medRxiv preprint doi: https://doi.org/10.1101/2024.07.11.24310210; this version posted July 16, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Consistency in Large Language Models Ensures Reliable Patient Feedback Classification

Zeno Loi*, David Morquin, Xavier Derzko, Xavier Corbier, Sylvie Gauthier, Laurine Moniez, Emilie Prin-Lombardo, Grégoire Mercier, Kévin Yauy*, University Hospital Center of Montpellier. * Corresponding authors. Email: z-loi@chu-montpellier.fr and kevin.yauy@chu-montpellier.fr

Abstract

Evaluating hospital service quality depends on analyzing patient satisfaction feedback. Human-led analyses of patient feedback have been inconsistent and time-consuming, while natural language processing approaches have been limited by constraints in handling large contexts. Large Language Models (LLMs) offer a potential solution, but their hallucination tendency hinders widespread adoption.

Here we show that Global Consistency Assessment (GCA), a method directing LLM to produce a structured chain of thought as a logical argument and evaluate their reproducibility across two independent predictions, enhances the reliability of LLMs in patient feedback analysis without the use of fine-tuning or annotated dataset.

GCA applied to GPT-4 successfully eliminated GPT-4's 16% hallucination rate, achieving a precision of 87% while keeping a recall of 75% in analyzing 100 patient feedback samples. Furthermore, this method markedly outperforms state-of-the-art models in a benchmark of 1170 feedbacks, with a precision-recall AUC of 89%, compared to the highest score of 59% with standalone models like GPT-4, Llama 3 and classical machine learning.

Consistency assessment provides a reliable and scalable solution for identifying areas of improvement in hospital services and shows promise for any text classification task.

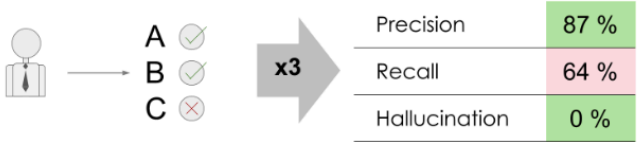
NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

Graphical Abstract

Objective:
Classify 12,600 categories among 100 patient feedback (42 existing categories, 3 independent agents) with:

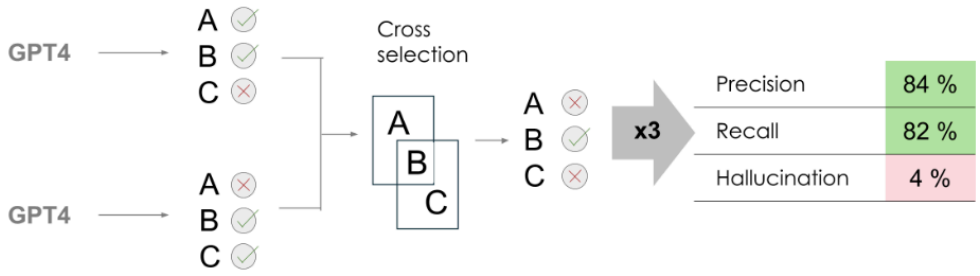
- High precision
- No hallucination
- No annotated dataset

A) Humans quality of care experts

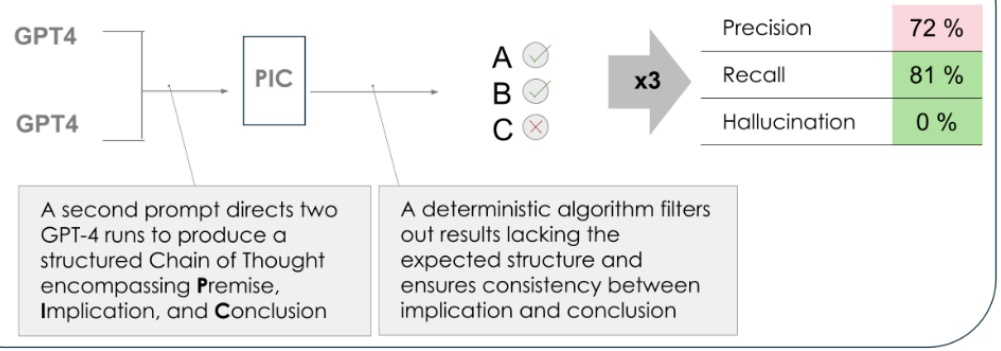


B) GPT-4 standalone

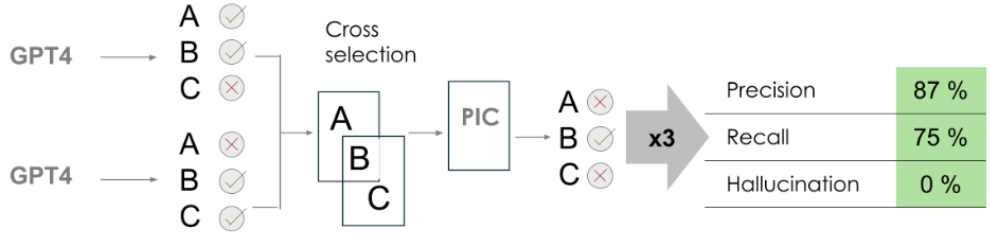
C) External consistency assessment (ECA)



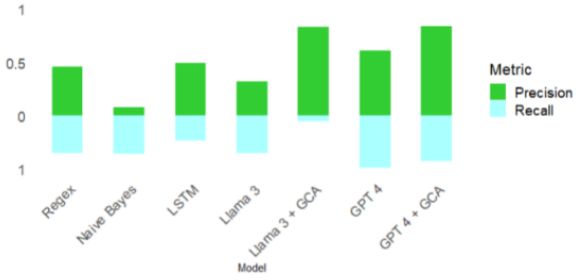
D) Internal consistency assessment (ICA)



E) Global consistency assessment (GCA) : ECA + ICA



F) Models Benchmark on 49,140 classifications among 1,170 feedbacks



Conclusion :

GPT-4 + GCA outperforms both human experts and other models, introducing the concept of consistency assessment as an applicable method for the implementation of LLMs in hospital settings.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.11.24310210; this version posted July 16, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Introduction

Patient satisfaction feedback is a crucial metric for identifying areas of improvement in hospital services, directly impacting the quality of care [1].To effectively manage the substantial volume of feedback, it is essential to structure and classify this information to prioritize improvement efforts. 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, the human-led classification process remains time-consuming and inefficient, highlighting the need for more effective solutions.

Automated methods for analyzing patient feedback have historically fallen short due to the technical limitations of natural language processing (NLP) algorithms. Models such as Naive Bayes and BERT have struggled to accurately classify nuanced feedback due to their inability to handle complex language contexts effectively [2,4–12]. These limitations necessitate the development of more advanced and reliable tools.

Large Language Models (LLMs) offer a promising alternative, with their superior ability to understand natural language and identify subtle nuances in patient feedback [5,13]. Both states of the art, the proprietary model GPT-4 and open source model Llama-3 can be considered. In addition, recent works showed that evaluating the External Consistency of LLMs predictions, the availability of a model to provide the same classification over multiple independent attempts, greatly enhances the performances of LLMs in classification tasks [14]. However, it does not address the tendency of LLMs to produce hallucinations, which is incompatible in sensitive applications such as patient feedback analysis. As the Chain of Thoughts (CoT)

medRxiv preprint doi: https://doi.org/10.1101/2024.07.11.24310210; this version posted July 16, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

methods enhance the explainability of these models [15], they are a potential solution to the hallucinations issue.

Here we describe the Global Consistency Assessor (GCA), a method designed to improve the reliability of LLM-generated predictions. The GCA combines the External Consistency assessment, the evaluation of the LLM ability to provide the same predictions in similar conditions, with a Internal Consistency assessment, which evaluates the capability of an LLM to produce a logically valid chain of thought (CoT).

Results

GPT-4 classification is more exhaustive than humans, but is unsuitable due to hallucinations

To our knowledge, no evaluation study was performed to evaluate human-led classification of patient satisfaction feedback to date. We evaluated the ability of 3 human quality-of-care experts, assessed by a blind investigator, to accurately classify 100 patient feedbacks (Table 1) among 21 categories and two 2 tones (positive/negative) (Appendix 1) adapted from previous unsupervised classifications suggestions [3,12,16] to fit operational standards for hospital quality of care improvement for a total effective of 12,600 classifications. We found that humans were precise (mean precision of 0.87) but not exhaustive (recall : 0.64). Moreover, the classification has been time consuming (3h per 100 feedbacks).

To assess GPT-4 performance to classify patient feedback, we performed 3 independent runs of GPT-4 standalone on the same task (prompt provided in

medRxiv preprint doi: https://doi.org/10.1101/2024.07.11.24310210; this version posted July 16, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Appendix 2). GPT-4 was less precise (mean precision of 0.72) but more exhaustive (recall : 0.87) than humans (McNemar ϱ<1e-15). However, GPT-4 exhibited a significant hallucination rate, representing 16% of all generated identifications. These hallucinations were detected and identified through a human review of all results.

External consistency assessment enhance precision, while only internal consistency reliably mitigates hallucinations

Previous works suggest the use of an External Consistency Assessor (ECA) to enhance LLMs precision by only selecting categories identified by two independent runs.

We evaluated the performances of 3 independent runs of GPT-4+ECA on this task (i.e. 6 generations from GPT-4 to produce 3 predictions). ECA increased GPT-4's precision by 12%. GPT-4+ECA was still less precise (mean precision of 0.84), and more exhaustive (recall : 0.82) than humans (McNemar ϱ<1e-15). However, GPT-4+ECA still presented a 4% hallucination rate.

To address this hallucination deal breaker issue, we developed an Internal Consistency Assessor (ICA) to evaluate the LLM Chain of Thought (CoT) structure without the need for fine-tuning or annotated datasets. Two independent GPT-4 standalone predictions were directed in 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 category identified). A list of valid implications and their compatibility with the categories have been established priorly by the three human experts and were also indicated to the LLM. The LLM CoT accordance to the instructions reflects the

medRxiv preprint doi: https://doi.org/10.1101/2024.07.11.24310210; this version posted July 16, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Internal Consistency of the prediction and only categories identified at least once with valid implications were kept (Figure 1).

Then we evaluated the performances of 3 independent runs of GPT-4+ICA on this task (i.e. 12 generations from GPT-4 to produce 3 predictions). GPT-4+ICA was less precise (mean precision of 0.72) and more exhaustive (recall : 0.81) than humans (McNemar ϱ<1e-15). Notably, GPT-4+ICA successfully removed every hallucination from its predictions.

Global Consistency Assessment applied to GPT-4 outperforms human experts and state-of-the-art models

We investigated the performances of the Global Consistency Assessor (GCA) combining both ECA and ICA. GCA was associated with GPT-4 on the same task. This approach delivered better performances than humans. GPT-4+GCA was equally precise (mean precision of 0.87) and more exhaustive (recall : 0.75) than humans (McNemar ϱ<1e-6), without presenting any hallucination. A thorough human review of the results confirmed the absence of hallucinations, highlighting the robustness and reliability of this combined method.

Finally, we compared GPT-4+GCA to other automated solutions in a large scale benchmark (n=49,140 classifications over 1,170 feedbacks) : GPT-4 standalone, Llama-3+GCA, Llama-3 standalone, Regex (decision tree used in production in our establishment), Long Short Term Memory (LSTM) and Naive Bayes (NB). GPT-4+GCA achieved a precision-recall AUC (pr-AUC) of 0.89, outperforming every other model (Figure 3). Standalone GPT-4 was the second-best solution, showing poor precision (0.67) despite a high recall (0.97), for a pr-AUC of 0.59. Llama-3+GCA ranked third, with an AUC of 0.5. Historical models like Regex and LSTM had lower AUCs of 0.32 and 0.28, respectively. Standalone Llama-3 performed poorly, with a

medRxiv preprint doi: https://doi.org/10.1101/2024.07.11.24310210; this version posted July 16, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

recall similar to Regex (0.70) but with decreased precision (from 0.46 to 0.32). Naive Bayes was the worst option, with an AUC of 0.13 and a maximum precision of 0.3 at any threshold. All p-values for comparison are individually <1e-3.

Performances depend on Internal Consistency metrics quality

In the benchmark subgroup analyses (Appendix 4), the performance of LLMs+GCA in the "Medical and paramedical care" category was particularly low (AUC ϵ [0.02;0.19]). This is likely due to the lower quality of available implications provided for these specific categories, which have been difficult for our teams to describe. This issue was consistent for both GPT-4+GCA and Llama-3+GCA. Conversely, categories such as "Meals and snacks" and "Humanity and availability of professionals - positive" were easier to identify, as evidenced by the high performance of all algorithms in these subgroups (AUC ϵ [0.69;1]). These results could be explained by the more pronounced recurrence of specific vocabulary, resulting in a reduced geometric distance between the feedbacks mentioning them. Moreover, LLMs+GCA tend to perform better in identifying negative tones, with an AUC for GPT-4+GCA of 0.95 and Llama-3+CA of 0.54, compared to 0.88 and 0.48 respectively for positive tones. This effect is inverted for some historical models like LSTM (AUC negative: 0.21, AUC positive: 0.35) and Regex (AUC negative: 0.26, AUC positive: 0.39). It is hard to explain this phenomenon clearly and further research should be conducted for investigation.

Discussions

In this study, we present the Global Consistency Assessment method that enhances the reliability of LLMs in patient feedback classification without the use of fine-tuning

medRxiv preprint doi: https://doi.org/10.1101/2024.07.11.24310210; this version posted July 16, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

or annotated dataset. GPT-4 alone is unsuitable for medical grade patient feedback classifications as it produces hallucinations. The association of External and Internal Consistency Assessment over GPT-4 outperforms both humans and other machine learning models and LLMs.

One of the strengths of our results is demonstrating that we have developed a highly effective model capable of completely avoiding hallucinations on a sample of 100 feedbacks. Although we cannot ensure that our system would eliminate a hallucination occurring after 12,600 classifications, we consider the performance in our use case to be sufficient in clinical context.

Our study highlights an interesting point commonly encountered in the field of text classification: the Gold Standard is not fixed but relative. We observed limited inter-expert reproducibility (Krippendorff's alpha between 3 human experts : 0.67), which makes establishing a consensus Gold Standard challenging. This situation reflects numerous real-life scenarios where evaluating the results of automated language processing is difficult due to the absence of a definitive Gold Standard or its highly subjective nature. Consequently, the evaluation of the large-scale benchmark depends on the evaluator constructing the Gold Standard, potentially leading to variations with different assessors. This relativity underscores the need for adaptable and robust evaluation methods in automated patient feedback analysis, as well as in other domains where subjective judgment plays a significant role. Our approach allows the production of classifications shown as human-leveled in terms of precision and recall with the addition of the possibility to define a more stable and reproducible consensus (Krippendorff's alpha between 3 GPT-4+GCA : 0.80).

Moreover, our method offers the advantage of not requiring fine-tuning, prompt engineering, or prompt tuning, thereby conserving computational resources.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.11.24310210; this version posted July 16, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Additionally, it could be applied to any model and perhaps any text classification task, enhancing its versatility.

As described in the sub-group analysis, some groups present lower performances. This highlights the major stake using internal consistency : the quality of the consistency assessment seems to highly depend on the ability of the engineering team to describe each category, to give a semantic scope of them and thus to list exhaustively all valid implications the LLM will be authorized to use.

One of the most significant limitations of our study is the uncertainty regarding the applicability of our results to all large language models (LLMs). While this approach enhanced the performance of the two models we tested, LLAMA-3+GCA did not match the performance of GPT-4 alone. Further studies are needed to compare the effectiveness of these techniques, particularly with smaller models or those fine-tuned for specific tasks. Additionally, we have not yet explored how our approach complements traditional methods for improving LLMs, such as fine-tuning or other optimization techniques. Additional research is required to investigate these potential synergies and fully understand the broader applicability of our methods."

Although the analysis material was in French while LLMs tend to perform better in English [17], we didn't measure the impact of the language over the different models' performances. Further research is necessary to extrapolate the advantages of our approach among languages.

Our study represents an important step towards the professionalization of large language models (LLMs) in complex text classification tasks, which is particularly useful in the medical field. These results could have significant implications in

medRxiv preprint doi: https://doi.org/10.1101/2024.07.11.24310210; this version posted July 16, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

numerous other domains, paving the way for broader applications of LLMs in various industries.

Our approach is designed to be robust and adaptable, transcending specific tasks and models without the need for additional tuning or development beyond defining the scope of each category. We have shown that this method can classify patient feedback with high accuracy and without hallucination. Future research should investigate its potential as a general-purpose, task-agnostic tool to further affirm its effectiveness. Moreover, this approach could consistently generate human-level reference standards across new datasets, tackling challenges previously unaddressed by humans due to the extensive time required. Consistency might be all you need to classify texts.

Methods

Data and Inclusion Criteria

In France, the national system E-Satis systematically collects patient feedback after hospital stays and subsequently provides extensive data back to the concerned healthcare institutions. This study is based on feedbacks collected via this platform from adult patients hospitalized in the University Hospital Center of Montpellier between 2022 and 2024. Exclusions include data from patients refusing their use, feedbacks too lengthy for analysis by all models, compensation claims, and feedbacks with overly extreme content (Appendix 5). Selected feedbacks are systematically pseudonymized to ensure the protection of personal data.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.11.24310210; this version posted July 16, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Categorization

The classification task in this study always corresponds to the following : classify the feedback among non-exclusive 21 categories and 2 non-exclusive tones (positive and negative) (Appendix 1). The categories are adapted from the categorization proposed by the works of the High Society of Health of France, with the addition of the category “Patient’s Rights” to fulfill the operational purpose of this classification : defining local healthcare quality improvement axes.

Gold standards

Two Gold standards are used in this study. The main analyses (experience 1 to 5) have been conducted over 100 feedbacks selected in accordance with the three human experts to be dense in information and to reflect a large spectrum of compliments and critics. A list of all categories found by human experts and GPT-4 standalone have been blindly evaluated by a fourth, external human quality of care expert to validate or invalidate each category identification.

The Gold standard for the benchmark (n=49,140 classifications over 1,170 feedbacks) has been established by a human quality of care expert alone due to its high time consumption. The sample results from a randomized selection from E-satis 2023 database attached to our facility. The number of feedback to include have been estimated to be able to put in evidence a difference of 2% pr-AUC over 6 bilateral tests (comparing 7 models), with a total alpha error rate of 1% (alpha error rate control method of Bonferroni) and a power of 80%, supporting an attrition up to 10% of the initial materials. The number of categories to include is 44370 and therefore, with a rate of 42 categories identification per feedback, 1057 feedbacks.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.11.24310210; this version posted July 16, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Metrics of interest

The main analyses are focused on three metrics : the precision, which is a prerequisite for medical-grade classification, the recall and the hallucination rate. Additionally, the main analyses were conducted over a rather low number of feedback and with three agents (including GPT-4 runs) to avoid high individual human time-consumption and in order to evaluate reproducibility of the results (see Discussion).

As the benchmark is meant to make abstraction from the threshold chosen for the models capable of estimating probabilities, the metric of focus will be the precision-recall area under the curve (pr-AUC).

Experiment 1 : human experts

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

Experiment 2 : GPT-4 standalone

Three runs of GPT-4 have 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 Appendix 2.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.11.24310210; this version posted July 16, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Experiment 3 : GPT-4+ECA

Three additional GPT-4 runs have classified each feedback. Every classifications were independent. The classifications have been associated in 3 groups of 2 to perform External Consistency assessment. For each one of the three GPT-4+EC, only categories identified twice by the GPT-4 standalone were kept as identified by GPT4+EC. This cross selection aims at identifying only the category consistently identified by GPT-4 (Figure 2).

Experiment 4 : GPT-4+ICA

On top of the six GPT-4 runs, a second, more structured prompt is issued (Appendix 2). This prompt directs the LLM to refine its CoT into a structured logical argument consisting of a premise, an implication, and a conclusion. Specifically, the premise should directly cite the patient feedback, and the implication should come from a predefined valid implications list that delineates the scope for each category (Appendix 1). The conclusion must then logically deduce the appropriate category based on the premise and implication. As typos errors or white spaces trims can skew feedback citations, this part of the CoT is not evaluated. The Internal Consistency is assessed only if the implication is included in the provided list and if the category identified corresponds to this very implication. This structure evaluation approach aims to assess the LLM's reasoning as a coherent argument. Three groups of two GPT-4+IC are formed, and only classifications that present a valid CoT at least once are retained. (Figure 2).

medRxiv preprint doi: https://doi.org/10.1101/2024.07.11.24310210; this version posted July 16, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Experiment 5 : GPT-4+GCA

The three groups of two GPT-4 runs apply successively the External then the Internal method, selecting only the identification that fulfill the two condition :

- Both GPT-4 standalone must have identified the category
- At least one must have provided a valid CoT

This two-step process allows to test the LLM's Internal Consistency—how logically structured and valid the argument is—but also its External Consistency, which evaluates the reproducibility of the LLM's classifications across two independent attempts.

Experiment 6 : large scale benchmark

The benchmark compared 7 models : Naive Bayes (NB), Long Short Term Memory (LSTM), Regex, Llama-3 standalone, Llama-3+GCA, GPT-4 standalone and GPT-4+GCA over the classification of 49,140 categories among 1,170 feedbacks. NB and LSTM have been trained and evaluated in ten fold cross validation. Regex was used in its production version available in our facility. Llama-3 and GPT-4 standalone were not designed to provide multiple thresholds.

The two LLMs in association with the GCA provided thresholds depending on which consistency has been assessed. Every consistency metric fulfilled as follows increased the epistemic probability attached to the mention of the given category :

- Each one of the two category identification (External Consistency) : 12/35
- Each one of the two CoT validated (Internal Consistency) : 4/35
- In addition, if the two runs provided exactly the same implication : 2/35
- If the two runs provided exactly the same citation : 1/35

medRxiv preprint doi: https://doi.org/10.1101/2024.07.11.24310210; this version posted July 16, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Training sets

The training sets for the models were not strictly equivalent. LLMs and Regex benefited from the expertise of quality professionals during their prompt engineering and the decision tree generation, while LSTM and Naive Bayes were trained on a database of only 1,170 feedbacks. Despite the already consequent 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 enhance 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/Consistency-in-Large-Language-Models-Ensures-Reliable-Patient-Feedback-Classification.

Data Availability

The datasets used during the current study are available from the corresponding author on reasonable request. Anonymized 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 University Hospital Center of Montpellier policies.

Ethical Considerations

This study complies with French regulations according to data protection laws. Patients were informed of the usage of their data and had the option to withdraw

medRxiv preprint doi: https://doi.org/10.1101/2024.07.11.24310210; this version posted July 16, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

access at any time. The ethical approval of this study has been waived by the Scientific Ethical Committee of the University Hospital Center of Montpellier (Cécile Yriarte, Yrina Gilhodes, Sandrine Mas, Caroline Dunoyer) as requesting the approval by an ethics committee for this type of research is not possible according to French law.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.11.24310210; this version posted July 16, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Figures and tables

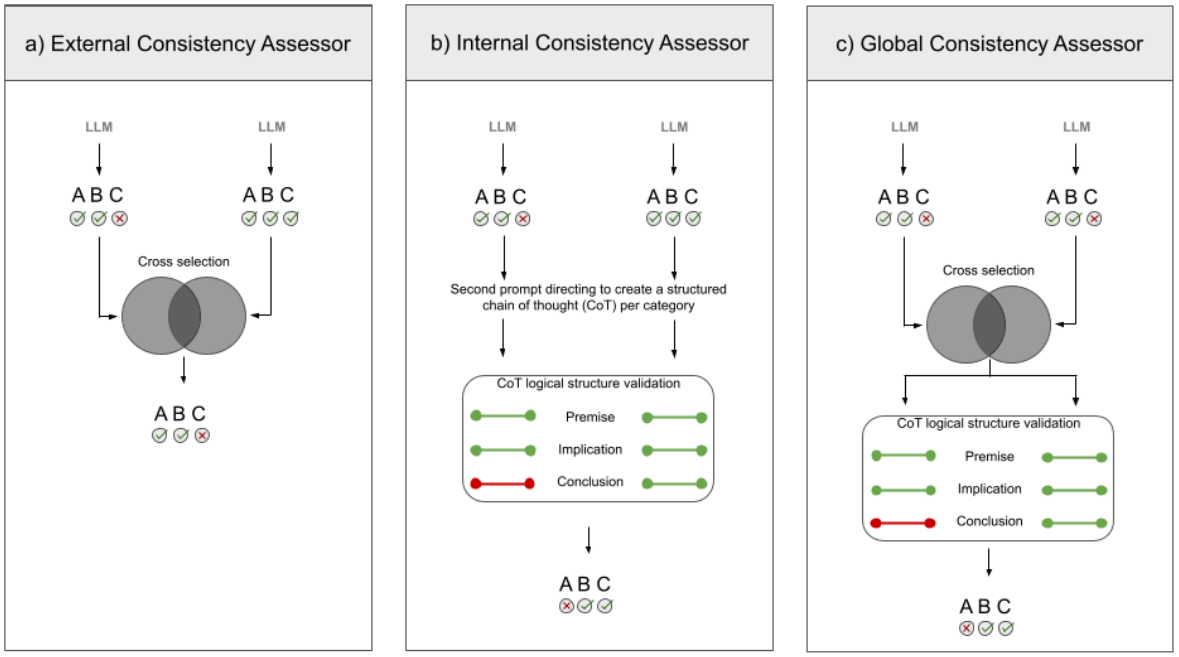
Table 1 : Experiences 1 to 5 - Humans, GPT-4 and GPT-4 consistency assessed performances.

		Mean performances over three independent agents		
Experience	Agents	Precision (%)	Recall (%)	Hallucination rate (%)
1	Human experts	87	64	0
2	GPT-4 standalone	72	87	16
3	GPT-4 + EC	84	82	4
4	GPT-4 + IC	72	81	0
5	GPT-4 + CA	87	75	0

These results are provided analyzing 100 patient feedback independently three times by each agent. As each feedback classification is composed of 21 categories and 2 tones identification, the size of sample for 95% confidence intervals (95%CI) computation is 12,600. All precision and recall 95%CI present a range <1%. ECA stands for external consistency assessment, ICA for internal consistency assessment and GCA for global consistency assessment.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.11.24310210; this version posted July 16, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Figure 1 : Consistency Assessor engineering



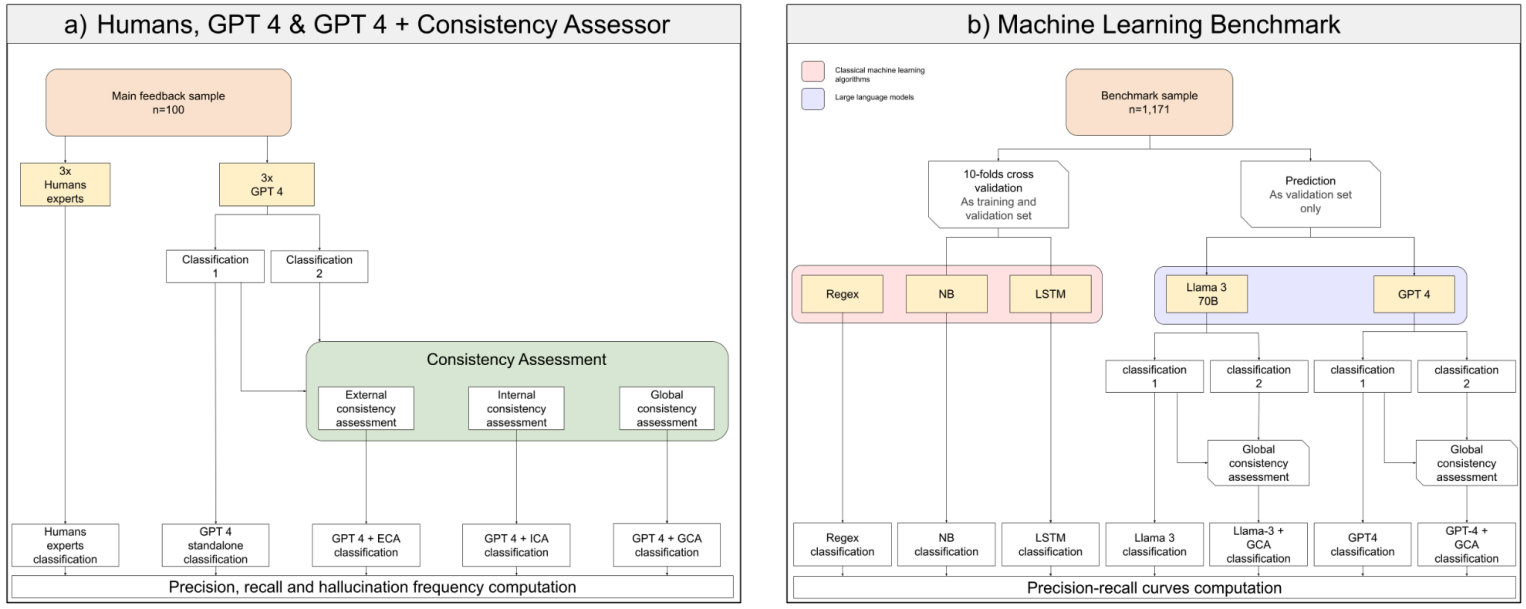
All Consistency methods take in entry two independent LLM predictions.

a) External Consistency assessor (ECA) proceeds to a straightforward cross selection. Only categories identified twice are kept.

b) Internal consistency assessor (ICA) 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 if the implication given by the LLM can be found attached to the adequate category in the provided list. At least one CoT must present a valid structure to be accepted.

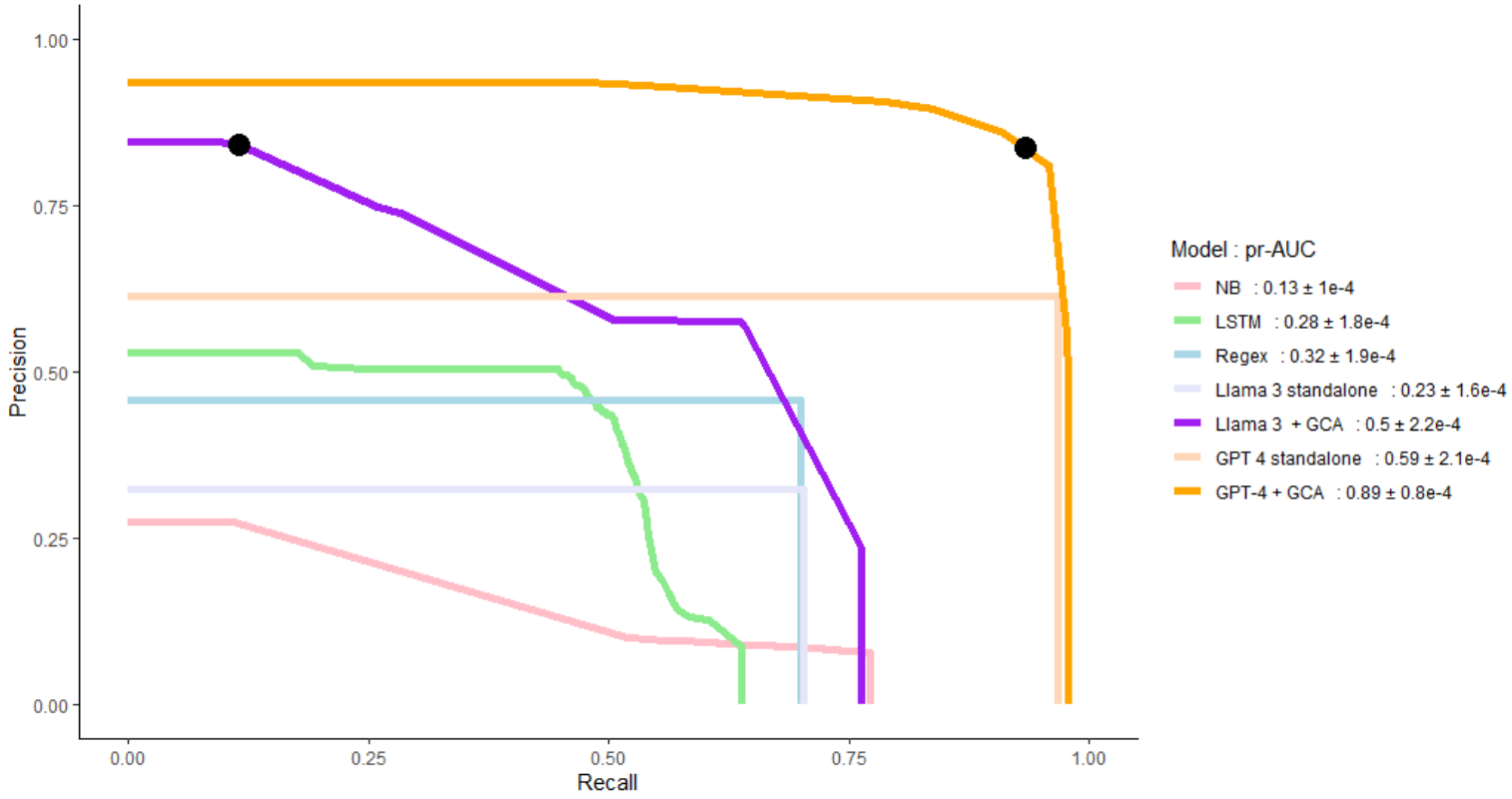
c) Global Consistency assessor applies the ICA to the identifications validated by the ECA.

Figure 2 : Consistency assessor validation method



a) The performances of humans, GPT-4 and GPT-4 consistency assessed are explored through experiences 1 to 5. Every agent type runs 3 independent classifications, allowing to precisely compute precision, recall and hallucination rate over 12,600 categories and tones identifications. b) Seven models are evaluated : Regex is the decisional tree used in production in our establishment, NB stands for Naive Bayes, LSTM for Long Short Term Memory, Llama-3 is the state of the art open-source LLM and GPT-4 is the state of the art LLM, both LLM being tested with and without GCA. their thresholds and their corresponding performances are explored 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 : Experience 6 - Benchmark (n=1,170)



The graph shows the precision-recall curves performed by the 7 tested models and their respective area under the curve (pr-AUC), based on comparison with a human expert produced gold standard. The black dot for the LLMs enhanced by the Consistency Assessor represents the threshold from which every category identification is based on a valid global consistency assessment. GPT-4 + GCA greatly outperforms other models.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.11.24310210; this version posted July 16, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

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 et al.. *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 et al.. *Expérience Des Patients Hospitalisés En France : Analyse Nationale Des Commentaires Libres Du Dispositif E-Satis*. (2022).
4. Ranard, B. L. *et al.* Yelp Reviews Of Hospital Care Can Supplement And Inform Traditional Surveys Of The Patient Experience Of Care. *Health Aff.* **35**, 697–705 (2016).
5. Martino, Steven C. et al.. Using Natural Language Processing to Code Patient Experience Narratives: Capabilities and Challenges.
6. Hawkins, J. B. *et al.* Measuring patient-perceived quality of care in US hospitals using Twitter. *BMJ Qual. Saf.* **25**, 404–413 (2016).
7. Gallan, A. S., Girju, M. & Girju, R. Perfect ratings with negative comments: Learning from contradictory patient survey responses. *Patient Experience Journal* **4**, 15–28 (2017).
8. Fairie, P. *et al.* Categorising patient concerns using natural language processing techniques. *BMJ Health Care Inform* **28**, (2021).
9. Doing-Harris, K., Mowery, D. L., Daniels, C., Chapman, W. W. & Conway, M. Understanding patient satisfaction with received healthcare services: A natural language processing approach. *AMIA Annu. Symp. Proc.* **2016**, 524–533 (2016).
10. 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).
11. Chekijian, S., Li, H. & Fodeh, S. Emergency care and the patient experience: Using sentiment analysis and topic modeling to understand the impact of the COVID-19 pandemic. *Health Technol.* **11**, 1073–1082 (2021).

medRxiv preprint doi: https://doi.org/10.1101/2024.07.11.24310210; this version posted July 16, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

12. Assmann, K. *et al.* Expérience des patients: valorisation et analyse nationale des commentaires des patients recueillis dans le cadre du dispositif national e-Satis. *Rev. Epidemiol. Sante Publique* (2022) doi:10.1016/j.respe.2022.01.005.
13. 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).
14. Wang, X. *et al.* Self-Consistency Improves Chain of Thought Reasoning in Language Models. *arXiv [cs.CL]* (2022).
15. Wei, J. *et al.* Chain of thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **abs/2201.11903**, (2022).
16. Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A. & Donaldson, L. Use of Sentiment Analysis for Capturing Patient Experience From Free-Text Comments Posted Online. *J. Med. Internet Res.* **15**, e2721 (2013).
17. Huang, H. *et al.* Not All Languages Are Created Equal in LLMs: Improving Multilingual Capability by Cross-Lingual-Thought Prompting. *arXiv [cs.CL]* (2023).

medRxiv preprint doi: https://doi.org/10.1101/2024.07.11.24310210; this version posted July 16, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Author informations

Authors and affiliations

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

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

Infectious and tropical diseases department (University Hospital Center) - Montpellier, France

David Morquin

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

Emilie Prin-Lombardo, Xavier Derzko, Sylvie Gauthier.

Department of Epidemiology, Health Data, and Medical Information (University Hospital Center) - Montpellier, France

Zeno Loi, Grégoire Mercier

University of Montpellier, France

Zeno Loi, Kévin Yauy, Xavier Corbier

Data Sciences Institute of Montpellier, France

Xavier Corbier

LIRMM, CNRS, Reference center for congenital anomalies, Clinical Genetic Unit, (University Hospital Center) - Montpellier, France

Kévin Yauy

medRxiv preprint doi: https://doi.org/10.1101/2024.07.11.24310210; this version posted July 16, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

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. L.M. contributed to figures design. G.M. contributed to the ethical and legal framework of the study. K.Y. contributed to study coordination, study design, figures design and writing of the manuscript. All authors contributed to the critical review of the manuscript.

Competing interests

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

Acknowledgments

The study was conceived, funded, and executed entirely by the Hospital University Center of Montpellier. There was no industry support of any kind.