

<https://doi.org/10.1038/s44387-025-00036-4>

Evaluating GPT-4's ability to generate informed consent material for genetic testing



Eirini Petrou^{1,4}✉, Kelly E. Ormond^{1,4}, Dominik Stambach^{2,3}, Elliott Ash², Oliver Buchholz¹ & Effy Vayena¹

The informed consent (IC) process is essential in genetic testing, yet IC materials are often difficult to read and understand, influencing patients' decision-making. Large language models may improve the accessibility and clarity of these materials. We used GPT-4 to generate IC materials for Non-Invasive Prenatal Testing (NIPT) and hereditary breast and ovarian cancer testing (BRCA) in English, German, Italian, and Greek, using zero-shot prompting and retrieval-augmented generation. Healthcare providers evaluated GPT-4-generated and human-generated materials using a previously published framework. GPT-4 performed well on structured components, such as explaining the purpose and benefits of testing, but struggled with nuanced ethical and contextual content. Respondents overall preferred human-written materials, underscoring limitations in current GPT-4-generated material for health communication in the genetic testing context. GPT-4's performance in German, Italian and Greek was generally weaker than in English, highlighting potential language-specific challenges in GPT-4-generated IC content.

Generative Artificial Intelligence (GenAI) is being explored for its potential to enhance patient-facing healthcare communication. Large language models (LLMs) could significantly impact the informed consent (IC) process, a critical step in medical decision-making that requires clear, accurate, and accessible patient information. However, traditional IC documents often exceed recommended readability levels, limiting patient understanding and leading to a suboptimal consent process^{1–4}.

LLMs, such as ChatGPT3.5 and GPT-4, may help address these challenges by adapting content to patient literacy levels and reducing the explanatory burden on healthcare providers^{5,6}. However, their effectiveness in generating IC materials remains underexplored, with existing studies reporting mixed results. While GPT-4 can enhance readability with structured prompting^{7–11}, its performance varies based on disease complexity, prompt specificity, and information accuracy^{12–15}. Some studies suggest GPT-4-generated IC materials meet minimum standards¹¹, while others highlight critical omissions, particularly regarding procedural risks, benefits, and testing alternatives¹⁶.

Most studies assessing GPT-4's performance focus on IC forms for surgical procedures^{7,10,11}. However, the IC process differs from an IC form: the former is a dynamic, interactive exchange between patients and providers, while the latter is a static document with predefined content. This

distinction raises questions about GPT-4's adaptability across different IC structures. Additionally, concerns remain about LLM performance across languages^{17,18}.

This study evaluates GPT-4's ability to generate IC materials for two widely administered genetic tests: Non-Invasive Prenatal Testing (NIPT) and hereditary breast and ovarian cancer (BRCA) testing. While LLMs have shown some utility in conveying genetic information^{19,20}, they remain limited, particularly in understanding inheritance patterns¹⁹. To our knowledge, no prior study has systematically evaluated the accuracy, quality, and accessibility of GPT-4-generated IC materials for genetic testing. Additionally, no research has directly compared GPT-4-generated and human-generated IC materials across different languages. By addressing these gaps, we aim to determine whether LLMs can support the IC process across different genetic tests and linguistic backgrounds by replacing the human-made IC materials.

Results

Demographics

A total of 65 participants started the survey and 25 completed it (completion rate: 38%; dropout rate: 62%). The average completion time was approximately 30 minutes, excluding responses with survey times exceeding

¹Health Ethics and Policy Lab, Department of Health Sciences and Technology, ETH Zurich, Zurich, Switzerland. ²Center for Law and Economics, Department of Humanities, Social and Political Sciences, ETH Zurich, Zurich, Switzerland. ³Center for Information and Technology Policy, Princeton University, Princeton, New Jersey, US. ⁴These authors contributed equally: Petrou Eirini, Kelly E. Ormond. ✉e-mail: Eirini.petrou@hest.ethz.ch

Table 1 | Demographics

		NIPT (N = 12)	BRCA (N = 13)
Professional affiliation	Genetic Counselor	2 (16.7%)	6 (46.2%)
	MD Geneticist	3 (25%)	2 (15.4%)
	OBGYN	3 (25%)	1 (7.7%)
	Midwife	2 (16.7%)	0
	Oncologist	0	1 (7.7%)
	MD Other	2 (16.7%)	1 (7.7%)
	Laboratory Geneticist	0	2 (15.4%)
Gender	Female	9 (75%)	10 (77%)
	Male	3 (25%)	3 (23%)
Language of medical practice	English	2(17%)	1 (8%)
	German	3 (25%)	6 (46%)
	Italian	3 (25%)	1 (8%)
	Greek	4 (33%)	5 (38%)
Age	Mean age of participants: 46.7 (SD: 12.6)		
Years of practice	Mean years of practice: 17.5 (SD: 11.5)		

24 hours ($N = 6$). Table 1 summarizes participant demographics. Our sample included 25 participants, 48% offering NIPT and 52% BRCA testing. Most (76%) were female, with an average of 17.5 years of clinical experience across genetic and non-genetic specialties. Greek and German were the most common language groups; English was the least represented, with three participants (12%).

Readability

Table 2a (NIPT) and 2b (BRCA) present readability scores assessed using standard metrics (see Methods section), along with GPT-4-generated material, which was consistently shorter than human-generated ones. Among the analyzed languages, Italian materials were the most difficult to read, followed by German. Greek materials were the easiest, with GPT-4 achieving a lower readability score (3.13) than human-generated material (4.71).

For NIPT (Table 2a, excluding English), GPT-4-generated material was consistently more readable than human-generated material, with the largest difference in German (9.9 vs. 11.3, respectively).

In contrast, readability varied by language for BRCA (Table 2b). GPT-4-zero-shot materials, that is, material generated without feeding GPT-4 prior examples or domain-specific context²¹, were slightly less readable in English and Italian than GPT-4-RAG-generated material. In German, both GPT-4-Zero-Shot and GPT-4-RAG had lower readability scores than human-generated material.

Error evaluation

For NIPT, participants identified a mean of 1.0 errors (range: 0–3) in GPT-4-generated material, with one outlier reporting more than four errors in the patient-facing material. Participants identified more errors in the human-generated material, with a mean of 2.0 (range: 0–4).

For BRCA, participants reported a mean of 1.0 errors in both GPT-4-Zero-Shot (range: 0–3) and GPT-4-RAG (range: 0–2). One participant reported more than four errors in GPT-4-RAG-generated material. Errors in the human-generated material were relatively infrequent (mean 0.7, range 0–2).

IC components

Participants evaluated the extent to which various IC components for genetic testing were included in the NIPT (Fig. 1a) and BRCA materials (Fig. 1b), based on a predefined framework (see Table 3 and Methods section). The figures present data combined across the four languages.

For NIPT, GPT-4- and human-generated materials performed similarly in *Test reason*, *Test aim*, *Test benefit*, *General results*, and *Clinical*

Table 2 | a Readability scores for NIPT: b Readability scores for BRCA

LANGUAGE	MATERIAL	SCORE AND READING LEVEL	WORD COUNT
a			
NIPT			
English	GPT-4 Generated	10.9, college	481
English	Human-generated	9.4, 10th to 12th grade	627
German	GPT-4 Generated	9.9, College	446
German	Human-generated	11.3, College	664
Greek	GPT-4 Generated	3.13, 5th grade or lower	294
Greek	Human-generated	4.71, 5th grade or lower	728
Italian	GPT-4 Generated	14.9, College graduate	312
Italian	Human-generated	15.9, College graduate	632
b			
BRCA			
English	GPT-4-Generated (Zero-shot)	10.7, college	465
English	GPT-4-Generated (RAG)	10.4, college	502
English	Human-generated	9.5, college	754
German	GPT-4-Generated (Zero-shot)	11, College	394
German	GPT-4-Generated (RAG)	11.7, College	338
German	Human-generated	13.2, College graduate	767
Italian	GPT-4-Generated (Zero-shot)	18.4, College graduate	335
Italian	GPT-4-Generated (RAG)	17.4, College graduate	359
Italian	Human-generated	15.7, College graduate	803
Greek	GPT-4-Generated (Zero-shot)	3.13, 5th grade or lower	272
Greek	GPT-4-Generated (RAG)	3.13, 5th grade or lower	238
Greek	Human-generated	3.13, 5th grade or lower	839

Table 3 | List of “necessary and critical” concepts for Informed Consent for genetic testing (modified from Ormond et al.⁴⁶)

– Genetic testing is always voluntary
– Patient is required to give informed consent
– Why this test is being done (generally)
– What is the aim of the test?
– What are the benefits associated with the test
– What are the risks associated with the test
– What results will be returned (generally)
– What other types of non-primary results will be returned, if any
– Patients can decide for the return of non-primary results
– How will prognosis and management be impacted, if at all
– The results may also impact family in different ways (e.g emotions, relationships) if the patient wants to share the result
– Patients’ options for future steps after the test
– What are the clinical limitations of the test
– What happens after the patient gets the results
– To whom are the results reported

limitations. However, human-generated material provided more comprehensive coverage in *Test risk* and *Actions after the results* components. Conversely, GPT-4-generated material scored higher in covering the *Prognosis and management*, *Impact on families*, *Future steps*, and *Who gets the results* components.

Both approaches covered key components well up to *General results*, though human-generated material received lower ratings in *Voluntariness*

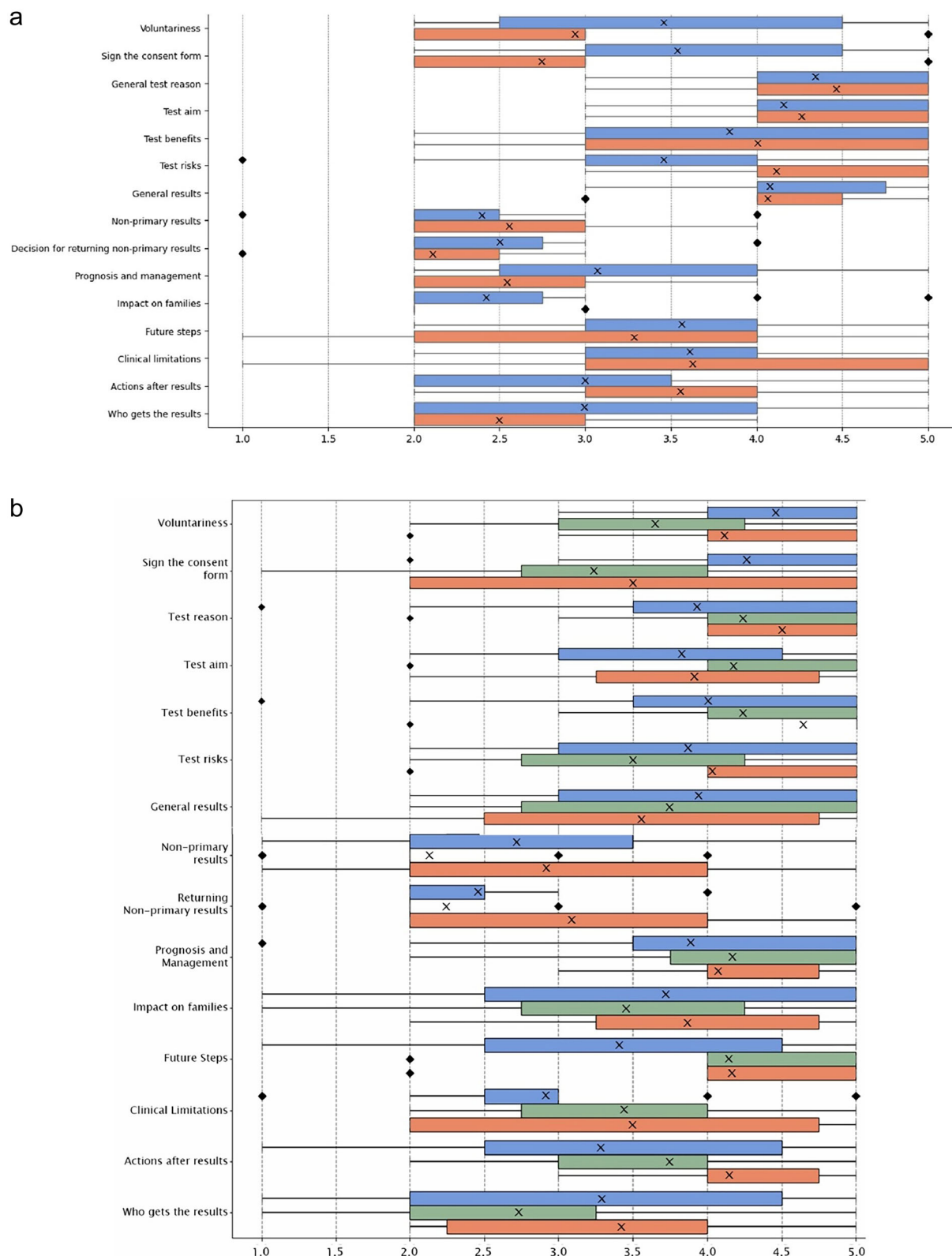


Fig. 1 | **a** Inclusion of components of Informed Consent on NIPT Materials. 1 – this topic is not applicable to material related to this genetic test; 2 – Applicable but not addressed at all in the material; 3 – Indirectly addressed; 4 – Briefly addressed but not clearly stated; 5 – Sufficiently addressed. X – Median ♦ – Outliers. Blue boxes: GPT-4-generated material. Orange boxes: Human-generated material. Variables on the y-axis correspond to the 15-point list we developed based on Ormond et al.⁴⁶ (pg. 8). The sequence of the variables on the boxplot follows both the original framework and our survey design. **b** Inclusion of components of Informed Consent on BRCA

Materials. 1 – this topic is not applicable to material related to this genetic test; 2 – Applicable but not addressed at all in the material; 3 – Indirectly addressed; 4 – Briefly addressed but not clearly stated; 5 – Sufficiently addressed. X – Median ♦ – Outliers. Blue boxes: GPT-4-Zero-shot generated material. Green boxes: GPT-4-RAG generated material. Orange boxes: Human-generated material. Variables on the y-axis correspond to the 15-point list we developed based on Ormond et al.⁴⁶ (pg. 8). The sequence of the variables on the boxplot follows both the original framework and our survey design.

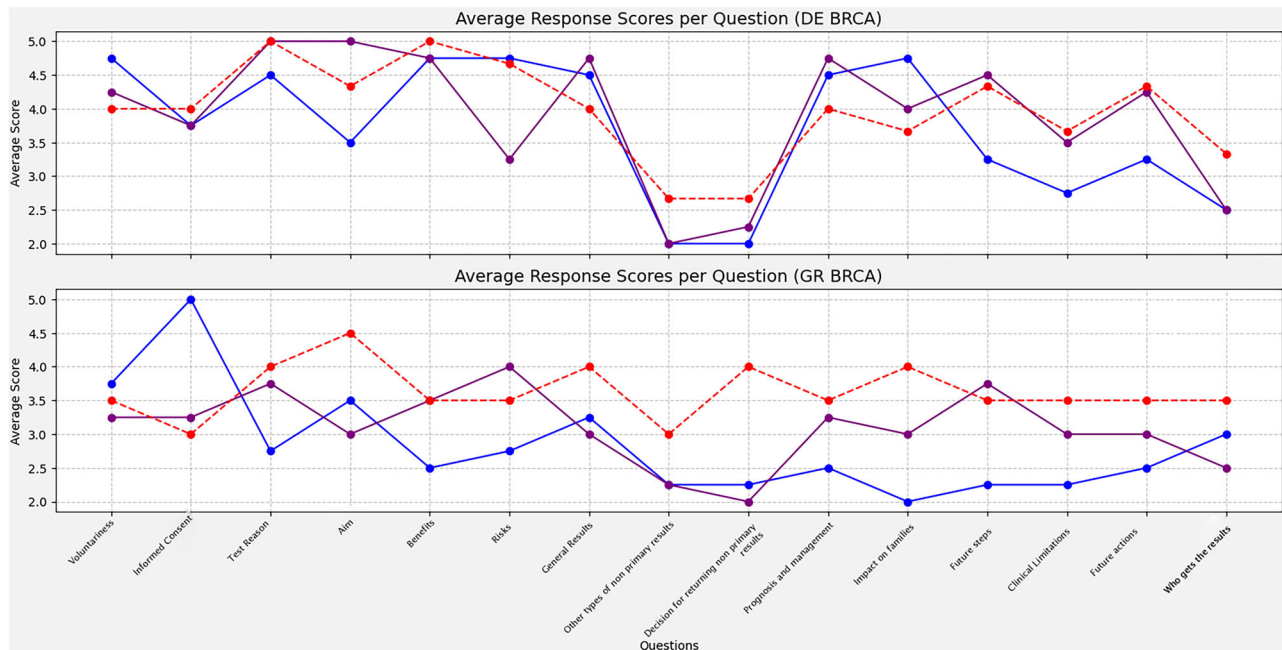


Fig. 2 | Inclusion of components of Informed Consent: German versus Greek this topic is not applicable to material related to this genetic test; 2 – Applicable but not addressed at all in the material; 3 – Indirectly addressed; 4 – Briefly addressed but not clearly stated; 5 – Sufficiently addressed. **Blue line:** GPT-4-Zero-shot generated

material. **Purple line:** GPT-4-RAG generated material. **Red dotted line:** Human-generated material. Figure 2 compares the average responses to Likert-scale questions in German (top-line chart) and Greek (bottom-line chart) regarding including IC components in the BRCA materials.

and *Sign the consent form*. The downward shift in the boxplots indicates a clear decline in overall ratings for the *Non-primary results*, *Decision for returning non-primary results*, *Prognosis and management*, and *Impact on families* components. While GPT-4-generated material exhibited greater variability in ratings, human-generated material maintained more stable ratings across participants.

For BRCA, participant ratings varied considerably across GPT-4- and human-generated material. The GPT-4-Zero-Shot approach exhibited the greatest variability, particularly in *Test benefits*, *Actions after the results*, and *Who gets the results*. GPT-4-RAG-generated showed more consistent evaluations, though it received slightly lower ratings overall.

Despite this variability, GPT-4-Zero-Shot outperformed GPT-4-RAG in eight of 15 categories, particularly in core components (*Voluntariness*, *Sign the consent form*), risk and result disclosure (*Test risk*, *General results*, *Non-primary results*, *Decision for returning non-primary results*), and personal and familial implications (*Impact on families*, *Who gets the results*). However, while GPT-4-Zero-Shot outperformed GPT-4-RAG in covering *Non-primary results* and *Decision for returning them*, both approaches still received low ratings in these areas, suggesting a shared weakness in handling these IC components.

Human-generated material performed best in *Test reason* (median: 4.5), *Test risk* (median: 4.0), and *Actions after the results* (median: 4.0). By contrast, RAG-generated material showed inconsistencies, with extreme outliers in *Test reason*, *Non-primary results*, *Decision for returning non-primary results*, and *Future steps*, where ratings fluctuated significantly from the medians.

Participants' preferred choice

Participants preferred human-generated material for both NIPT (3.45 ± 1.63 vs. 3.36 ± 1.37 for GPT-4) and BRCA (4.00 ± 1.33 vs. 3.9 ± 1.16 for GPT-4 Zero-Shot, 3.5 ± 1.04 for RAG). For NIPT, participants were more accurate in distinguishing human- from GPT-4-generated material. However, for BRCA, they were more likely to misidentify GPT-4-RAG-generated material as human-written (66.6%), suggesting that GPT-4-RAG outputs more closely resembled human writing in this context.

Case study: Analysis of the effect of language on BRCA materials

The corresponding results are presented in Fig. 2. In the German BRCA case, human-generated material received the highest mean scores for *Test reason* (4.6 ± 0.8) and *Test Benefit* (4.8 ± 0.4). In Greek, it consistently scored higher across IC components and showed less variability in participants' responses (range: 3.0–4.5).

Among GPT-4-generated material, GPT-4-RAG scored highest in German for *Test aim*, *General results*, and *Prognosis and management*. GPT-4-Zero-Shot outputs were consistently inadequate in covering IC components in Greek, while GPT-4-RAG's scores aligned more with human-generated material. Notably, GPT-4-RAG outperformed other methods in *Test risk* (mean: 4.0, SD: 0.63) and *Future steps* (mean: 3.75, SD: 1.09), but this was not observed across all IC components.

Non-Primary results and *Decision-making for returning non-primary results* components were consistently underrepresented in both languages. In German BRCA material, all three versions (human, GPT-4-Zero-Shot, and GPT-4-RAG) scored poorly, with little to no discussion of these topics. A review of the translated texts confirmed their absence. In contrast, Greek human-generated material provided more coverage, scoring higher (*Non-Primary results*: 3.0 ± 1.0 ; *Decision-Making*: 3.5 ± 0.87) than both GPT-4-Zero-Shot and GPT-4-RAG-generated material.

Discussion

Our findings showed that GPT-4-generated materials for both NIPT and BRCA cases remained difficult to read according to established readability thresholds for patient-facing information^{22,23}. However, when evaluated using the same metrics, they were, in some cases, easier to read than the human-generated materials. The model did not hallucinate across its outputs. Our analysis showed notable differences in how respondents evaluated the materials, particularly those generated by GPT-4. These inconsistencies raise key questions about the model's reliability and potential role in the IC process. Part of this variation could also reflect the inherent variability in both human- and GPT-4-generated IC materials. That is, human-generated text can be shaped based on the author's background, communication style, and institutional norms. This can significantly limit its standardization as a benchmark. Similarly, GPT-4-generated outputs are subject to variability

mainly due to the probabilistic nature of the model per se, meaning outputs can differ across sessions even with identical prompts. This dual variability should be considered when interpreting these types of findings, as it reflects the real-world variation that occurs in both clinical communication and LLM-generated text. Acknowledging this dual variability strengthens the transparency of our methodological approach and contextualizes the inconsistencies observed in both content coverage and evaluation.

GPT-4 struggled to generate NIPT and BRCA materials at the readability levels recommended by leading health organizations^{24,25} when no specific instructions were provided. Readability varied widely, particularly in GPT-4-generated material. These results align with previous research in general medical fields^{7,26–30} but contrast with studies where ChatGPT-3.5 and GPT-4 were explicitly instructed to simplify consent forms^{7,10,11}. Since GPT-4 was not given pre-written text to simplify, it may have generated material at a higher reading level than expected. GPT-4 Zero-shot learning resulted in slightly harder-to-read material than GPT-4 RAG in both English and Italian, partially aligning with Lai et al.³¹, who found that GPT-4 zero-shot learning underperforms across different languages. GPT-4 generated the most readable texts in Greek. However, this is possibly due to limitations in existing readability assessment tools (e.g., SMOG) that are not optimized for Greek's morphology and semantic density. Overall, GPT-4's readability scores closely matched those of human-generated IC material, suggesting that its default text generation mimics human writing unless explicitly instructed otherwise.

GPT-4-generated materials included some IC components but omitted others, with variation across languages. In German, GPT-4-RAG-generated material outperformed human-written materials, likely due to RAG's ability to retrieve structured knowledge from reliable sources^{32,33}. This is particularly relevant for BRCA testing, where rising demand^{34,35} may have expanded databases, improving RAG's accuracy. However, GPT-4-RAG struggled with non-primary results, yielding lower scores in this area. These findings align with research showing ChatGPT's difficulty in addressing complex genetics-related questions³⁶ further highlighting its limitations in capturing nuanced IC components regardless of the prompting technique used. GPT-4's challenges in these components may stem from the distinction between an IC form and an IC process. As a model trained on large text corpora, GPT-4 may be more suited to generating static IC forms that follow standardized formats, such as those commonly used in medical settings where risks, benefits, and procedures are relatively consistent and widely documented. Moreover, as previously noted, most studies evaluating LLMs for IC have focused on surgical settings. These studies^{7,11} typically involve feeding the model existing IC forms, often sourced from large medical centers¹¹, and prompting it to simplify the content. This may contribute to GPT-4's stronger performance with such materials, as it aligns closely with both its training data and the structure of the evaluation tasks. In contrast, genomic testing often requires more individualized, context-sensitive information, raising concerns about GPT-4's ability to generate consent materials that go beyond the scope of standard form templates.

Additionally, GPT-4 underperformed in Greek, highlighting language biases in AI development. Similar disparities have been observed in Japanese¹⁷ and Spanish¹⁸ compared to English. This supports the assumption that less commonly spoken languages receive weaker AI performance across domains^{17,18,37,38}. Language-based disparities in healthcare have been linked to reduced primary care utilization and poorer health outcomes^{37,38}, raising concerns about the potential for AI-driven inequities in medical information accessibility.

Participants' preference for human-generated materials for both NIPT and BRCA was modest. This trend was more pronounced for BRCA, where GPT-4-RAG-generated material was frequently misidentified as human-written. This suggests that RAG may not only improve informational content but also enhance tone, structure, and style in ways that closely resemble human-generated materials. This, however, also implies that users need to provide relevant information, which can be more challenging for those new to the task or those with limited time. Nevertheless, it overall indicates that prompting should be viewed not merely as a technical step to

produce an output, but as a critical design decision that can significantly influence both content quality and audience reception.

While participants preferred the human-generated material, the relatively small differences in ratings and the difficulty distinguishing GPT-4-generated from human text suggest that, in some contexts, GPT-4 may already be producing content that meets user expectations at least at the level of surface communication. This reinforces our approach to evaluating materials not only in terms of readability, but also in terms of content completeness and clinical relevance. Finally, the variation in participants' ability to identify GPT-4-generated text in NIPT compared to BRCA further suggests that topic complexity or familiarity may influence how GPT-4-generated materials are interpreted. These findings challenge the assumption that human authorship is inherently superior across different contexts⁹. However, they also align with existing literature that demonstrates human difficulty in distinguishing between ChatGPT-generated medical manuscripts and those written by humans. This has important implications for the medical community, particularly regarding the circulation of inaccurate material and the risk of increased public distrust⁴⁰.

This study has several limitations. First, we deliberately sought evaluations from healthcare providers. While this ensured expert assessments, it excluded patients, the primary end-users of IC material. This is a critical limitation, as patient feedback is essential to evaluate whether GPT-4-generated content is clear, relevant, and accessible to its intended audience. Future studies should incorporate patient-centered evaluation. For example, small-scale cognitive interviews or think-aloud sessions could help assess how patients interpret and engage with GenAI-generated IC materials. We are currently conducting a separate study using a think-aloud protocol with patients, which directly builds on this limitation by exploring how patients engage with GenAI-generated IC materials.

Second, our analysis focused on two genetic tests (NIPT and BRCA), limiting the generalizability of these findings to other genetic contexts. However, we observed key patterns, including the omission of certain IC components without explicit prompting, elevated readability levels, and variable expert confidence in patient-facing material quality. These challenges are not unique to NIPT or BRCA and may similarly affect GPT-4-generated materials in other contexts, including carrier screening or whole-genome sequencing, suggesting broader relevance. Future research should assess whether these patterns persist across additional testing scenarios.

Third, our small sample size ($N = 25$) and uneven distribution across languages pose constraints, and readers should consider the findings preliminary. Specifically, the sample size limits the ability to capture the full range of variation in GPT-4-generated material and provider assessments, thereby reducing the generalizability of the results. Also, the uneven representation of languages limits our ability to draw robust conclusions about language-specific patterns or to generalize across linguistic and cultural contexts. Some findings, therefore, may reflect features unique to specific language groups rather than broader trends in language use. Also, convenience sampling may have attracted individuals with a stronger interest in IC processes or genetic education, introducing selection bias. This could have influenced how participants engaged with the materials, possibly resulting in evaluations that are not fully representative of the broader clinical community. Moreover, the small sample size limits the ability to explore variation across professional roles, language groups, and test types. As a result, generalizability is limited, and the findings should be interpreted with caution. Future studies should use larger and more diverse samples to capture broader perspectives.

Another limitation of this study is the exclusive use of the GPT-4 model. We did not include other LLMs, such as Gemini, Copilot, or medical models like Med-PaLM and Med-Mistral, as they were outside the scope of the study. Such LLMs are likely to differ in clinical accuracy, terminology, and style. For instance, the Mistral 8 x 22 B LLM has shown promise in enhancing the readability, understandability, and actionability of IC forms without compromising accuracy or completeness⁴¹. While this highlights the potential of domain-specific models, our focus on a general-purpose model like GPT-4 strengthens the relevance of our findings to broader, real-

world clinical contexts where fine-tuned models may not be readily accessible. Finally, while we carefully designed our prompts, they did not account for patient-specific factors, such as literacy level, clinical history, gender, or age. Although our approach enabled a controlled comparison between GPT-4- and human-generated IC material, it did not allow for personalized content generation. Future research should explore LLM-generated IC material tailored to individual patient needs through personalized prompts.

To conclude, GPT-4 struggled to produce comprehensive IC material, failing to address *all* IC components for NIPT and BRCA testing. Similar results were observed across both testing scenarios and all examined languages, including English. Despite these limitations, the model performed well in structured IC components, such as explaining the purpose of the test, its intended benefits, and the general aim for testing. These components often follow standardized formats and appear in publicly available patient-facing health materials. Considering this, GPT-4 may be most effective in generating standardized patient instructions, medical protocols, or discharge summaries rather than IC materials. GPT-4-RAG-generated materials were more often perceived as human-authored, showed better readability than human-written materials in German and zero-shot outputs in English and Italian, and received more consistent evaluations from participants. Although these differences were not statistically significant, they suggest that RAG may offer practical advantages over zero-shot prompting in complex clinical communication tasks, such as IC for genetic testing, particularly in non-English languages. Integrating explicit instructions through RAG may improve model performance by ensuring more complete coverage of IC components. Its performance in German, Italian, and Greek was poorer compared to English. If LLM-generated IC materials favor English-language content, non-English-speaking patients may receive lower-quality health information, further exacerbating existing inequities. Addressing these challenges requires a multifaceted strategy: improving dataset curation, applying multilingual fine-tuning using high-quality, domain-specific texts from underrepresented languages, and designing culturally adapted prompts that reflect local examples, idioms, and healthcare structures. These, along with post-generation validation techniques, should be prioritized as technical, methodological, and ethical imperatives. For now, a hybrid approach, where GPT-4 generates material and clinicians review and refine it, may be more effective for the IC process in genetic testing.

Methods

We generated IC materials using GPT-4, first in English and then in German, Italian, and Greek. GPT-4 was selected because it was the most current model available at the time of the study, claimed to surpass its predecessors in reasoning capabilities²¹. Most prior studies relied on ChatGPT-3 or ChatGPT-3.5^{7,8,12,18,19,26}. We then conducted an online survey in which healthcare providers evaluated GPT-4-generated IC materials compared to human-generated ones. The study was approved by the ETH Zurich ethics review board (EK 2024-N-154).

Justification for test and language selection

We used GPT-4⁴² to generate accessible information for two types of genetic tests: NIPT and BRCA. We selected these genetic test scenarios because they are common forms of genetic testing, including in Switzerland⁴³. Additionally, referrals for these tests are relatively straightforward²³ and both genetic specialists (medical geneticists and genetic counselors) and non-genetic specialists (obstetricians, midwives, and oncologists) can offer them⁴⁴. This information was produced in English, German, Italian, and Greek. We selected these languages for three reasons. First, English is the most well-documented language in the literature on GenAI and the primary language of most scientific publications. Second, we included German, Italian, and Greek, as these are the primary languages spoken by our research team and are commonly spoken languages in Switzerland (German and Italian) and throughout Europe. Third, research suggests that GenAI systems typically perform worse in languages other than English^{17,18} highlighting a gap that needs to be addressed further.

Generation of IC materials

We conducted all experiments using the ‘gpt-4-turbo-2024-04-09’ model. We briefly experimented with GPT-4o, but our experience was that this model generated less accurate and comprehensive IC sheets. For increased reproducibility, we have set the temperature in all experiments to zero. Prompt development was an iterative process led by a Language Processing (NLP) engineer (DS) and a genetic counselor (KEO) with expertise in genetic testing, IC and bioethics. All prompts used in this study are detailed in the Supplemental Material.

GPT-4 generated adequate patient information for NIPT without requiring additional instructions, indicating that zero-shot prompting, which refers to generating output without prior examples or domain-specific context²¹, was sufficient. However, the output for BRCA testing lacked clinical relevance and included vague statements, such as “ask questions” or “consider your options” without addressing key elements of IC. Moreover, outputs often focused narrowly on BRCA1 or BRCA2, failing to reflect current multigene testing practices. These limitations prompted us to introduce retrieval-augmented generation (RAG)^{21–23}. RAG is a technique used to enhance AI-generated responses by enabling the model to access information from external sources, such as databases or documents. By adopting RAG, the generated outputs are usually more accurate and reliable, as they derive from the model’s built-in language abilities, which utilize real-world information retrieved as needed²³. In this case, by using RAG, we supplemented the system prompt with information from the National Cancer Institute’s website⁴⁵.

English prompts were first translated into German, Italian, and Greek using DeepL Pro, with accuracy verified by native speakers (Supplemental Material). These translated prompts were then input into GPT-4 to generate the corresponding information in each language, reflecting how healthcare providers might use LLMs in real-world settings. The model consistently responded in the language of the input prompt.

Development of the survey

We developed the survey drawing on Ormond et al.⁴⁶ framework for IC components in genetic testing (see Table 3). The survey (see Supplemental Material) evaluated the accuracy, relevance, accessibility, and inclusion of these IC components in the GPT-4-generated and human-generated materials. We used five-point Likert scales and multiple-response formats to assess the materials comprehensively. Participants were not informed that they were partially evaluating GPT-4-generated material, but were debriefed upon survey completion per the study protocol. Not disclosing the origin of the patient-facing material ensured that participants were not influenced by preconceptions about the source of the text when evaluating the content.

Participants and recruitment

Our target population included medical and laboratory geneticists, genetic counselors, obstetricians, midwives providing NIPT, oncologists, breast specialists, and surgeons offering BRCA testing. Participants were required to obtain an IC for at least one of these tests and see patients in English, German, Italian, or Greek. Recruitment was conducted via convenience sampling from June 2024 to January 2025. We used available staff lists of Swiss and Greek hospitals. We also distributed study information through the Transnational Alliance for Genetic Counselors (TAGC) emailing list, the Swiss Genetic Counselor Association email list and the SNPPT Newsletter ($N=498$). To supplement these recruitment approaches, members of the research team personally recruited potential participants attending the European Society of Human Genetics (ESHG) conference (June 2024) and a Rare Disease Justice Workshop held in January 2025. Finally, study information was posted on several researchers’ LinkedIn pages in January 2025. All participants provided IC electronically prior to survey participation.

Administration of the survey

We administered the survey using Qualtrics, an online tool. We developed four survey versions, each tailored to one of the languages assessed. We used

the software's branch, display, skip and randomization features to ensure that participants only evaluated the consent material for the genetic test they offer (either NIPT or BRCA) and that the human- and GPT-4-generated scenarios were randomized in their order of presentation. This randomization minimized bias and increased the responses' validity.

Data analysis

We conducted a descriptive analysis of completed surveys ($N = 25$; NIPT: 12, BRCA: 13) to evaluate expert assessments. Readability was assessed using standard metrics: Flesch-Kincaid⁴⁷ for English, German, and Italian material and the Simple Measure of Gobbledygook⁴⁸ (SMOG) for Greek.

We evaluated participants' evaluations of accuracy, relevance and inclusion of IC components by summarizing their responses using means, standard deviations, and interquartile ranges (IQRs). Given the small sample size in each language, we opted against comparative or inferential analyses due to low statistical power. Data analysis was conducted using IBM SPSS Statistics (Version 29.0.2.0) and Excel (Version 16.93.1), with Python (Version 3.12.0) used to generate line charts for the BRCA case study.

A genetic counselor and native English speaker (KEO) reviewed all materials, checking human-generated material for accuracy and GPT-4-generated material for hallucinations, meaning confidently presented information that diverges from source inputs, lacks factual grounding, and may be misleading, inaccurate, or irrelevant, often due to encoding and decoding errors in the LLM^{49–51}. We present a case study comparing Greek and German BRCA materials, as both had equal sample sizes ($N = 9$). This enabled a structured comparison of GPT-4-generated and human-generated materials.

Data availability

The datasets generated and analyzed during the current study (Table 1, Fig. 1a, b, and Fig. 2) are available in a Figshare repository at the following private link: <https://figshare.com/s/02823eae72ec0676fee>. The Excel file includes separate tabs corresponding to the genetic tests and IC materials examined in each language. Researchers can access the link supporting this study's findings by submitting a reasonable request to the corresponding author.

Received: 6 May 2025; Accepted: 18 August 2025;
Published online: 03 November 2025

References

- Albright, J. et al. Readability of patient education materials: implications for clinical practice. *Appl. Nurs. Res.* **9**, 139–143 (1996).
- Badarudeen, S. & Sabharwal, S. Readability of patient education materials from the American Academy of Orthopaedic Surgeons and Pediatric Orthopaedic Society of North America websites. *J. Bone Jt. Surg.-Am. Vol.* **90**, 199–204 (2008).
- Wang, S. W., Capo, J. T. & Orillaza, N. Readability and comprehensibility of patient education material in hand-related web sites. *J. Hand Surg.* **34**, 1308–1315 (2009).
- Mirza, F. N. et al. Using ChatGPT to facilitate truly informed medical consent. *NEJM AI* **1**, <https://doi.org/10.1056/aics2300145> (2024).
- Vaira, L. A. et al. Evaluating AI-Generated informed consent documents in oral surgery: a comparative study of ChatGPT-4, Bard gemini advanced, and human-written consents. *J. Cranio-Maxillofacial Surg.* **53**, <https://doi.org/10.1016/j.jcms.2024.10.002> (2024).
- Allen, J. W., Schaefer, O., Mann, S. P., Earp, B. D. & Wilkinson, D. Augmenting research consent: should large language models (LLMs) be used for informed consent to clinical research? *Res. Ethics* <https://doi.org/10.1177/17470161241298726> (2024).
- Decker, H. et al. Large language model-based chatbot vs surgeon-generated informed consent documentation for common procedures. *JAMA Netw. Open* **6**, e2336997 (2023).
- Gill, B., et al. ChatGPT is a promising tool to increase readability of orthopedic research consents. *J. Orthop. Trauma Rehabil.* <https://doi.org/10.1177/22104917231208212> (2024).
- Abreu, A. A. et al. Enhancing readability of online patient-facing content: the role of AI chatbots in improving cancer information accessibility. *J. Nat. Comprehensive Cancer Netw.* **22**, <https://doi.org/10.6004/jnccn.2023.7334> (2024).
- Patel, I., Om, A., Cuzzzone, D. & Garcia Nores, G. Comparing ChatGPT vs. surgeon-generated informed consent documentation for plastic surgery procedures. *Aesthetic Surg. J. Open Forum.* <https://doi.org/10.1093/asjof/ojae092> (2024).
- Ali, R. et al. Bridging the literacy gap for surgical consents: an AI-human expert collaborative approach. *npj Digit. Med.* **7**, 1–6 (2024).
- Currie, G., Robbie, S. & Tually, P. ChatGPT and patient information in nuclear medicine: GPT-3.5 versus GPT-4. *J. Nucl. Med. Technol.* **51**, 307–313 (2023).
- Truhn, D., Reis-Filho, J. S. & Kather, J. N. (2023). Large language models should be used as scientific reasoning engines, not knowledge databases. *Nat. Med.* **29**, 2983–2984 (2023).
- Horiuchi, et al. Comparison of the diagnostic accuracy among GPT-4 based ChatGPT, GPT-4V based ChatGPT, and radiologists in musculoskeletal radiology. *MedRxiv (Cold Spring Harbor Lab.)* <https://doi.org/10.1101/2023.12.07.23299707> (2023).
- Ahimaz, P., Bergner, A. L., Florido, M. E., Harkavy, N. & Bhattacharyya, S. Genetic counselors' utilization of ChatGPT in professional practice: a cross-sectional study. *Am. J. Med. Genet. - Part A.* <https://doi.org/10.1002/ajmg.a.63493> (2023).
- Hofmann, H. L. & Vairavamurthy, J. Large language model doctor: assessing the ability of ChatGPT-4 to deliver interventional radiology procedural information to patients during the consent process. *CVIR Endovasc.* **7**, <https://doi.org/10.1186/s42155-024-00477-z> (2024).
- Ando, K. et al. A comparative study of English and Japanese ChatGPT responses to anaesthesia-related medical questions. *BJA Open* **10**, 100296 (2024).
- Gonzalez Fiol, A. et al. Accuracy of Spanish and English-generated ChatGPT responses to commonly asked patient questions about labor epidurals: a survey-based study among bilingual obstetric anesthesia experts. *International J. Obstetric Anesthesia* **61**, 104290 <https://doi.org/10.1016/j.ijoa.2024.104290> (2024).
- Walton, N. et al. Evaluating ChatGPT as an agent for providing genetic education. *BioRxiv (Cold Spring Harbor Laboratory)*. <https://doi.org/10.1101/2023.10.25.564074> (2023).
- Nazareth, S. et al. Hereditary cancer risk using a genetic chatbot before routine care visits. *Obstet. Gynecol.* **138**, 860–870 (2021).
- Shneider, P., Chernysheva, A., Govorov, A., Khlopotov, M. & Nikiforova, A. Applying retrieval-augmented generation for academic discipline development: insights from zero-shot to tree-of-thought prompting. In *Proc. 36th Conference of Open Innovations Association*, 741–747 (FRUCT, 2024).
- Borgeaud, S. et al. Improving language models by retrieving from trillions of tokens. *ArXiv:2112.04426 [Cs]*. <https://arxiv.org/abs/2112.04426> (2022).
- Gao, Y. et al. Retrieval-augmented generation for large language models: a survey. *ArXiv.org*. <https://doi.org/10.48550/arXiv.2312.10997> (2023).
- Wasir, A. S., Volgman, A. S. & Jolly, M. Assessing readability and comprehension of web-based patient education materials by American Heart Association (AHA) and CardioSmart online platform by American College of Cardiology (ACC): How useful are these websites for patient understanding?. *Am. Heart J. Cardiol. Res. Pract.* **32**, 100308 (2023).
- CDC. Health Literacy. Health Literacy. https://www.cdc.gov/health-literacy/?CDC_AAref_Val=https://www.cdc.gov/healthliteracy/pdf/simply_put.pdf (2024).
- Cocci, A. et al. Quality of information and appropriateness of ChatGPT outputs for urology patients. *Prostate Cancer Prostatic Dis.* **23**, 103–108 (2023).

27. Sahin, S. et al. Evaluating ChatGPT-4's performance as a digital health advisor for otosclerosis surgery. *Front. Surg.* **11**, <https://doi.org/10.3389/fsurg.2024.1373843> (2024).
28. McCarthy, C. J., Berkowitz, S. A., Ramalingam, V. & Ahmed, M. Evaluation of an artificial intelligence chatbot for delivery of interventional radiology patient education material: a comparison with societal website content. *J. Vasc. Interventional Radiol.* **34**, 1760–1768, <https://doi.org/10.1016/j.jvir.2023.05.037> (2023).
29. Paran, M., Almog, A., Dreznik, Y., Nesher, N. & Kravarusic, D. A new era in medical information: ChatGPT outperforms medical information provided by online information sheets about congenital malformations. *J. Pediatric Surg.* **60**, <https://doi.org/10.1016/j.jpedsurg.2024.161894> (2024).
30. Walker, H. L. et al. Reliability of medical information provided by ChatGPT: assessment against clinical guidelines and patient information quality instrument. <https://doi.org/10.2196/47479> (2023).
31. Lai, V. D. et al. ChatGPT Beyond English: towards a comprehensive evaluation of large language models in multilingual learning. *Arxiv.* <https://doi.org/10.48550/arxiv.2304.05613> (2023).
32. Lewis, P. et al. Retrieval-augmented generation for knowledge-intensive NLP Tasks. *ArXiv.org.* https://arxiv.org/abs/2005.11401?utm_source=chatgpt.com (2020).
33. Lakatos, R., Pollner, P., Hajdu, A. & Tamás, J. Investigating the performance of retrieval-augmented generation and domain-specific fine-tuning for the development of AI-driven knowledge-based systems. *Mach. Learn. Knowl. Extract.* **7**, 15–15 (2025).
34. Desai, S. & Jena, A. B. Do celebrity endorsements matter? Observational study of BRCA gene testing and mastectomy rates after Angelina Jolie's New York Times editorial. *BMJ* **355**, i6357 (2016).
35. Lippi, G. The risk of unjustified BRCA testing after the "Angelina Jolie effect": how can we save (laboratory) medicine from the Internet?. *Clin. Chem. Lab. Med.* **56**, e33–e35 (2018).
36. Khosravi, T., Sudani, A. & Morteza O. To what extent does ChatGPT understand genetics? *Innov. Educ. Teach. Int.* **61**, 1320–1329 (2023).
37. Jaeger, F. N., Pellaud, N., Laville, B. & Klausner, P. The migration-related language barrier and professional interpreter use in primary health care in Switzerland. *BMC Health Serv. Res.* **19**, <https://doi.org/10.1186/s12913-019-4164-4> (2019).
38. Eslier, M. et al. Association between language barrier and inadequate prenatal care utilization among migrant women in the PreCARE prospective cohort study. *Eur. J. Public Health* **33**, 403–410 (2023).
39. Wang, S. & Huang, G. The impact of machine authorship on news audience perceptions: a meta-analysis of experimental studies. *Commun. Res.* **51**, 815–842 (2024).
40. Helgeson, S. A. et al. Human reviewers' ability to differentiate human-authored or artificial intelligence-generated medical manuscripts. *Mayo Clin. Proc.* **100**, 622–633 (2025).
41. Shi, Q. et al. Transforming informed consent generation using large language models: insights, best practices, and lessons learned for clinical trials. *JMIR Med. Inform.* <https://doi.org/10.2196/68139> (2025).
42. OpenAI, Achiam, S. & Adler, S. GPT-4 Technical Report. *ArXiv:2303.08774 [Cs]*. <https://arxiv.org/abs/2303.08774> (2023).
43. Manegold-Brauer, G. et al. A new era in prenatal care: non-invasive prenatal testing in Switzerland. *Swiss Med. Weekly.* <https://doi.org/10.4414/smw.2014.13915>.
44. Nelson, H. D. et al. Risk assessment, genetic counseling, and genetic testing for BRCA-related cancer: systematic review to update the U.S. preventive services task force recommendation. In PubMed. Agency for Healthcare Research and Quality (US). <https://pubmed.ncbi.nlm.nih.gov/24432435/>.
45. National Cancer Institute. (2024, July 19). BRCA mutations: cancer risk & genetic testing. *Nat. Cancer Inst.* <https://www.cancer.gov/about-cancer/causes-prevention/genetics/brca-fact-sheet>
46. Ormond, K. E. et al. Defining the critical components of informed consent for genetic testing. *J Pers. Med.* **11**, 1304 (2021).
47. Flesch, R. A new readability yardstick. *J. Appl. Psychol.* **32**, 221–233 (1948).
48. Mc Laughlin, G. H. SMOG grading-a new readability formula. *J. Read.* **12**, 639–646 (1969). <https://www.jstor.org/stable/40011226>.
49. Salamin, A.-D. Russo, D. & Rueger, D. ChatGPT, an excellent liar: how conversational agent hallucinations impact learning and teaching. *Proceedings of the 7th International Conference on Teaching, Learning and Education.* <https://doi.org/10.33422/6th.iaceti.2023.11.100> (2023).
50. Bruno, A., Mazzeo, P. L., Chetouani, A., Tiiba, M. & Kerkouri, M. A. Insights into Classifying and Mitigating LLMs' Hallucinations. *ArXiv.org.* <https://arxiv.org/abs/2311.08117> (2023).
51. Beutel, G., Eline G. & Kielstein, J. T. Artificial hallucination: GPT on LSD? *Crit. Care* **27**, <https://doi.org/10.1186/s13054-023-04425-6> (2023).

Acknowledgements

We would like to thank Dr Mattia Andreoletti for proofreading the GPT-4- and human-generated materials in Italian. No special funding was obtained for this paper.

Author contributions

Paper conception: E.V. and E.A. Development of methodology: E.V., E.A., K.E.O., E.P., D.S. Planning and instrument development: E.P., K.E.O., and D.S. Data analysis: E.P. and K.E.O. Data validation: E.P. and O.B. Figures and tables preparation: E.P. Drafting: E.P. and D.S. Review and editing: E.P., K.E.O., D.S., O.B., E.V. and E.A.

Funding

Open access funding provided by Swiss Federal Institute of Technology Zurich.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s44387-025-00036-4>.

Correspondence and requests for materials should be addressed to Eirini Petrou.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025