Self-Logical Consistent GPT-4 Enables Human-Level Classification of Patient Feedback

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

* Corresponding Authors Email: z-loi@chu-montpellier.fr and kevin.yauy@chu-montpellier.fr

Abstract

Patient satisfaction feedback is crucial for hospital service quality, but human-led reviews are time-consuming and traditional natural language processing remains ineffective. Large Language Models (LLM) offer potential, but their tendency to generate illogical thoughts limits their use in healthcare. Here we describe Self-Logical Consistency Assessment (SLCA), a method ensuring a reproducible LLM classification explained by a logically-structured chain of thought. In an analysis targeting extrinsic faithfulness hallucinations, SLCA mitigated the 16% GPT-4 hallucination rate, leaving only three residual cases across 12,600 classifications from 100 diverse patient feedbacks. In a benchmark designed to evaluate classification accuracy, SLCA applied to GPT-4 outperformed best algorithms, with a 88% precision rate and a 71% recall rate across 49,140 classifications from 1,170 sampled patient feedbacks. This method provides a reliable, scalable solution for improving hospital services and shows potential for accurate, explainable text classifications without fine-tuning.

Introduction

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 methods for analysing patient feedback have historically fallen short due to the technical limitations of Natural Language Processing (NLP) algorithms. Previous works performed an unsupervised classification of 2.5 million patient feedbacks, offering a comprehensive overview of patient concerns and defining 20 categories for classification [2,3]. Despite the value of these insights, models such as Naive Bayes and BERT have struggled to accurately classify nuanced feedback due to their inability to handle complex language contexts effectively [4,5].

Large Language Models (LLMs) such as the proprietary model GPT-4 and open source model Llama-3.1 offer a promising alternative[6], with their superior ability to understand natural language and pinpoint subtle nuances in patient feedback [5,7]. Recent work illustrated that by evaluating the self-consistency of LLMs predictions, the availability of a model to provide the same classification over multiple independent attempts, greatly optimises their performances in classification tasks [8]. However, these models tend to produce hallucinations, that consist of factual mistakes or logical flaws in the generated texts [9]. Even though factual hallucinations may be acceptable to certain extent within a hospital context, the loss of explainability due to extrinsic faithfulness hallucinations — the generation of illogical thought processes — is incompatible in sensitive applications such as patient

feedback analysis. Given that the Chain of Thoughts (CoT) generation methods ameliorate the explainability of these models[10], validation of CoT logical structures might be a potential solution to the issue of hallucinations.

Here we describe and review Self-Logical Consistency Assessment (SLCA), a method designed to bolster the reliability of LLM-generated predictions. The SLCA combines Self-Consistency Assessment (SCA) [8] and 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.

Results

GPT-4 Classification is More Exhaustive than Humans, but is Unsuitable Due to Extrinsic Faithfulness Hallucinations

To our knowledge, evaluation studies have not been performed to ascertain human-led classification of patient satisfaction feedback to date, nor GPT-4 faithfulness hallucinations tendencies. Three human quality-of-care experts and 3 independent GPT-4 agents (prompt provided in Supplementary Note 2) were directed to classify 100 patient feedbacks (Table 1) among 21 categories and two 2 tones (favourable/unfavourable) (Supplementary Note 1) for a total effective of 12,600 classifications. All responses were blindly affirmed or infirmed a posteriori as gold standard by the Investigator. In addition, the Investigator determined all wrongly identified categories (n =462) made by GPT-4 agents and assessed the presence or absence of any extrinsic faithfulness hallucination. We found that humans were

precise but not exhaustive with a precision-recall of 0.87-0.64. Moreover, the classification proved to be time-consuming : 3h per 100 feedbacks. In constrat, GPT-4 was less precise but more exhaustive than humans with a precision-recall of 0.71-0.88 (McNemar ϱ<1e-15). However, GPT-4 exhibited a significant extrinsic faithfulness hallucination rate, representing 16% of all generated categories identifications.

Self-Consistency Enhances Precision, while Logical Consistency Reliably Mitigates Most Extrinsic Faithfulness Hallucinations

Self-consistency consists of generating two independent LLM predictions and proceeds into a cross selection of their results. Only categories recognised twice by GPT-4 are confirmed as identified by GPT-4+SCA (Figure 1). We evaluated the performances of 3 independent runs of GPT-4+SCA on this task (**i.e.** a pair of outputs from GPT-4 to produce 1 prediction). SCA increased GPT-4 precision by 12%. GPT-4+SCA was still less precise and more exhaustive than humans with a precision-recall of 0.83-0.83 (McNemar ϱ<1e-15). However, GPT-4+SCA still presented a 6% faithfulness hallucination rate.

To address this “deal breaker” issue pertaining to faithfulness hallucination, we developed a Logical Consistency Assessment (LCA) method to evaluate the LLM CoT structure without the need for fine-tuning or annotated datasets. A GPT-4 standalone prediction was directed into a second prompt to produce a CoT with a detailed structure encompassing premise (a citation from the feedback), implication (a logical link between feedback citation and categories), and conclusion (the identified category) as defined in philosophy of logic reasoning[11]. The LLM CoT in

accordance with the instructions reflects the logical consistency of the prediction and only categories identified with a valid implication were retained (Figure 1).

We appraised the performances of 3 independent runs of GPT-4+LCA on this task (i.e. 2 chained generations from GPT-4 to produce 1 prediction, three times). GPT-4+LCA was also less precise and more exhaustive than humans with a precision-recall of 0.76-0.80 (McNemar ϱ<1e-15). Notably, GPT-4+LCA successfully removed most hallucinations from its predictions as it represented only 1% of identified categories. Moreover, the 11 faithfulness hallucinations all concerned the category “Medical and Paramedical Care” and occurred with the invocation of the implication “Quality and Speed of Response from Calls to Regulatory Services and Emergency Services (EMS, emergency department)”, making them identifiable in daily hospital use.

Self-Logical Consistency Assessment Applied to GPT-4 Enables Human-Level Performances

We investigated the performance of SLCA combining both SCA and LCA applied with GPT-4 on the same task (Figure 1). GPT-4+SLCA was equivalently precise and more exhaustive than humans with a precision recall of 0.86-0.75 (McNemar ϱ<1e-7). It presented a total of only 3 faithfulness hallucinations over the 12,600 classifications, highlighting the robustness and reliability of this combined method. These hallucinations occurred under the exact same circumstances as for its nested method LCA. Moreover, GPT-4+SLCA tended to be more reproducible than human experts with a Krippendorff’s alpha between 0.85 versus 0.67 for 3 GPT-4+SLCA agents and 3 human agents respectively. Overall, this approach delivered better performances than humans.

To provide extrinsic validation of SLCA focusing on precise accuracy evaluation, we compared GPT-4+SLCA to other automated solutions in a large scale benchmark of 49,140 category classifications over 1,170 feedbacks : GPT-4 standalone, its consistency assessed variations : GPT-4+SCA, GPT-4+LCA; Llama-3.1 70B standalone (referred simply as “Llama-3”), its consistency assessed variations : Llama-3+SCA, Llama-3+LCA, Llama-3+SLCA; Regex (decision tree used in production in our establishment), Long Short Term Memory (LSTM) and Naive Bayes (NB).

GPT-4 combined with SLCA achieved the most optimal performance among all models, with precision of 0.88, recall of 0.71, and a global accuracy of 0.98. Other GPT-4 variants performed slightly less effectively: GPT-4+LCA produced precision of 0.78 and recall of 0.73, while GPT-4+SCA provided precision of 0.75 and recall of 0.72. Standalone GPT-4 demonstrated poor precision (0.61) despite a high recall (0.75).

In contrast, Llama-3 models generally underperformed compared to GPT-4 models. Llama-3+SLCA had a high precision rate of 0.82 but low recall of 0.30 whereas Llama-3+LCA presented a precision rate of 0.76 and recall of 0.31. Llama-3+SCA offered a more balanced performance with precision at 0.72 and recall at 0.62. Standalone Llama-3 performed poorly, with low precision (0.45) in spite of an acceptable recall level (0.70).

Historical models like Regex and LSTM exhibited lower performance, with precision-recall pairs of 0.39–0.73 and 0.53–0.39, respectively. Naive Bayes was the

worst option, with a precision rate of 0.06 and recall rate of 0.79. All differences between models were statistically significant ($p < 1e-3$).

Performances Varies Across Categories

The benchmark sub-group analyses offer valuable insights into the varying performance of different models across patient feedback categories. For the categories “Humanity and Availability of Professionals — favourable” and “Medical and Paramedical Care — favourable,” all LLMs displayed relatively lower performance, with F1 scores ranging from 0.66 to 0.95. In contrast, categories such as "Room Temperature — favourable" and "Patient Rights — favourable" exhibited higher model performance, with F1 scores between 0.97 and 1.00. This observation underscores the presence of four outlier categories (out of 42) that are either particularly challenging or notably easier to classify — a trend consistent across all evaluated models. All subgroup analysis results are detailed in Supplementary Table 1.

It is important to note that the low number of LCA faithfulness hallucinations identified in the first experience only occurred in a difficult-to-identify category as described in the benchmark — “Medical and Paramedical Care”.

Discussions

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 GPT-4, 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 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 with SLCA reached 98%, thereby illustrating a substantial improvement over existing approaches.

Additionally, SLCA effectively addresses LLM limitations by nearly eliminating GPT-4's extrinsic faithfulness hallucinations. In a sample of 100 patient feedback entries, these hallucinations were almost entirely absent. Only 33 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 3 detected extrinsic faithfulness hallucinations. Systematic human review of these classifications is therefore feasible. Although performance on larger datasets remains uncertain, the results are clinically promising.

Although our approach improved GPT-4 performance, similar enhancements were not observed with Llama-3.1 70B, indicating that the effectiveness of SLCA may vary across models. Additionally, while we focused on SLCA, other forms of consistency evaluation could have been incorporated to further boost LLM performance. We were

unable to review the internal states of the models in our context, which is a limitation in itself, especially considering that recent studies have illustrated that internal consistency assessments can reduce hallucination rates [12,13]. Moreover, the clarity and precision of classification categories also impacted the model's performance. In particular, semantically complex categories posed challenges such as “medical and paramedical care”, leading to inconsistent results and some logical implications were ambiguous and poorly defined, contributing to hallucinations and emphasising the need for precise wording and lucidity when using SLCA methodologies.

To address these limitations, we reviewed 902 false positive category identifications for extrinsic faithfulness hallucinations across 100 feedback samples. This manual review was supplemented by formal guidelines to help define the scope of such hallucinations. However, the detection process remained subject to the Investigator's judgement, which introduces subjectivity. In terms of performance variability, we closely examined the clarity of classification categories, ensuring meticulous definition of semantic boundaries and logical implications where possible.

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. Moreover, future research should consider the selection of models in the context of institutional constraints, such as data sovereignty, which may limit the use of proprietary models like GPT-4. Open-source alternatives, such as future iterations of Llama, may also

offer a solution. 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.

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 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.

Methods

Data and Inclusion Criteria

The French national system E-Satis routinely collects patient feedback following hospital stays and provides this data to healthcare institutions. This study focused on feedback collected through E-Satis from adult patients hospitalised at the Montpellier University Hospital Centre between 2022 and 2024. Exclusions were made for feedback from patients who declined data use, feedbacks too lengthy for model analysis, compensation claims, and feedback containing extreme content (see Supplementary Figure 1). All selected feedback was pseudonymised to protect personal data. Two feedback samples were chosen for analysis: a human-curated sample portraying feedback diversity and a large randomly selected sample representative of the general population. These two samples are detailed in Supplementary Tables 2 and 3.

Two gold standards (one per sample) were built for this study:

- Gold standard 1: Human vs. GPT-4 Consistency

This comparison was conducted over 12,600 classifications from 100 feedbacks. The feedback sample was selected by three human experts to include a wide range of information-rich feedback that represented both compliments and criticisms. A list of categories identified by both human experts and GPT-4 was blindly evaluated by a fourth quality-of-care expert to either validate or invalidate each category. Details of this gold standard are outlined in Supplementary Table 4.

- Gold Standard 2: Benchmark Analysis

For the benchmark analysis, 49,140 classifications were conducted on 1,170 feedbacks. Due to the high time demands, the gold standard was created by a single

quality-of-care expert. This sample was randomly selected from the 2023 E-Satis database specific to our facility. The sample size was determined to detect a 2% difference in precision or recall across 20 bilateral tests (comparing 11 models), with a total alpha error rate of 5% (using Bonferroni correction) and 80% statistical power. The estimated precision and recall levels ranged from 0.60 to 0.95, with a maximum attrition rate of 10%. Based on these parameters, 40,407 classifications were required, equating to 962 feedbacks with 42 category identifications per feedback. Supplementary Table 5 outlines the contents of this gold standard.

Categorisation

The classification task in this study always corresponds to the following method : categorising the feedback among non-exclusive 21 categories and 2 non-exclusive tones (favourable and unfavourable) (Supplementary Note 1). The categories are adapted from the categorization proposed by the works of the Haute Autorité de Santé (French National Healthy Authority) [2,3], with the addition of the category “Patient’s Rights” to fulfill the operational purpose of this classification : defining local healthcare quality improvement axes. The favourable tone describes a category in a positive way, such as a compliment or an aspect of the hospitalisation the patient appreciated. In contrast, the unfavourable tone negatively characterises a category such as criticism over something the patient did not like.

Metrics of Interest

The comparisons between humans and GPT-4 consistency assessment are centered on four metrics : precision, which is a prerequisite for medical-grade classification, recall, reproducibility and extrinsic faithfulness hallucination rate. Precision is characterised as how many selected documents are relevant while recall is described

as how many relevant items are selected. Extrinsic faithfulness hallucinations are stated as described in previous work[9] : it corresponds to the LLM creating information that is not inferable from the original text or the presence of unjustified logical steps in the LLM reasoning. The Investigator assessed hallucinations by cross-referencing the LLM generation with the patient feedback, verifying that every piece of information in the CoT leading to a category identification was either explicitly stated or inferable from the original text. If the CoT included additional information or logical inconsistency, it was flagged as containing an extrinsic faithfulness hallucination. Blinding evaluation was only partially possible due to the distinctive structure of the CoT produced by LCA and SLCA compared to LLMs standalone and SCA.

In addition, the comparisons between humans and GPT-4 consistency assessment were conducted over a rather low number of feedback and with three agents of each type, to avoid high individual human time-consumption and in order to evaluate reproducibility of the results via Krippendorff's alphas (see Discussion).

The benchmark focused on precision and recall estimations for global performances. Benchmark sub groups analysis focused on F1 score for readability purposes.

Philosophy of Logic Approach for Logical Consistency

The evaluation of logical human-led argumentation is a fundamental and well explored field in philosophy of logic[11]. Despite the existence of previous studies attempting to structure LLMs Chains of Thought (CoT)[14,15], no research to our knowledge has directly applied philosophical methodologies to evaluate the validity of the logical components constituting the CoT generated by LLMs. These methods oriented the core of our work, associating multidisciplinary knowledge.

An argument is classically described as having three components : premise, logical implication and conclusion. The premise is an unverified axiomatic statement. In our case, it is a quote from the patient's feedback. The conclusion is a statement that is deduced from the premise, here, it is a category identification. The implication is supposed to logically link the premise and the conclusion by a causality relationship. If the trueness of the premise always implies the conclusion by the mechanism described in the implication, the argument is described as “valid”. Evaluating the validity of a LLM CoT structured as an argument would therefore allow us to ascertain the logical consistency of the prediction. Given that it seems to be very difficult to deterministically identify the quality of a premise and as the scope of valid conclusions is trivial (corresponding to the 21 defined categories), we concentrated on defining the scope of valid implications. The three human experts and the Principal Investigator constructed an exhaustive list of implications considered as valid for each category of identification conclusion. This implication list is provided to the LLM with the indications to produce a valid argument. To identify a category, the logical consistency is evaluated on the capacity of the LLM to generate an implication from the adequate implication sub-list.

Study Design on Patient Feedback Classification

Human Quality-of-Care Experts Evaluation

Three human quality of care experts were asked independently to classify the main 100 feedback samples. No communication was established between participants during the exercise. Human experts were considered as not subject to hallucinations.

GPT-4 Standalone

Three runs of GPT-4 turbo (version March 05, 2024) classified each feedback. All classifications were independent. The prompt contained detailed information about the output structure, the conditions of classification, a list of available categories, and a structure-free CoT was instructed to generate as general prompt engineering good practices. No valid classification example was given. The prompt provided is available in Supplementary Note 2.

Self-Consistency Assessment Applied to GPT-4

Self-consistency describes the LLM's ability to produce similar results in independent predictions under the same conditions. Three additional GPT-4 runs classified each feedback. Every classification was independent. The classifications were assigned in 3 groups of 2 in order to perform self-consistency assessment. For each one of the three GPT-4+SCA, only categories identified twice by the GPT-4 standalone were retained as identified by GPT4+SCA. This cross selection aims to determine only the category consistently identified by GPT-4 (Figure 2).

Logical Consistency Assessment Applied to GPT-4

To assess GPT-4 logical consistency, we produced a practical implementation of the philosophical method presented above. As outlined in recent works, output format restrictions can impact logical reasoning [16]. To avoid this effect, an initial prediction was produced without restriction on the CoT structure. A second, chained, prompt instructs the LLM to structure its previous generation. On top of the three GPT-4 runs, a second, more structured prompt is issued (Supplementary Note 2). This prompt directs the LLM to refine its CoT into a structured logical argument consisting of a premise, an implication, and a conclusion. More specifically, the premise should directly cite the patient feedback, and the implication should come from the

predefined valid implication list that delineates the scope for each category (Supplementary Note 1). The conclusion must then logically deduce the appropriate category based on the premise and implication. Since typos errors or white spaces trims can skew feedback citations, this part of the CoT is not evaluated. The logical consistency is assessed only if the implication is included in the provided implication list and if the identified category corresponds to this very implication. This structure evaluation approach aims to rate the LLM's reasoning as a valid and coherent argument. For each GPT-4+LCA agent, only classifications that put forward a valid CoT were retained (Figure 2).

Self-Logical Consistency Assessment Applied to GPT-4

The three groups of two GPT-4 runs successively apply the self and the logical method, selecting only the identification that fulfills the two conditions :

- Both GPT-4 standalone runs must have identified the category
- At least one must have produced a valid CoT when directed to do so

This two-step process enables the evaluation of the LLM's logical consistency — the capacity of the model to generate a coherently structured and valid argument — with increased sensitivity. By requiring the generation of two independent logically structured CoT, the method not only tests the feasibility of producing such reasoning but also assesses self-consistency, which measures the reproducibility of LLM classifications across both attempts.

Large Scale Benchmark of Machine Learning Models

The benchmark compared 11 models : Naive Bayes (NB), a 1.5 million parameters Long Short Term Memory (LSTM), Regex; Llama-3.1 70B unquantised standalone and its consistency assessed variations : Llama-3+SCA, Llama-3+LCA,

Llama-3+SLCA; GPT-4 turbo version 5 March 2024 standalone and its variations : GPT-4+SCA, GPT-4+LCA and GPT-4+SLCA over the classification of 49,140 categories among 1,170 feedbacks. NB and LSTM have been trained and evaluated in ten fold cross validation. NB was performed after a dimensional reduction by a non-negative matrix factorisation. Regex was used in its production version available at our facility.

Training Sets

The training sets for the models were not strictly equivalent. LLMs and Regex benefitted from the expertise of quality professionals during their prompt engineering and decision tree generation, while LSTM and Naive Bayes were trained on a database of only 1,170 feedbacks. Despite the already considerable size of this sample, expanding the training data could enhance the performance of these models. Additionally, improving LSTM by integrating an efficient input embedding, such as a BERT encoder, could potentially optimise its performance further.

Code Availability

The underlying code for this study is available in Github and can be accessed via the following link : https://github.com/ERIOS-project/SLCA_LLM4Quality

Data Availability

The datasets used during the current study are available from the Corresponding Authors on reasonable request. Pseudonymised patient feedback may contain personal health information and their access can only be granted with traceability

according to the French National Commission on Informatics and Liberty and Montpellier University Hospital Centre policies.

Ethical Considerations

This study complies with French regulations relating to data protection laws. Patients were informed about the usage of their data and had the option to withdraw access at any time. The ethical approval of this work was been given by the Ethical and Scientific Committee of the Montpellier University Hospital Centre (registration number : A015/2024-05-050/001)

Figures and Tables

Figure : Graphical Abstract

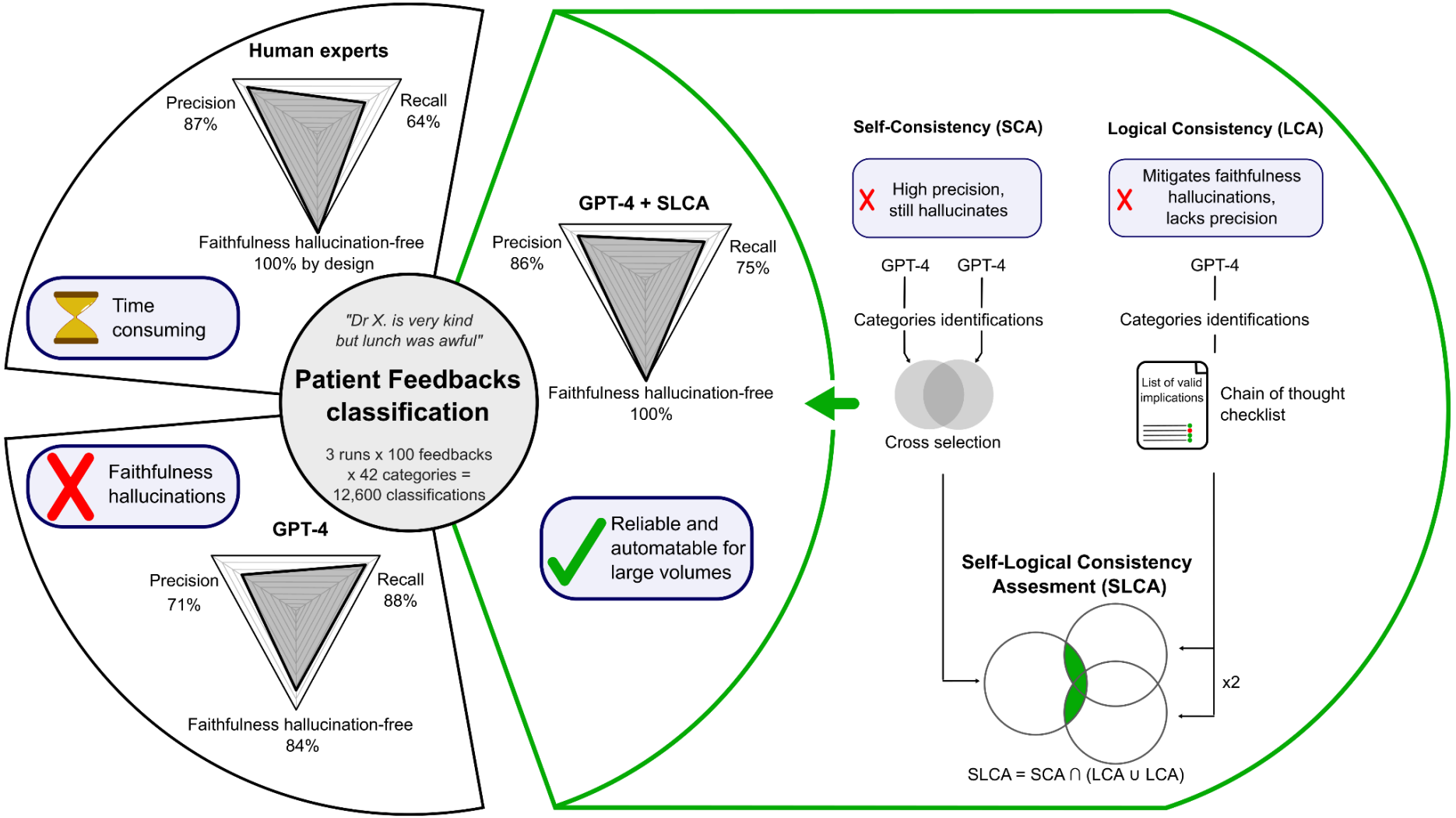
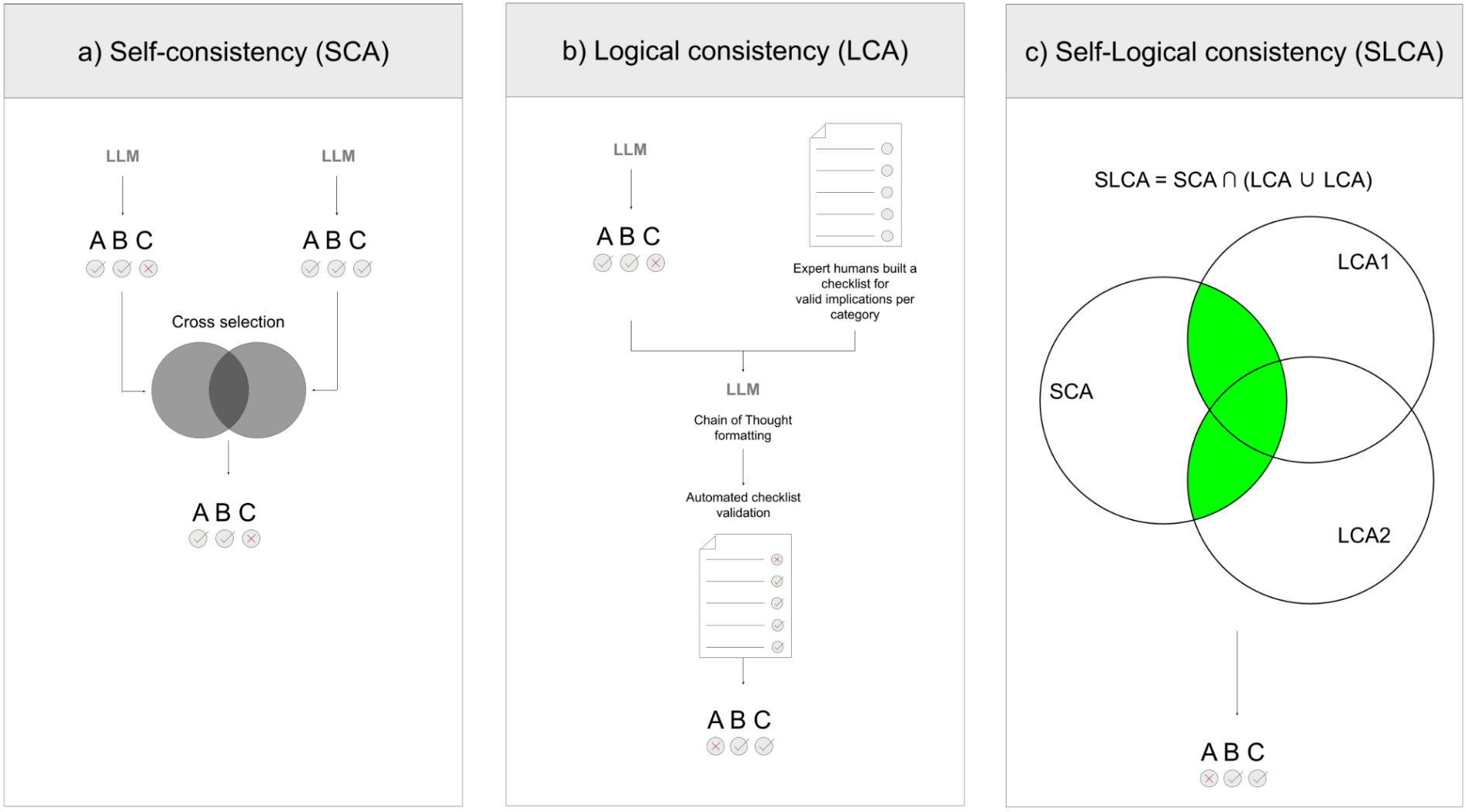


Table 1 : Patient Feedback Classification Performances of Humans, GPT-4 and Consistency Applied to GPT-4

Agents	Mean Performances over Three Independent Agents			Classification Reproducibility (Krippendorff's Alpha)
	Precision (%)	Recall (%)	Hallucination rate (%)	
Human Experts	87	64	0	0.67
GPT-4 Standalone	71	88	16	0.81
GPT-4 + SCA	83	83	6	0.87
GPT-4 + LCA	76	80	1	0.82
GPT-4 + SLCA	86	75	0	0.85

These results are provided from analysing 100 patient feedback independently three times by each agent. As each feedback classification is composed of 21 categories and identification of 2 tones, the size of sample for 95% confidence intervals (95%CI) computation is 12,600. All precision and recall 95%CI present a range <1%. SCA stands for Self-Consistency Assessment, LCA for Logical Consistency Assessment and SLCA for Self-Logical Consistency Assessment. Classification reproducibility has been estimated by the accordance level between agents of the same type with a Krippendorff's alpha.

Figure 1 : Consistency Assessment Layouts

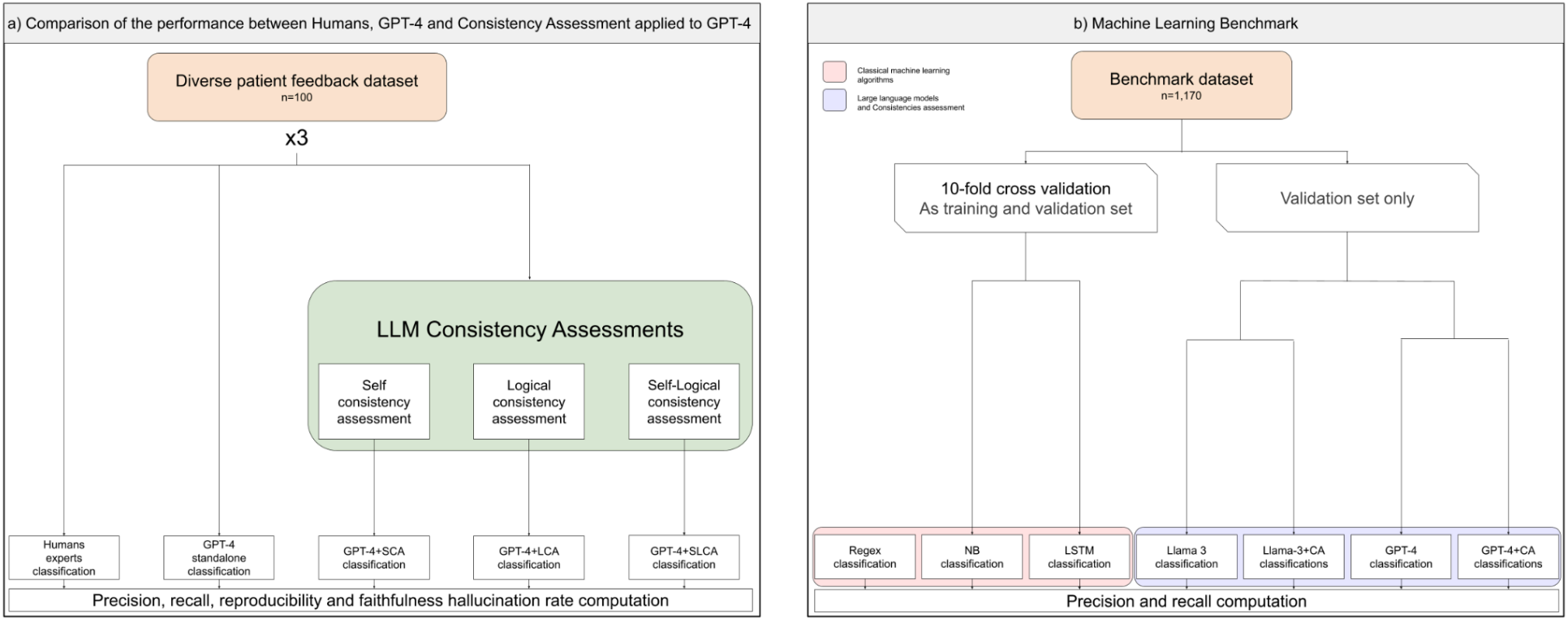


a) SCA proceeds to a straightforward cross selection. Only categories identified twice are kept.

b) LCA directs the LLM directing it to produce two structured Chain of Thought (CoT) encompassing a premise (a citation from the feedback), an implication selected from a predefined list and a conclusion (the identified category). A deterministic algorithm evaluates whether the implication given by the LLM can be found attached to the adequate category in the provided list. The CoT must present a valid structure to be accepted.

c) SLCA applies the two assessments, ensuring detected categories are logically and self consistent. SCA is applied once to enhance precision of the result. The union of two LCA allows the model to have two opportunities to show the possibility to create a logically structured CoT, enhancing recall. SLCA is the intersection of the two sub-methods.

Figure 2 : Consistency Assessment Evaluation Study Design



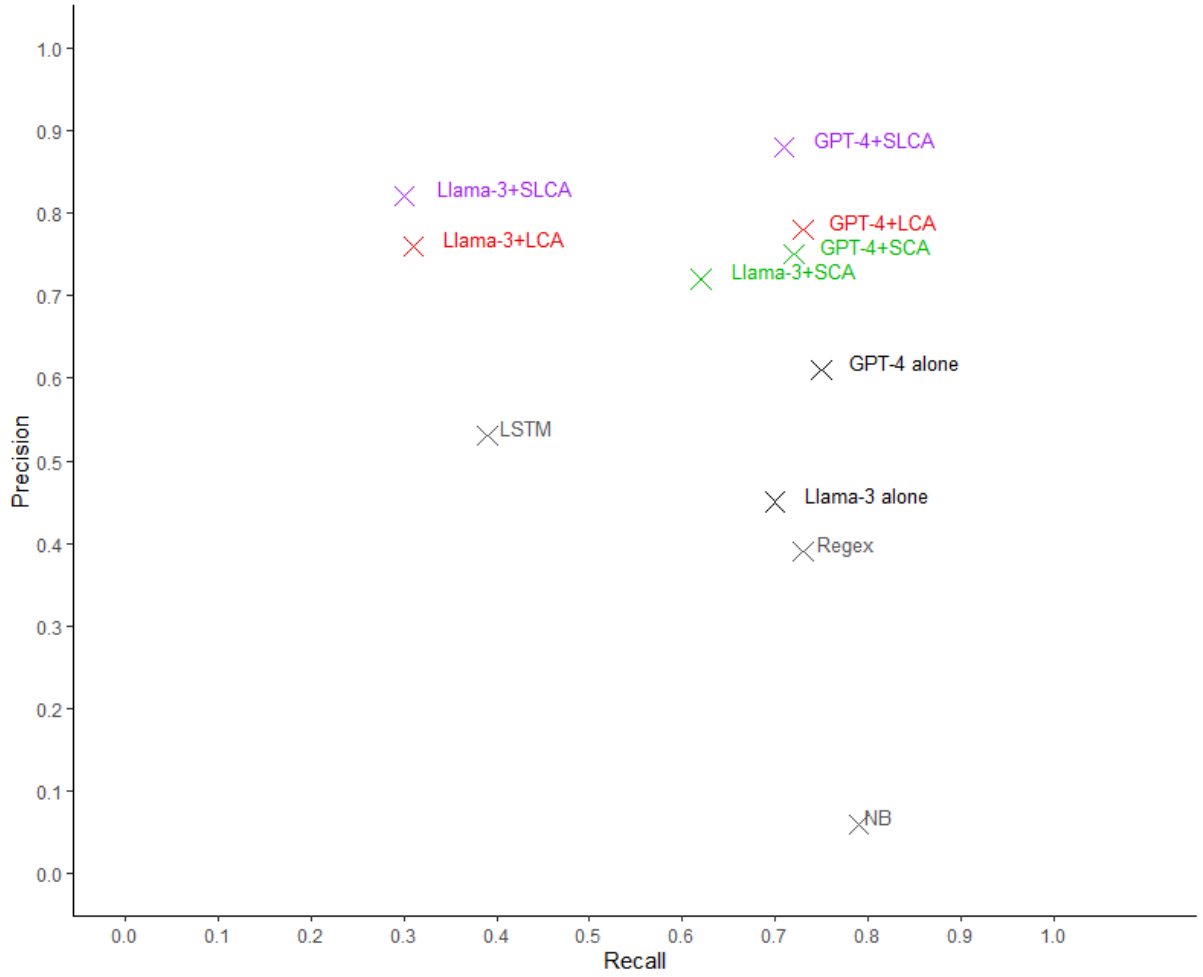
The performances of humans, GPT-4 and GPT-4 consistency assessed are explored through two main experiments.

a) Every agent type runs 3 independent classifications, allowing accurate computation of precision, recall, reproducibility and hallucination rates over 12,600 categories and tones identifications.

b) Eleven models are evaluated. Regex is the decisional tree used in production at our establishment, NB stands for Naive Bayes, LSTM for Long Short Term Memory, Llama-3.1 70B unquantised is the state-of-the-art open-source LLM reasonably usable with

standard hospital computational capabilities and GPT-4 is the state of the art LLM, both LLM being tested with and without SLCAn and its consistency assessment (CA) sub-methods. Their corresponding performances are investigated through a large-scale benchmark of 1,170 feedbacks, i.e. 49,140 categories and tones identifications. This methodology allows to effectively rank available solutions for real care use.

Figure 3 : Benchmark of Machine Learning Models on Patient Feedback Classification (n=1170)



Seven models are compared through a large-scale benchmark of 1170 feedbacks corresponding to 49,140 categories and tones identifications. Precision represents the fraction of correctly identified categories among selected ones. Recall represents the exhaustivity of this selection. Best thresholds are defined for the algorithms accepting thresholds (Naive Bayes and LSTM) as the maximisation of precision time recall.

Bibliography

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* **3**, (2013).
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*. https://www.has-sante.fr/upload/docs/application/pdf/2022-11/iqss_outil_verbatim_note_cadrage_2022.pdf (2022).
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*. (2022).
4. Cammel, S. A. *et al.* How to automatically turn patient experience free-text responses into actionable insights: a natural language programming (NLP) approach. *BMC Med. Inform. Decis. Mak.* **20**, 97 (2020).
5. Khanbhai, M. *et al.* Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review. *BMJ Health Care Inform* **28**, (2021).
6. LMSys Chatbot Arena Leaderboard - a Hugging Face Space by lmsys. https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard.
7. 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.
8. Wang, X. *et al.* Self-Consistency Improves Chain of Thought Reasoning in Language Models. *arXiv [cs.CL]* (2022).
9. Maynez, J., Narayan, S., Bohnet, B. & McDonald, R. On Faithfulness and Factuality in Abstractive Summarization. in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* 1906–1919 (2020).
10. Wei, J. *et al.* Chain of thought prompting elicits reasoning in large language models. *Adv. Neural*

Inf. Process. Syst. **abs/2201.11903**, (2022).

11. Copi, I. M., Cohen, C. & McMahon, K. *Introduction to Logic*. (Routledge, 2016).
12. Chen, C. *et al.* INSIDE: LLMs' Internal States Retain the Power of Hallucination Detection. (2024).
13. 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* https://arxiv.org/pdf/2407.03282 (2024).
14. Jia Li, Ge Li, Yongmin Li, Zhi Jin. Structured Chain-of-Thought Prompting for Code Generation. https://arxiv.org/pdf/2305.06599 (2023).
15. Sultan, M. A., Ganhotra, J. & Astudillo, R. F. Structured Chain-of-Thought Prompting for Few-Shot Generation of Content-Grounded QA Conversations. (2024).
16. Tam, Z. R. *et al.* Let Me Speak Freely? A Study on the Impact of Format Restrictions on Performance of Large Language Models. (2024).

Author Information

Authors and Affiliations

1) Erios : Research and Integration Space for Digital Health Tools (University Hospital Centre) - Montpellier, France

Zeno Loi, Kévin Yauy, Xavier Corbier, David Morquin, Laurine Moniez

2) Infectious and Tropical Diseases Department (University Hospital Centre) - Montpellier, France

David Morquin

3) Care, Quality, Pathways and Users Division (University Hospital Centre) - Montpellier, France

Emilie Prin-Lombardo, Xavier Derzko, Sylvie Gauthier.

4) Epidemiology, Health Data, and Medical Information Department (University Hospital Centre) - Montpellier, France

Zeno Loi, Grégoire Mercier

5) University of Montpellier, France

Zeno Loi, Kévin Yauy, Xavier Corbier

7) LIRMM, Reference Centre for Congenital Anomalies, Clinical Genetic Unit, (University Hospital Centre) - Montpellier, France

Kévin Yauy

8) **CHU Lapeyronie, University of Montpellier - Montpellier, France**

Patrice Taourel

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. X.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.

Acknowledgments

The study was conceived, funded, and executed entirely by Montpellier Hospital University Centre. We acknowledge all participants and professionals involved in the patient feedback processing. Special thanks to the ERIOS team: Cécile Yriarte, Louise Robert, 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 Hospital University Centre of Montpellier for their impactful discussions and contributions.