation was the inability to access or review the internal states of the models, which could have provided insights into their reasoning processes. Recent studies have demonstrated that internal consistency assessments can effectively reduce hallucination rates, indicating a missed opportunity for further refinement [16],[17].

The clarity and precision of classification categories also probably played a critical role in model performance. Semantically complex categories, such as “medical and paramedical care,” posed particular challenges, leading to inconsistent results. These issues highlight the importance of further methodological refinements to address such complexities effectively.

There are several ways we could further upgrade this approach. Expanding the dataset in future studies would help establish the external validity of SLCA, particularly for mitigating factual and other faithfulness variants hallucinations, which were not assessed due to the lack of a definitive gold standard. Cross-validation of hallucination detection by multiple investigators could reduce the subjectivity of this process, improving robustness. Additionally, exploring the impact of language on

model performance—especially since our study material was in French — could impart valuable insights since LLMs often perform better in English. Finally, the environmental cost of using LLMs with SLCA should be weighed up against the advantages of not requiring fine-tuning for classification tasks.

Finally, our study highlights a key issue in text classification when the gold standard is relative and not fixed. We found limited inter-expert reproducibility, making consensus difficult. This mirrors real-world scenarios where evaluating automated language processing is challenging due to the subjective nature of the gold standard. Our approach produces human-level classifications with precision and recall, opening up a path to more stable, reproducible consensus and AI-generated gold standards where human-made ones are impractical.

Conclusions

Overall, SLCA presents practical benefits in healthcare, where patient feedback is crucial for quality improvement [3,12]. Its ability to achieve high accuracy with greater control over extrinsic faithfulness hallucinations and control over computational cost while saving the possibility of complete sovereignty of the patients data makes it a valuable tool for large-scale feedback processing. Assessing consistency, it seems, might be all you need to classify complex text data with reliability and precision.

Abbreviations

AI: Artificial Intelligence

BERT: Bidirectional Encoder Representations from Transformers

CHU: University Hospital Center (from the French “Centre Hospitalier Universitaire”)

CoT: Chain of Thought

EMS: Emergency Medical Services

ERIOS: Research and Integration Unit for Digital Tools in Healthcare (from the French “Espace de Recherche et d’Intégration des Outils numériques en Santé”)

F1: F1 Score, the harmonic mean of precision and recall

GPT-4: Generative Pre-trained Transformer 4

HAS: French National Health Authority (from the French “Haute Autorité de Santé”)

INSERM: French National Institute of Health and Medical Research (from the French “Institut National de la Santé et de la Recherche Médicale”)

ISDM: Montpellier Data Science Institute (from the French “Institut de Science des Données de Montpellier”)

LCA: Logical Consistency Assessment

LIRMM: Montpellier Laboratory of Informatics, Robotics, and Microelectronics (from the French “Laboratoire d’Informatique, de Robotique et de Microélectronique de Montpellier”)

LLM (or **LLMs**): Large Language Model(s)

LSTM: Long Short-Term Memory

NB: Naive Bayes

NLP: Natural Language Processing

SCA: Self-Consistency Assessment

SLCA: Self-Logical Consistency Assessment

95% CI: 95% Confidence Interval

Code and Data Availability

The code is available on GitHub as presented in references [18].

Datasets are available from the corresponding authors on reasonable request. Pseudonymized patient feedback may contain personal health information; access is granted with traceability according to French National Commission on Informatics and Liberty and Montpellier University Hospital Centre policies.

Acknowledgments

The study was conceived, funded, and executed entirely by CHU de Montpellier. We acknowledge all participants and professionals involved in the patient feedback processing. Special thanks to the ERIOS team: Cécile Yriarte, Louise Robert, Laurine Moniez, Quentin Luzurier, Marin Portalez, Letizia Pala, Loïc Fontaine, Mylene Fernandes, Rita Pires, Yrina Gilhodes, Anne Laurent from the ISDM, Hugo Loi and Leo Giorgis from Pixminds, the LIRMM, Reference Centre for Congenital Anomalies, Clinical Genetic Unit and the Data Sciences ward of the CHU de Montpellier for their impactful discussions and contributions.

References

1. Doyle C, Lennox L, Bell D. A systematic review of evidence on the links between patient experience and clinical safety and effectiveness. BMJ Open 2013 Jan 3;3(1). PMID:23293244

2. Marie Gloanec, Karen Assmann, Caroline Prunet , Sandrine Morin, Laetitia May-Michelangeli, Pavel Soriano, Thimotée Chehab, Pierre-Alain Jachiet. Expérience des patients : Développement d'un outil d'analyse des verbatim de patients issus d'e-Satis. Haute Autorité de Santé; 2022 Oct. Available from: https://www.has-sante.fr/upload/docs/application/pdf/2022-11/iqss_outil_verbatim_note_cadrage_2022.pdf

3. Karen Assmann, Marie Gloanec, Laetitia May-Michelangeli, Sandrine Morin, Caroline Prunet, Daniel Benamouzig (CNRS-SciencesPo), Henri Bergeron (CNRS - SciencesPo), Arnaud Fouchard (EIR Conseil Santé). Expérience des patients hospitalisés en France : Analyse nationale des commentaires libres du dispositif e-Satis. Haute Autorité de Santé; 2022 Jun.

4. Cammel SA, De Vos MS, van Soest D, Hettne KM, Boer F, Steyerberg EW, Boosman H. How to automatically turn patient experience free-text responses into actionable insights: a natural language programming (NLP) approach. BMC Med Inform Decis Mak 2020 May 27;20(1):97. PMID:32460734

5. Khanbhai M, Anyadi P, Symons J, Flott K, Darzi A, Mayer E. Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review. BMJ Health Care Inform 2021 Mar;28(1). PMID:33653690

6. Martino, Steven C., Ish, Daniel, Shaller, Dale, Elliott, Marc N., Parker, Andrew M., Schlesinger, Mark, Grob, Rachel, Osoba, Osonde A., Hays, Ron D. Using Natural Language Processing to Code Patient Experience Narratives: Capabilities and Challenges. RAND Corporation; Available from: https://policycommons.net/artifacts/4836513/using-natural-language-processing-to-code-patient-experience-narratives/5673236/ [accessed Jul 3, 2024]

7. LMSys Chatbot Arena Leaderboard - a Hugging Face Space by lmsys. Available from: https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard [accessed Jul 24, 2024]

8. Boussina A, Krishnamoorthy R, Quintero K, Joshi S, Wardi G, Pour H, Hilbert N, Malhotra A, Hogarth M, Sitapati AM, VanDenBerg C, Singh K, Longhurst CA, Nemati S. Large language models for more efficient reporting of hospital quality measures. NEJM AI Massachusetts Medical Society; 2024 Oct 24;1(11). doi: 10.1056/aics2400420

9. Rydzewski NR, Dinakaran D, Zhao SG, Ruppin E, Turkbey B, Citrin DE, Patel KR. Comparative Evaluation of LLMs in Clinical Oncology. NEJM AI 2024 May;1(5). PMID:39131700

10. Wang X, Wei J, Schuurmans D, Le Q, Chi E, Narang S, Chowdhery A, Zhou D. Self-Consistency Improves Chain of Thought Reasoning in Language Models. arXiv [csCL]. 2022. Available from: http://arxiv.org/abs/2203.11171

11. Maynez J, Narayan S, Bohnet B, McDonald R. On Faithfulness and Factuality in Abstractive Summarization. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics 2020. p. 1906–1919. doi: 10.18653/v1/2020.acl-main.173

12. Wei J, Wang X, Schuurmans D, Bosma M, Chi E, Xia F, Le Q, Zhou D. Chain of thought prompting elicits reasoning in large language models. Adv Neural Inf Process Syst 2022 Jan 28;abs/2201.11903. Available from: https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html

13. Copi IM, Cohen C, McMahon K. Introduction to Logic. Routledge; 2016. Available from: https://books.google.com/books/about/Introduction_to_Logic.html?hl=&id=1iUDDQAAQBAJ ISBN:9781315510873

14. Jia Li, Ge Li, Yongmin Li, Zhi Jin. Structured Chain-of-Thought Prompting for Code Generation. 2023. Available from: https://arxiv.org/pdf/2305.06599 [accessed Sep 26, 2024]

15. Sultan MA, Ganhotra J, Astudillo RF. Structured Chain-of-Thought Prompting for Few-Shot Generation of Content-Grounded QA Conversations. 2024 Feb 19; Available from: https://paperswithcode.com/paper/structured-chain-of-thought-prompting-for-few [accessed Jul 24, 2024]

16. Chen C, Liu K, Chen Z, Gu Y, Wu Y, Tao M, Fu Z, Ye J. INSIDE: LLMs' Internal States Retain the Power of Hallucination Detection. 2024. Available from: http://arxiv.org/abs/2402.03744 [accessed Oct 17, 2024]

17. Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie, Pascale Fung. LLM Internal States Reveal Hallucination Risk Faced With a Query. arxiv.org. 2024. Available from: https://arxiv.org/pdf/2407.03282 [accessed Sep 23, 2024]

18. Website. Available from: https://github.com/ERIOS-project/SLCA_LLM4Quality.

Author Information

Contributions

Z.L. contributed to study design, data analysis, data interpretation, figures design and writing of the manuscript. D.M. contributed to study coordination, study design and figures design. FX.D, S.G and E.P.L contributed to data collection and data management. X.C. contributed to data management and large language models predictions. P.T. contributed to the writing of the manuscript. G.M. contributed to the ethical and legal framework of the study and writing of the manuscript. K.Y. contributed to study coordination, study design, data interpretation, figures design and writing of the manuscript. All Authors contributed to the critical review of the manuscript.

Competing Interests

All authors declare they have no financial or non-financial competing interests.