Computers in Human Behavior: Artificial Humans 10 (2026) 100410

Contents lists available at ScienceDirect

Computers in Human Behavior: Artificial Humans

journal homepage: www.journals.elsevier.com/computers-in-human-behavior-artificial-humans

Evaluating Large Language Models as artificial patients: An exploratory English and German silicon sampling study in oncology decision-making

Charles Ma [a,*], Niklas Hoerner [b], Miriam Meckel [a]

[a] Institute for Media and Communications Management, University of St. Gallen, Blumenbergplatz 9, St. Gallen, 9000, Switzerland
[b] Clinical Neurotechnology Lab, Charité Universitätsmedizin Berlin, Virchowweg 19, Berlin, 10117, Germany

ARTICLE INFO

Keywords:
Silicon sampling
Large language models
Quantum AI in oncology
Artificial patients
Patient treatment decision-making
Synthetic respondents

ABSTRACT

Whether persona-conditioned large language models (LLMs) can reproduce individual responses in clinical scenarios with unfamiliar emerging technologies remains uncertain. This exploratory study evaluated LLM-based human simulation (silicon sampling) in an oncology role-play involving a hypothetical quantum AI recommendation source. Demographic and psychographic profiles from 23 participants at Charité Universitätsmedizin Berlin were used to prompt GPT-5.1, Claude 4.5 Opus, DeepSeek V3.2, and Kimi K2 Thinking. Models chose between two treatments and rated decision confidence under conflicting recommendations from a physician, tumor board, and purported quantum AI system. Each model was prompted in English and German on six dates over approximately five weeks. Modal choices and mean confidence ratings were compared with human responses, with uncertainty estimated using participant-level bootstrapping. Across model–language conditions, modal decision accuracy ranged from 0.57 to 0.65, with weak chance-corrected agreement (Cohen's κ: 0.11–0.29). Confidence mean absolute error ranged from 11.00 to 18.06 points on the 0–100 scale, while concordance was near zero (concordance correlation coefficient: −0.16 to 0.06). All 95% confidence intervals for Cohen's κ and the concordance correlation coefficient included zero. Prompt language showed no consistent advantage in alignment, and temporal consistency was high (decision stability of 89.13% to 97.83%). These exploratory findings suggest that persona-conditioned LLMs may capture aggregate response structures but fall short of the individual-level fidelity required to replace human participants in high-stakes clinical scenarios.

1. Introduction

Artificial intelligence (AI) is already reshaping medicine, and quantum technologies (QT) are increasingly promoted as a possible next computational layer for digital health and precision medicine. The convergence of AI and QT in medicine has been proposed for applications spanning advanced diagnostic accuracy, optimal personal treatments, and real-time clinical decision-making support (Flöther & Griffin, 2023; Jeyaraman et al., 2024). At the same time, systematic assessment of the field shows that the empirical evidence base remains limited; only a small minority of studies use realistic operating conditions, and no consistent performance advantage over classical methods has yet been established (Gupta, Wood, et al., 2025). This combination of high promise and limited evidence makes so-called quantum AI a particularly revealing case for studying how patients may respond to emerging clinical technologies.

Whether such systems are accepted in practice will depend on more than technical performance alone. AI-enabled decision aids can support shared decision-making by providing personalized recommendations and, in some contexts, by creating more time for doctor-patient discussions (Hassan et al., 2024). However, patients and clinicians also report concerns about understandability, bias, transparency, loss of human contact, and how AI systems fit within informed consent (Longoni et al., 2019; Witkowski et al., 2024). These concerns are likely to be amplified for a prospective quantum AI system. Unlike familiar medical technologies, quantum AI is not yet a routine clinical object that patients can evaluate from everyday experience. In this study, we use a prospective quantum AI system as a hypothetical source cue and quantum AI itself is not evaluated. Judgments about it must be assembled from broader dispositions, including trust in doctors, trust in AI, technological affinity, and previous healthcare experiences. In that sense, a hypothetical quantum AI scenario offers a stringent test of how people reason about an emerging technology under uncertainty.

One promising way to study such questions is silicon sampling.

* Corresponding author. Blumenbergplatz 9, 9000, St. Gallen, Switzerland.
E-mail addresses: charles.ma@unisg.ch (C. Ma), niklas.hoerner@charite.de (N. Hoerner), miriam.meckel@unisg.ch (M. Meckel).

https://doi.org/10.1016/j.chbah.2026.100410
Received 9 July 2026; Received in revised form 6 October 2026; Accepted 21 October 2026
Available online 22 October 2026
2949-8821/© 2026 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Argyle et al. (Argyle et al., 2023) introduced the term to describe conditioning a large language model (LLM) on real participants' backstories in order to generate "silicon samples" or artificial respondents, and showed that properly conditioned GPT-3 could emulate response distributions across human subgroups. Subsequent work extended this approach to experimental replication and public opinion, but it also identified important constraints. Demographic-only prompting often performs poorly, richer issue-relevant covariates can improve fidelity substantially, and performance can vary across populations, domains, and response formats (Lee et al., 2024; Ma, 2025). These findings suggest that silicon sampling is most informative when persona prompts capture who a respondent is demographically, as well as what they believe and have experienced.

Within healthcare, related LLM research shows both opportunity and caution. General-purpose models can perform some diagnostic tasks at levels approaching physicians while remaining weaker on triage, and recent work argues that LLMs in medicine should be evaluated as agents embedded in realistic clinical simulations rather than only through static benchmarks (Levine et al., 2024; Mehandru et al., 2024). Other studies show that multi-agent simulated patient systems can produce accurate and stable medical encounters for educational use (Yu et al., 2025). However, this literature largely evaluates clinical reasoning or workflow support; it does not directly address whether a persona-conditioned model can reproduce a specific patient's treatment choice when physicians and a novel technological system differ on recommendations. At the same time, critical work warns that psychologically plausible outputs can still be unreliable under prompt variation or on novel items, making direct validation against human responses essential (Schröder et al., 2025) (Schröder et al., 2025).

Furthermore, despite this growing body of work, there are an additional couple of gaps that remain. The question of how prompting language affects LLM fidelity in this context has received little attention. While V. Gupta et al. (Gupta et al., 2025b) and Vadlapati (Vadlapati, 2024) have shown that LLM performance tends to be better in English than in other languages, the implications for silicon sampling accuracy in German, the language in which the original interviews were conducted, remain unexplored. The temporal consistency of LLM outputs is also an underexamined issue; as Barrie et al. (Barrie et al., 2025) have warned, LLM findings often cannot be re-run, let alone replicated, yet few silicon sampling studies have systematically tested replicability across multiple sampling dates.

Against this background, the present study provides an exploratory validation of silicon sampling in a high-stakes clinical scenario involving an emerging technology. In this scenario, "quantum AI" serves as a recommendation source label and technology framing for a hypothetical oncology decision support system. Using demographic and psychographic profiles from 23 participants in a realistic oncology role-play at Charité Universitätsmedizin Berlin, we prompted four LLMs to simulate each participant's treatment choice and decision confidence in both English and German across six repeated sampling dates. Because the human validation sample was small, the outcome was binary, and the design contained three scenario configurations, the study was not designed to support confirmatory claims about individual-level LLM fidelity. We therefore treat model-human alignment metrics as descriptive effect-size estimates accompanied by uncertainty intervals.

The study contributes to the silicon sampling literature in three ways. First, it examines LLM-based human simulation in a high-stakes clinical scenario involving quantum AI in oncology. Because this emerging technology is not yet part of routine clinical experience, it functions as a sparse domain limit case in which participants' judgments likely depend on latent dispositions rather than direct experience. Second, it compares English and German prompting in a setting where the original human data were collected in German. Third, it evaluates the temporal consistency of LLM-generated decisions and confidence ratings across repeated sampling dates. After reviewing related work in the next section, we develop a theoretical framework that conceptualizes the vignette as a source-conflict task and derives the research questions guiding the analyses.

2. Related work

The conceptual foundation for this study comes from early work on LLM-based human simulation. Argyle et al. (Argyle et al., 2023) framed persona-conditioned LLM outputs as a question of algorithmic fidelity and showed that models could approximate subgroup response distributions when conditioned on thousands of real sociodemographic backstories. Aher et al. (Aher et al., 2022) (2023) broadened the agenda by introducing "Turing Experiments", which test whether models can reproduce patterns from human-subject research in economic, psycholinguistic, and social-psychology paradigms. Gerosa et al. (Gerosa et al., 2023), in turn, argued that synthetic text generation could augment data collection through persona-based interviews, focus groups, and surveys, while still cautioning that AI cannot replace the nuanced and empathetic understanding provided by real human participants. Sarstedt et al. (Sarstedt et al., 2024) reviewed this literature in the context of consumer and marketing research and found that results vary considerably across different domains, arguing that silicon samples hold promise in upstream parts of the research process such as qualitative pretesting and pilot studies. Together, these studies established silicon sampling as a plausible research method while also making clear that its success depends on the match between prompt design, task structure, and the target human behavior.

A recurring conclusion from later silicon-sampling work is that context-rich prompting matters. Lee et al. (Lee et al., 2024) found that GPT models conditioned only on demographics matched survey data on global warming poorly, whereas adding issue-specific covariates substantially improved fidelity. Ma (Ma, 2025) reported a similar pattern for facial recognition technology acceptance, with demographic-plus-context prompts aligning more closely with human responses than demographic-only prompts. Recent work on personalized survey modeling likewise suggests that synthetic sampling may be more promising for aggregate trends than for capturing the full variability and complexity of individual human opinions (Kaiser et al., 2025). These findings are especially relevant to the present study because the Charité personas include attitudinal and experiential covariates that should be more informative than demographics alone for a novel healthcare technology decision.

The use of LLMs in healthcare has expanded rapidly beyond traditional text processing tasks. Khaokaew et al. (Khaokaew et al., 2025) created demographic-based digital twins from survey respondents to model healthcare decisions and found that some LLMs produced unrealistic behavior, such as near-universal vaccine acceptance, while others captured subgroup differences more effectively but introduced new biases. Mehandru et al. (Mehandru et al., 2024) argued that LLMs in medicine should be evaluated as agents embedded in simulated clinical environments, including patient-physician interactions and hospital processes, rather than only by narrow benchmark scores. Yu et al. (Yu et al., 2025) further showed that multi-agent simulated patient systems can generate accurate, readable, and reliable medical encounters for training purposes. These studies demonstrate the feasibility of LLM-based clinical simulation, but they remain methodologically distinct from our study exploring whether LLMs can reproduce how a specific person, given their psychological and experiential profiles, would respond to a hypothetical quantum AI technology in a clinical scenario.

Our study sits at the intersection of LLM-based human simulation and the study of technology acceptance in healthcare. Longoni et al. (Longoni et al., 2019) showed that consumers often resist medical AI technology because they doubt its ability to account for their unique circumstances, but this resistance weakens when AI is framed as personalized or as supporting rather than replacing a human clinician. Robertson et al. (Robertson et al., 2023) found that patients' attitudes

toward AI diagnoses are mostly shaped by physician endorsement and demographic characteristics. Complementing these findings, Hassan et al. (Hassan et al., 2024) and Witkowski et al. (Witkowski et al., 2024) showed that AI-enabled decision aids can support shared decision-making and patient-centered care, but only if concerns about understandability, therapeutic alliance, transparency, patient choice, and informed consent are addressed. Overall, this underscores the complexity of the decision the silicon samples are being asked to simulate: technology acceptance in a clinical context is shaped by a web of psychological, demographic, and situational factors.

A strand of work regarding replicability is a relevant concern for studies using silicon sampling. Barrie et al. (Barrie et al., 2025) showed that LLMs can be accurate yet have unacceptably high variance, and that in many cases, LLM findings cannot be re-run, let alone replicated. Schröder et al. (Schröder et al., 2025) reinforced this concern by demonstrating that slight wording changes can produce notable divergences from human responses. Practical advice from Barrie et al. (Barrie et al., 2025) and Schröder et al. (Schröder et al., 2025) to consider open models, justify usage of closed models, run repeated routines over multiple days, weeks, or months, and keep a record of every prompt text, model version, and parameter setting directly informed our study design.

A final dimension of our study concerns the effect of prompting language on LLM fidelity. The original Charité interviews were conducted in German, and we prompt each LLM in both English and German to assess whether language affects simulation accuracy. V. Gupta et al. (Gupta et al., 2025b) tested multiple LLMs across nine languages, including German, on educational tasks and found that performance using English prompts continues to be better than in other languages, though the difference was not always large. Similarly, Vadlapati (Vadlapati, 2024) tested GPT-3.5 across English, German, Spanish, French, and Hindi and found that performance was clearly better when prompted in English. While these studies focused on task performance rather than silicon sampling fidelity, they raise the possibility that LLMs may simulate human respondents less faithfully when prompted in non-English languages, which is an important consideration given that the attitudes and decisions being simulated were originally expressed in German.

In summary, prior work establishes silicon sampling as a reasonable but fragile method. LLMs can sometimes recover aggregate human response patterns, especially when prompts include relevant covariates, but evidence remains mixed for individual-level prediction, high-stakes healthcare contexts, multilingual prompting, temporal replicability, and genuinely novel technologies. Existing studies have typically addressed these issues separately. Studies using silicon sampling on public opinions often focus on domains with established social scripts; healthcare LLM studies often evaluate clinical reasoning or simulated encounters rather than validated person-level treatment choices; multilingual evaluations usually examine task performance rather than silicon-sampling fidelity; and replicability work has rarely been embedded in clinical persona simulations. The present study brings these gaps together by treating quantum AI in oncology as a limit case for persona-conditioned silicon sampling and by embedding that test in an English–German prompt comparison and a temporal rerun design.

3. Theoretical framework and expectations

To motivate the theoretical expectations, we briefly summarize the scenario structure in our study; full procedural details are reported in the Methods section. Participants chose between two oncology treatments after receiving recommendations from three sources: an attending physician, a tumor board, and a purported quantum AI decision-support system. The design randomized the physician's recommended treatment between the two treatments and varied source alignment across three configurations: physician-isolated, where the tumor board and quantum AI opposed the physician; quantum-AI-isolated, where the physician and tumor board opposed the quantum AI; and tumor-board-isolated, where the physician and quantum AI opposed the tumor board.

This scenario is also a boundary-condition test for persona-sensitive silicon sampling. In many silicon sampling applications, LLMs may benefit from dense public discourse and repeated examples of human judgments about the target issue, such as elections, climate change, vaccines, or familiar consumer technologies. Here, however, the target is a patient-like decision in which a purported quantum AI oncology system conflicts or aligns with human medical authorities. Comparable real-world patient experience is likely sparse. The task therefore shifts the burden of simulation onto persona-sensitivity where the model must infer how a person with a particular trust profile, technology affinity, locus of control, personality profile, and healthcare history would assess the physician, tumor board, and quantum AI source. We use persona-sensitivity to refer to the model's ability to make outputs vary appropriately with such individual-specific variables instead of defaulting to generic scenario cues.

3.1. Dual-process theories

To evaluate whether persona-conditioned LLMs can reproduce individual treatment choices, we first specify what would make a decision simulable from structured persona information. We define a choice as simulable to the extent that the persona prompt captures the information participants rely on when making the decision. A decision should be more simulable when it is guided by relatively stable dispositions represented in the prompt, such as trust in physicians, trust in AI, prior healthcare experience, cancer experience, locus of control, or affinity for technology interaction. Conversely, a decision should be less simulable when it hinges on factors not captured by the prompt, such as impressions from the role-play encounter, perceived treatment burden, or idiosyncratic interpretations of the medical information.

This premise is grounded in dual-process theories of judgment and decision-making, which distinguish between fast, intuitive, heuristic judgment and slower, more deliberative reasoning (Kahneman, 2011; St et al., 2013). Dual-process theory helps specify which part of the task a structured persona prompt should be expected to capture. In the present source-conflict scenario, categorical treatment choice can plausibly be generated through relatively fast source-weighting heuristics, such as heeding the attending physician, deferring to human clinical authority over an unfamiliar AI system, or accepting the AI recommendation when it is clinically endorsed. These source-based heuristics can be approximated from structured persona variables, including trust in physicians, trust in AI, and affinity for technology interaction. Confidence, by contrast, is a metacognitive judgment about how settled the participant feels after considering the recommendation conflict, treatment burdens, emotional reactions to the cancer scenario, and residual uncertainty. These deliberative and affective inputs draw on richer, more idiosyncratic information that are only indirectly represented in the current structured questionnaire prompt. Thus, the dual-process account yields the expectation that, under the present persona-prompt design, LLMs may reproduce some categorical choices through source-based heuristics while struggling to preserve individual-level differences in confidence.

3.2. Technology acceptance theory

Technology acceptance theory provides a second basis for interpreting the scenario. In the Technology Acceptance Model (TAM), acceptance of a technology is shaped by perceived usefulness and perceived ease of use (Davis, 1989). The Unified Theory of Acceptance and Use of Technology (UTAUT) extends this logic by emphasizing performance expectancy, effort expectancy, social influence, and facilitating conditions (Venkatesh et al., 2003). However, the present scenario differs from conventional TAM and UTAUT applications because participants are not deciding whether to personally use a quantum AI

system, nor can they evaluate its ease of use, performance, or usefulness through direct interaction. Instead, they are deciding whether to accept a treatment recommendation generated by an unfamiliar and technically opaque system. Thus, TAM and UTAUT are used here as frameworks for understanding recommendation acceptance rather than as full adoption models.

In standard TAM applications, perceived usefulness can be formed or updated through use experience or concrete interaction with the technology. However, in the present vignette, the perceived usefulness of the quantum AI system must be inferred almost entirely from the clinical framing provided by the physician and from the system's relation to other medical authorities. The physician introduces the quantum AI as an evidence-based decision-support tool, but the design then varies whether the AI's specific recommendation is endorsed or contradicted by the attending physician and the tumor board. Acceptance of the AI-recommended treatment therefore depends less on an independently formed attitude toward quantum AI than on whether participants trust the endorsement structure surrounding its recommendation. This connects the scenario directly to the social influence component of UTAUT, where the physician and tumor-board agreement function as normative and epistemic cues that substitute for direct use experience. It is also compatible with the resistance model from Longoni et al. (Longoni et al., 2019), because resistance to medical AI should be reduced when the system is framed as personalized and clinician-mediated instead of replacing human judgment. Under these conditions, trust in AI and affinity for technology interaction may still matter, but mainly insofar as they shape whether participants accept the framing by physicians and clinically endorsed AI recommendation as useful and trustworthy.

3.3. Shared decision-making theory

Finally, the scenario is also a shared decision-making problem. Shared decision-making models emphasize that treatment decisions are preference-sensitive choices in which clinicians and patients jointly consider evidence, risks, burdens, and patient values (Elwyn et al., 2017). The current vignette complicates this process because the authoritative sources do not provide a single coherent recommendation. As a result, the decision may feel less like a choice between treatments than a choice about how much authority to grant the physician, the tumor board, and the quantum AI system. This distinction matters for confidence because a binary treatment choice may follow a dominant cue, such as physician trust or openness to AI, whereas confidence may reflect unresolved conflict, perceived understanding, and emotional burden. A participant may choose the physician-recommended option while still feeling uncertain because another authority disagrees, or accept the quantum AI recommendation while remaining uneasy about relying on an unfamiliar technology. For this reason, confidence should be harder to simulate than categorical treatment choice.

In combination, these frameworks generate bounded theoretical expectations for silicon sampling in this study. First, persona-conditioned LLMs may recover broad scenario-level structure if their training data encode general social schemas about doctors, AI, expert consensus, and medical uncertainty. Second, individual-level alignment should be strongest when the structured persona variables point clearly toward a dominant source preference. Third, acceptance of the treatment recommended by the quantum AI should primarily follow an endorsement gradient. Consequently, quantum AI-recommended treatments should be most acceptable when supported by the attending physician because physician endorsement serves as a social influence cue that reduces perceived risk of an unfamiliar system, and it is least acceptable when the quantum AI system stands alone against both human medical sources. Fourth, confidence should be harder to simulate than categorical treatment choice because confidence depends on source-weighting heuristics as well as on how participants internally resolve the tension between opaque AI recommendation, physician endorsement, and expert disagreement. A participant may choose the quantum AI-recommended treatment because it is clinician-endorsed while still feeling uncertain about relying on quantum AI, or may follow the physician while remaining unsettled by AI or tumor-board disagreement. These individual interpretations are shaped by deliberative and affective processing that the questionnaire-derived persona variables may not fully capture.

Because the present study is exploratory, these expectations are not tested as formal hypotheses. However, together with the empirical motivation and prior work reviewed in the previous sections, they provide a theoretical baseline for interpreting the alignment results and lead to the following research questions:

RQ1: What are the descriptive magnitude and uncertainty of alignment between persona-conditioned LLM outputs and observed human treatment decisions and confidence ratings in a clinical source-conflict scenario involving recommendations from a physician, a tumor board, and a purported quantum AI system?
RQ2: Do English and German prompts yield different LLM-human alignment estimates for simulated treatment decisions and confidence ratings?
RQ3: How temporally consistent are LLM-generated treatment decisions and confidence ratings across six repeated sampling dates over approximately five weeks?

4. Methods

4.1. Procedure at Charité

At the start of the silicon sampling phase, data were available for 25 participants from a realistic role-play simulation conducted at Charité Universitätsmedizin Berlin. Two participants had incomplete survey data, leaving a final sample of 23 participants with complete and valid data. All participants were healthy volunteers who were screened for contraindications based on self-reported cardiac conditions and relevant medical history before participation.

The following describes the standardized procedure. Participants first received an information sheet providing a brief introduction to quantum computing and its potential applications, followed by a case vignette describing the simulation context and their assigned patient role (both materials are in Appendix A). An attending physician then escorted the patients to the simulation room and presented a device described as a quantum computer with AI assistance, framed as an evidence-based decision-support tool for oncological treatment planning. The purported quantum AI system was part of the role-play framing and no quantum computation was performed during the role-play. Participants were informed that histopathology had confirmed a diagnosis of Stage IIIA non-small cell lung cancer with lymph node metastasis. The physician explained that the disease was advanced and potentially life-threatening, situated at the boundary between curative and palliative care, and therefore required an urgent treatment decision. Two treatment strategies were then presented: Treatment A consisted of neoadjuvant chemotherapy with radiotherapy, whereas Treatment B consisted of neoadjuvant chemotherapy followed by surgical tumor resection. The physician also explained the potential benefits and side effects of both options, after which participants were given time to reflect on the information and ask questions. Subsequently, the attending physician, the tumor board, and the purported quantum AI system each recommended one of the two treatment options according to the assigned scenario condition.

Participants were randomly assigned to one of three scenario groups: in group 1, both the tumor board and the quantum AI system recommended the alternative to the physician's recommendation; in group 2, the tumor board agreed with the physician whereas the quantum AI system recommended the alternative; and in group 3, the quantum AI system agreed with the physician whereas the tumor board recommended the alternative (see Fig. 1). To avoid confounding source effects

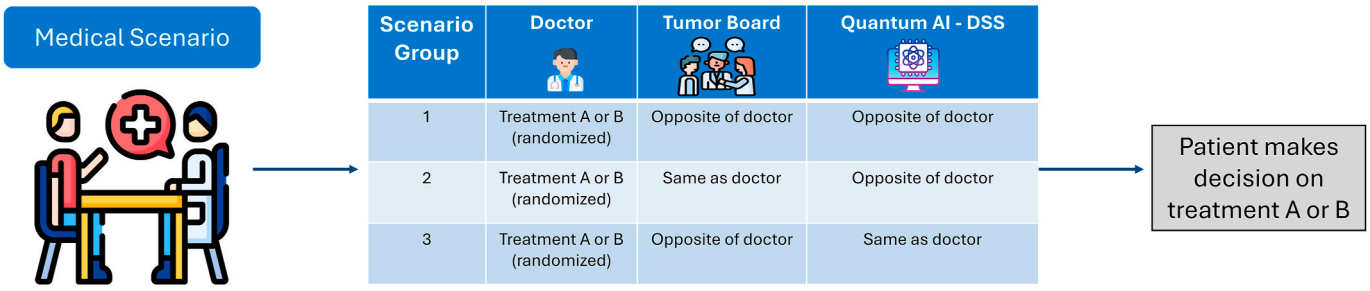


Fig. 1. Clinical role-play and recommendation constellations across the three scenario groups at Charité. The attending physician's recommendation was randomized between Treatment A and Treatment B. In group 1, both the tumor board and the purported quantum AI decision-support system recommended the alternative treatment to the physician's recommendation. In group 2, the tumor board agreed with the physician, whereas the quantum AI system recommended the alternative treatment. In group 3, the quantum AI system agreed with the physician, whereas the tumor board recommended the alternative treatment. Participants then chose between Treatment A and Treatment B. The figure was created using Microsoft PowerPoint; all icons are from Freepik at flaticon.com.

with a specific treatment option, the physician's recommendation was randomly chosen between Treatment A and Treatment B, and the other recommendations were determined relative to that recommendation according to scenario assignment. A team member at Charité used a computerized random number generator to assign participants to scenario groups and to select the physician's recommendation between Treatments A and B. Scenario assignment and physician recommendation were generated independently, without forced balancing. The final analyzed scenario groups contained six, nine, and eight participants. Participants then selected their preferred treatment option and rated their confidence in that decision on an integer scale from 0 (not confident at all) to 100 (completely confident). Finally, they completed questionnaires assessing trust in AI and in physicians, prior experiences with physicians and cancer, locus of control (Internale-Externale-Kontrollüberzeugung-4; IE-4), personality (BFI-10), demographics (age, gender, education, and employment status), and affinity for technology interaction (ATI).

4.2. Silicon sampling procedure

Silicon samples were generated using four LLMs: GPT-5.1 (OpenAI API), Claude 4.5 Opus (Anthropic API), DeepSeek V3.2, and Kimi K2 Thinking (both via the OpenRouter API). For models accessed through OpenRouter, the Novita, Groq, and SambaNova providers were excluded because prior reports indicated poorer performance in terms of accuracy relative to other available providers (Brand & Denain, 2025; Clark, 2025; Ermut & Şipi, 2026; MoonshotAI, 2025). GPT-5.1 was configured with the verbosity parameter set to low. Claude 4.5 Opus was given a thinking budget of 4600 tokens. For DeepSeek V3.2, Kimi K2 Thinking, and GPT-5.1, reasoning/thinking was set to high. The remaining parameters for every model were kept at their default values (the exact settings and values of the parameters are listed in Appendix B). Deterministic generation was not enforced and no fixed random seed was specified because Claude 4.5 Opus did not support a deterministic seed parameter at the time of this procedure. We included the closed-source models, GPT-5.1 and Claude 4.5 Opus, because they represented the frontier in complex reasoning at the time of this procedure. Furthermore, their inclusion allowed us to do a comparative analysis with the high-performing open-source models, DeepSeek V3.2 and Kimi K2 Thinking.

Each model received a system prompt instructing it to adopt the persona of a specific participant. The persona embedded in the system prompt was constructed from the participant's questionnaire data, including demographics (age, gender, education, and employment status), trust in AI, trust in physicians, prior experiences with physicians and with cancer, IE-4, BFI-10, and ATI. The user prompt presented the quantum computing information sheet and the clinical decision scenario, then instructed the model to select a treatment option and rate its confidence from 0 to 100. Responses were required to contain a treatment choice, A or B, followed by an integer confidence rating from 0 to 100. Responses failing this validation were discarded, and the same request was resubmitted until a valid response was obtained. The original materials were all in German, so to create the English version of the prompts, we referred to the official English versions for the ATI and BFI-10 surveys. For the rest, we used DeepL to translate to English, then we checked the translations with two native German speakers. The system and user prompts in both English and German are provided in Appendix A. The silicon sampling procedure was done once in English and again in German, on six different dates: 2025-12-06, 2025-12-13, 2025-12-19, 2025-12-26, 2026-01-02, 2026-01-09.

4.3. Alignment metrics and interpretive scope

Because the study was designed as exploratory with a sample of 23 participants, the number of human observations per scenario group was necessarily small. As a result, all participant-level metrics reported below should be interpreted as descriptive effect-size estimates rather than as confirmatory tests of individual-level LLM simulation fidelity, subgroup differences, or language effects. Accordingly, we also report bootstrapped confidence intervals to convey estimation uncertainty and interpret both participant-level estimates and scenario-group patterns descriptively.

At the scenario-group level, we first calculated descriptive rates of agreement with the attending physician's recommendation and descriptive mean confidence ratings. For the silicon data, these summaries were computed within each scenario group, model, and language condition, pooling the six sampling runs from each participant persona. The corresponding human summaries were based on one observed decision and one confidence rating per participant. We then compared these scenario-level silicon summaries with the corresponding human summaries.

Treatment choice and decision confidence were analyzed as distinct outcomes. To evaluate participant-level decision alignment, the six LLM sampling runs were aggregated to obtain a modal treatment choice for each participant persona, model, and language. The modal choice was the most frequently selected treatment; ties were resolved using the higher mean confidence for the tied treatments. Choice-frequency ties occurred in six of 184 participant-model-language combinations. Five cases were resolved by selecting the treatment with higher mean confidence; one case had both tied choice frequency and mean confidence, so treatment A was selected as a deterministic alphabetical tiebreaker. Decision alignment was evaluated using accuracy and Cohen's κ.

We assessed sensitivity to our confidence-based tie rule by assigning all frequency-tied profiles to Treatment A or Treatment B without using confidence. We also recalculated decision alignment after excluding frequency-tied profiles. In this sensitivity analysis, accuracy and Cohen's κ were evaluated by model and language, with participant-level bootstrap confidence intervals and evaluation denominators reported (full

metrics are in Appendix C).

We additionally calculated accuracy and Cohen's κ separately for each model, language, and date, to compare the 23 LLM choices with the corresponding human choices and to characterize decision alignment without aggregation across sampling dates. Participant-level bootstrap confidence intervals were calculated for these estimates, and the same human responses served as the reference on all dates (full metrics are in Appendix D).

Confidence was elicited on a scale from 0 ("not confident at all") to 100 ("completely confident"). For the main confidence analysis, we calculated LLM confidence as the mean raw confidence across all six sampling runs for each participant profile, model, and language, irrespective of the treatment selected. These means were compared with the corresponding human confidence ratings across all 23 participants using mean absolute error (MAE) and the concordance correlation coefficient (CCC). This analysis assesses agreement only in reported certainty about respondents' own treatment choices, and not agreement in treatment preference. As a secondary analysis, we evaluated confidence alignment among decision-concordant profiles, defined as profiles for which the LLM modal treatment choice matched the human choice. For these profiles, LLM confidence was calculated as the mean confidence across runs producing the modal choice, then MAE and CCC were calculated against human confidence. The number of decision-concordant profiles for each model-language condition is reported in Appendix E.

We estimated 95% confidence intervals using a clustered bootstrap with 10,000 resamples at the participant level. For metrics conditional on decision agreement, we verified post hoc that all bootstrap iterations produced valid conditional estimates.

4.4. Paired language comparisons

To evaluate RQ2 directly, we calculated within-model German minus English differences in decision accuracy, Cohen's κ, and all-participant confidence MAE and CCC. We estimated 95% confidence intervals using 10,000 paired participant-level bootstrap resamples. In each iteration, the same resampled participant IDs were used for both languages, so that the corresponding human observation and LLM outputs were retained, and the language difference was calculated within the resample. These comparisons were interpreted as exploratory and hypothesis-generating because no adjustments for multiple comparisons were applied, and none of the between-model or between-language contrasts was powered for hypothesis testing.

4.5. Temporal output consistency metrics

For each model, language, and participant profile, we computed within-profile consistency metrics across the six dates to assess output consistency. Decision stability was implemented by identifying the modal decision, the most frequently occurring treatment choice, across the six dates for that participant. In case of ties, we took the decision with the highest average confidence rating across runs. Then, we calculated the proportion of instances where the model matched this mode to the six sampling instances. Confidence volatility was defined as the standard deviation of the confidence ratings across runs. These participant-level metrics were then averaged across all 23 participant profiles within each model-language condition, yielding summary estimates of average decision stability, average confidence, and average confidence volatility. Decision stability was reported as a percentage, with higher values indicating greater consistency of categorical choices; lower confidence volatility indicated greater stability of confidence judgments across sampling dates.

5. Results

5.1. Decision alignment

Fig. 2 presents the decision agreement rates with the attending physician's recommendation across the three scenario groups for human participants and for the four silicon models under German and English prompting. Given the small human sample size (N = 23) and the correspondingly small number of human participants within each scenario group, the agreement rates shown in Fig. 2 should be interpreted as descriptive and exploratory. Human participants agreed with the physician's recommendation in 67% of cases in group 1, 44% in group 2, and 75% in group 3. Across model-language conditions, the silicon samples tended to show lower agreement with the physician in group 1 (0% to 39%) and higher agreement in groups 2 (48% to 67%) and 3 (88% to 98%) than the human participants. In all model-language conditions, physician agreement increased from group 1 to group 3, whereas the human data showed lower agreement in group 2 than in group 1. With respect to prompt language, English and German prompting yielded broadly similar scenario-specific patterns within models, and no consistent directional advantage for either language was apparent. The largest within-model language discrepancy was observed for GPT-5.1 in scenario group 1 (39% in German vs. 19% in English), whereas differences for the other models were generally small.

Table 1 reports participant-level accuracy and Cohen's κ for each model-language combination, based on the silicon modal decision obtained by aggregating the six sampling runs for each participant persona. Modal decision accuracy ranged from 0.57 to 0.65. Chance-corrected agreement was generally weak, with κ values ranging from 0.11 to 0.29. All 95% cluster-bootstrap confidence intervals for κ crossed zero and were wide, indicating that the present sample does not statistically resolve whether any model achieved agreement beyond chance at the participant level. Kimi K2 Thinking had the numerically highest κ point estimate in both English and German, κ = 0.29, but its confidence intervals also crossed zero and overlapped substantially with those of the other models. Prompt-language differences were small and inconsistent; the most visible descriptive difference was observed for GPT-5.1, whose accuracy decreased from 0.65 in English to 0.61 in German and whose κ decreased from 0.24 to 0.11, although the wide overlapping intervals preclude any strong conclusion about a language effect.

5.2. Confidence alignment

Fig. 3 shows the mean confidence ratings across the three scenario groups for human participants and for the four silicon models under German and English prompting. Given the limited human sample size (N = 23), these comparison results should be interpreted as exploratory. Human participants reported mean confidence ratings of 69.50 in group 1, which increased to a peak of 81.44 in group 2, and subsequently decreased to 75.13 in group 3. Most model-language conditions showed lower mean confidence than the human participants; however, DeepSeek V3.2 exceeded the human mean in scenario groups 1 and 3 under both prompt languages. Generally, confidence increased from group 1 to group 2 and then remained similar or decreased slightly in group 3, broadly resembling the human pattern, although the silicon data mostly showed less between-group variation. Regarding prompt language, English and German prompting yielded broadly similar scenario-specific patterns within models, and no consistent directional advantage for either language was apparent. The largest within-model language discrepancy was observed for DeepSeek V3.2 in scenario group 2 (80.15 in German vs. 74.61 in English), whereas differences for the other models were modest.

Fig. 4 presents participant-level confidence alignment on the original confidence scale. The main analysis included all 23 participants in each model-language condition, comparing each human rating with the corresponding LLM's mean confidence across all six sampling runs,

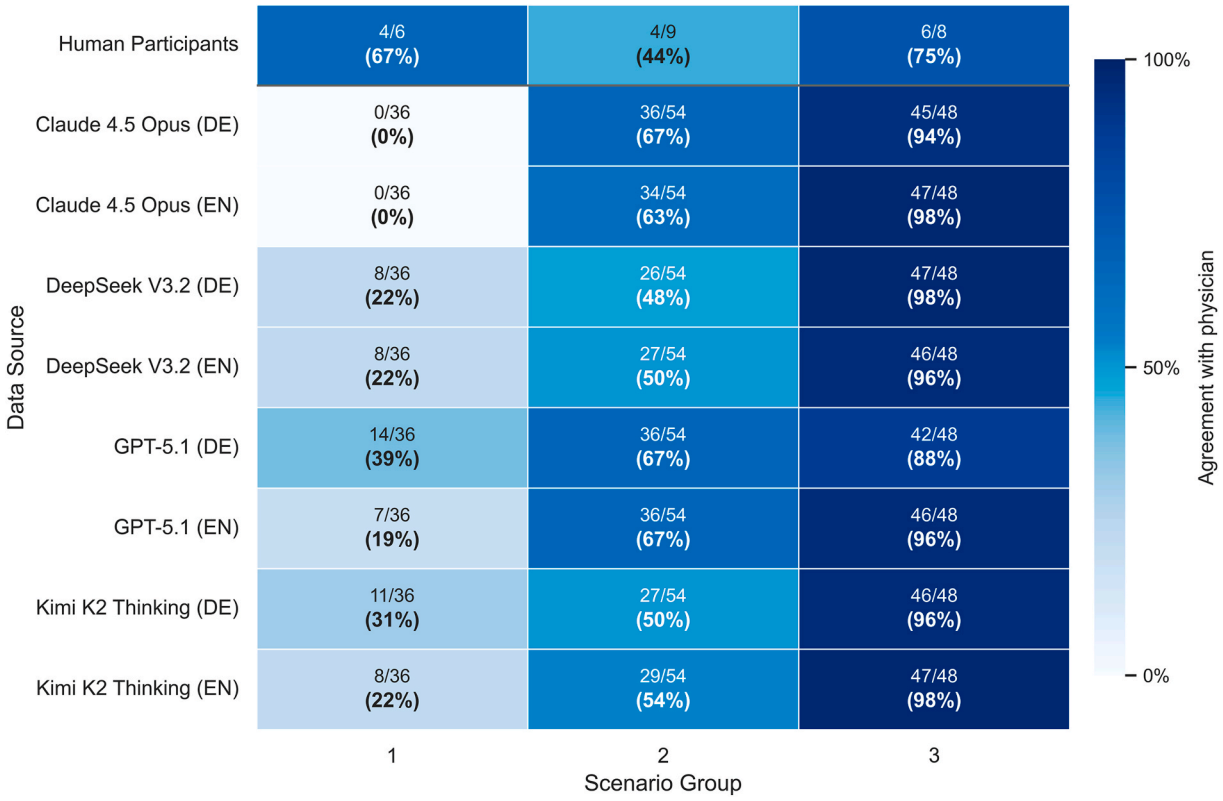


Fig. 2. Heatmap of agreement rate with the physician's recommendation by scenario group and data source. Cell annotations show the number of agreeing responses divided by the total number of responses, with percentages rounded to whole numbers in parentheses. The color shade describes the unrounded agreement proportion. Human denominators represent individual participants, and LLM denominators represent outputs pooled across six sampling runs per participant. DE and EN denote German and English prompting, respectively.

Table 1
Accuracy and Cohen's Kappa metrics for the modal decision of the four models grouped by English and German prompting. Values in brackets are 95% cluster-bootstrap confidence intervals based on resampling participants with 10,000 iterations.

Language	Model	Accuracy	Cohen's Kappa (κ)
English	Claude 4.5 Opus	0.57 [0.35, 0.74]	0.12 [-0.27, 0.48]
	DeepSeek V3.2	0.61 [0.39, 0.78]	0.23 [-0.14, 0.58]
	GPT-5.1	0.65 [0.44, 0.83]	0.24 [-0.18, 0.64]
	Kimi K2 Thinking	0.65 [0.48, 0.83]	0.29 [-0.10, 0.65]
German	Claude 4.5 Opus	0.57 [0.35, 0.74]	0.12 [-0.27, 0.48]
	DeepSeek V3.2	0.61 [0.39, 0.83]	0.23 [-0.14, 0.59]
	GPT-5.1	0.61 [0.39, 0.83]	0.11 [-0.28, 0.52]
	Kimi K2 Thinking	0.65 [0.44, 0.83]	0.29 [-0.10, 0.65]

irrespective of treatment choice. Confidence MAE ranged from 11.00 points for DeepSeek V3.2 in English to 18.06 points for Claude 4.5 Opus in German (Panel A). DeepSeek V3.2 had the lowest MAE point estimate in both languages, whereas Claude 4.5 Opus had the highest. Confidence CCC estimates ranged from −0.16 to 0.06, but all 95% confidence intervals included zero (Panel B), providing no clear evidence of positive individual-level confidence concordance. English prompting tended to yield lower all-participant confidence MAE for all four models, with differences ranging from 0.17 to 1.77 points between models. CCC differences did not show a consistent language advantage. These comparisons are descriptive and do not establish model superiority or a language effect.

For the secondary analysis among decision-concordant profiles, confidence MAE ranged from 10.66 to 16.25 points (Panel C), while CCC ranged from −0.13 to 0.19, with all 95% confidence intervals including zero (Panel D). Because the eligible participant subsets and the confidence-aggregation rule differ from the main analysis, these conditional estimates should not be interpreted as directly comparable measures of overall model performance. Overall, the analyses do not establish reliable reproduction of individual participants' reported confidence. Complete numerical results for all concordant-decision alignment metrics, including the number of concordant decisions per condition, are provided in Appendix E.

5.3. Paired language differences

Table 2 reports paired German-minus-English differences in participant-level alignment for modal decision and all-participant confidence. Decision accuracy and Cohen's κ point estimates were identical between languages for Claude 4.5 Opus, DeepSeek V3.2, and Kimi K2 Thinking. For GPT-5.1, the German-minus-English differences were −0.04 for accuracy and −0.13 for κ, with respective 95% confidence intervals of [−0.13,0.00] and [−0.40,0.00]. All-participant confidence MAE was higher under German prompting for every model, with differences ranging from 0.17 to 1.77 points. Confidence CCC differences ranged from approximately −0.07 to 0.09 and showed no consistent directional advantage. Notably, every paired-difference interval included zero.

5.4. Model consistency across sampling dates

All four models demonstrated high decision stability across the six sampling dates (Table 3). Claude 4.5 Opus had the highest average decision stability, producing the same modal decision on 97.83% of runs in both English and German. GPT-5.1 was the next most consistent (English: 94.93%; German: 94.20%), followed by Kimi K2 Thinking (English: 91.30%; German: 89.86%) and DeepSeek V3.2 (English: 89.13%;

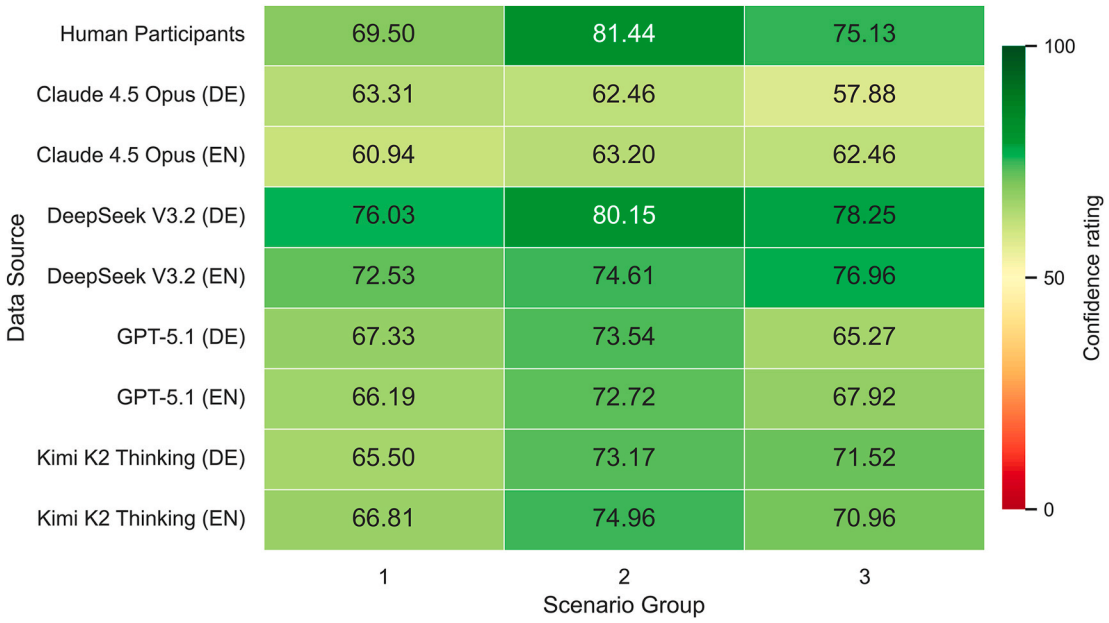


Fig. 3. Heatmap of the average confidence rating grouped by scenario and data source (human participants N = 23). (DE) means German and (EN) means English; these are the languages that the models were prompted with during silicon sampling. For the sources from models, average confidence ratings were calculated using all six sampling runs per participant profile.

German: 90.58%). Language differences in decision stability were small: Claude 4.5 Opus was identical across languages, GPT-5.1 and Kimi K2 Thinking were slightly more stable in English, while DeepSeek V3.2 was slightly more stable in German. Confidence volatility, measured as the average within-profile standard deviation of confidence ratings across runs, followed a pattern roughly consistent with the decision stability ranking. Claude 4.5 Opus exhibited the lowest confidence volatility (English: 3.28; German: 3.11), while DeepSeek V3.2 showed the highest volatility in English (6.21) and GPT-5.1 the highest in German (5.68). Kimi K2 Thinking fell in between (English: 4.44; German: 4.54). Across all model–language conditions, average confidence standard deviations remained below 6.21 points on the 0–100 scale, suggesting that confidence ratings were reasonably stable across sampling dates even where some run-to-run variability was present.

6. Discussion

Using 23 human participant profiles from a realistic oncology role-play at Charité Universitätsmedizin Berlin, we estimated the magnitude and uncertainty of LLM-human alignment in treatment choice and decision confidence, compared English and German prompting descriptively, and assessed temporal consistency across six sampling dates. The central finding is that the models showed aggregate regularities in how conflicting recommendations shaped scenario-group-level choices, but the available human sample was too small to statistically resolve individual-level fidelity. Accuracy point estimates were modest, Cohen's κ estimates were weak and imprecise, and confidence concordance was near zero. Against this overall picture, we revisit the four expectations from the theoretical framework. The first expectation that persona-conditioned LLMs may recover broad scenario-level structure was partially supported. The silicon samples did not produce flat choices across scenarios, and model physician-agreement rates generally increased from the physician-isolated configuration to the physician–AI configuration. However, this source-structure pattern did not fully reproduce the human data, especially the unexpectedly low human physician agreement in group 2. The second expectation that individual-level alignment should be strongest when persona variables indicate a dominant source preference was not supported by the present data. Despite prompts containing trust-in-physicians, trust-in-AI, and technology affinity scores, modal decision accuracy remained modest (0.57–0.65) and chance-corrected agreement was weak ($\kappa = 0.11$–0.29) with all confidence intervals crossing zero. The third expectation that quantum AI treatment acceptance should be most acceptable when physician-endorsed and least when isolated was mixed. The silicon samples often treated the AI recommendation as more acceptable when it was paired with the physician in group 3 and less acceptable when the AI stood alone in group 2, but several model conditions also strongly followed the AI–tumor-board consensus in group 1. Furthermore, the human data deviated, with the lowest physician agreement occurring in group 2 rather than group 1, indicating that a simple endorsement gradient rule may not fully account for participant choice. The fourth expectation was that confidence would be harder to simulate than categorical choice, but the findings did not establish this expectation. For instance, confidence concordance was weak in both the all-participant analysis (CCC: −0.16–0.06) and the decision-concordant analysis (CCC: −0.13–0.19), while chance-corrected choice agreement was also weak and imprecise. In addition, the choice and confidence metrics are not directly comparable, so the present analyses indicate limitations in both outcomes as opposed to confidence being harder to reproduce than choice. The study therefore provides useful pilot estimates and signals on aggregate structural patterns, but it does not establish that persona-conditioned LLMs can faithfully reproduce the decisions or confidence levels of specific individuals in this high-stakes clinical scenario.

6.1. Decision and confidence alignment

The first research question asked whether persona-conditioned LLMs reproduced participants' treatment decisions and confidence. As previously established, the models showed broad scenario-level structure but did not establish adequate individual-level fidelity. The models may have applied general schemas about medical authority and expert consensus drawn from the explicit source cues in the prompt such as trust in physicians, trust in AI, acceptance of doctor-reviewed AI, while missing unobserved factors such as fear, perceived treatment burden, or impressions from the role-play encounter. The descriptive human pattern in group 2, where physician agreement was lowest despite physician and tumor board consensus, also cautions against treating a simple authority-endorsement rule as an adequate account of

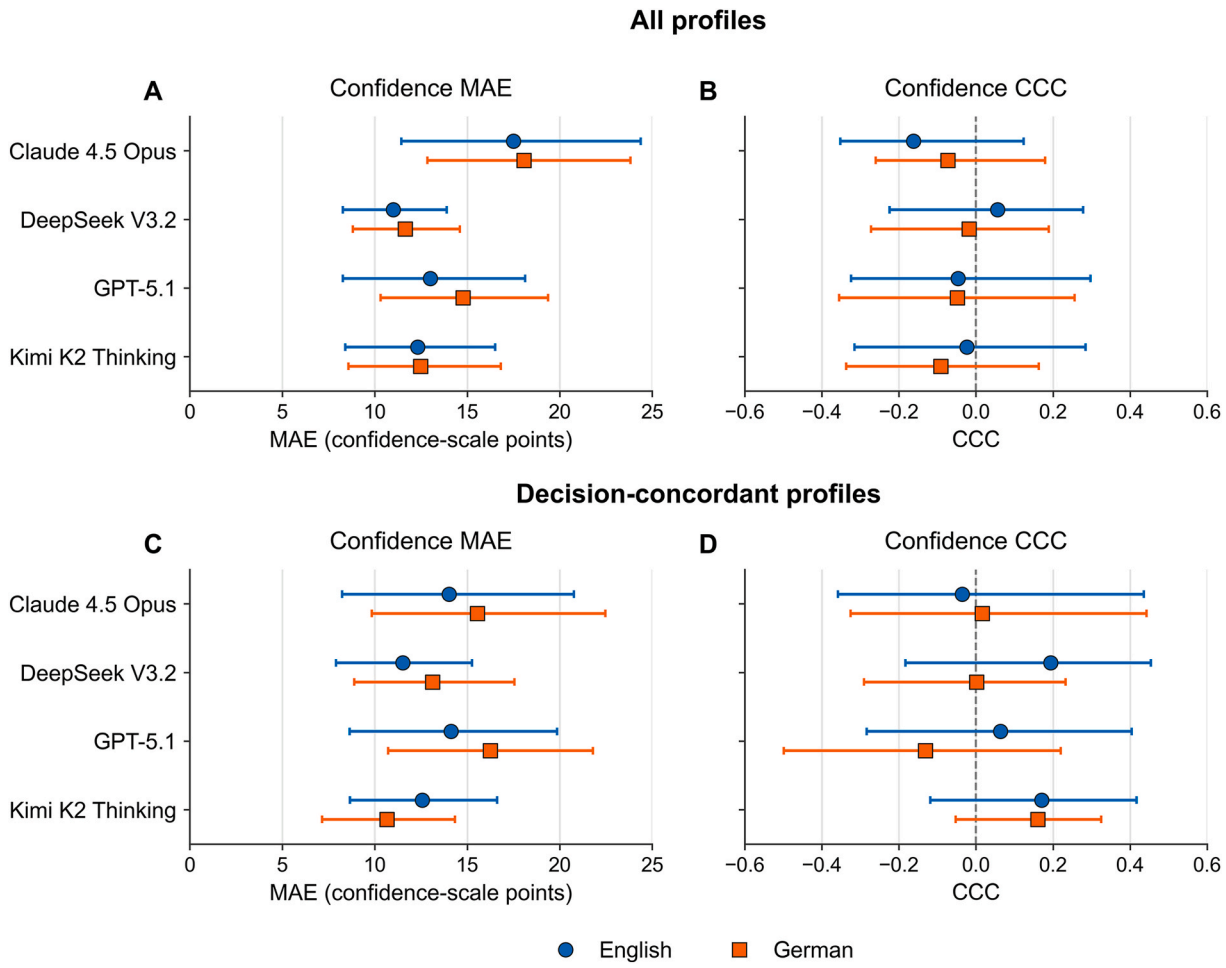


Fig. 4. Participant-level confidence alignment by model and prompt language. Symbols show point estimates, and horizontal lines show 95% participant-level percentile bootstrap confidence intervals based on 10,000 resamples. Blue circles and orange squares denote English and German prompts, respectively. Panels A and B compare the human confidence rating with the corresponding LLM's mean raw confidence across all six sampling runs, irrespective of treatment choice, for all 23 participants. Panels C and D present the analysis among profiles for which the LLM modal treatment choice matched the human choice; LLM confidence was averaged only across runs producing that modal choice. Conditional sample sizes ranged from 13 to 15 profiles and are reported by condition in Appendix E. Panels A and C show mean absolute error (MAE) in original confidence-scale points; Panels B and D show the concordance correlation coefficient (CCC) and the dashed vertical lines mark CCC equal to zero.

participant choice. Given the small number of participants per scenario group, however, this discrepancy should not be overinterpreted.

Near-zero confidence concordance in both analyses indicates that the present study does not establish individual-level confidence fidelity. This finding is compatible with the proposed explanation that decision confidence depends partly on deliberative and affective information that are not fully represented in the structured questionnaire prompt. The persona variables may be sufficient for the model to approximate which treatment a participant would choose via source-weighting heuristics, but insufficient to calibrate how settled that participant felt about the choice. A participant may follow the physician's recommendation because trust in doctors is the dominant cue, yet still feel deeply uneasy because the quantum AI dissented; another may accept the AI-recommended treatment because it is clinician-endorsed, yet remain uncertain about relying on a system they cannot evaluate. Because the validation sample was small, these findings remain exploratory and do not constitute definitive evidence of a specific failure mechanism.

6.2. *Effect of prompting language (English–German)*

Regarding the second research question, the paired analyses did not establish a consistent language advantage across alignment metrics. Confidence MAE was numerically lower under English prompting for all four models, but the paired confidence intervals included zero. GPT-5.1 also had lower decision accuracy and Cohen's κ under German prompting, but the corresponding intervals similarly extended to zero. Claude 4.5 Opus' identical modal choices across languages indicate stability of its aggregated decisions in the observed sample, but it cannot serve as concrete proof its responses are unaffected by prompt language. Overall, these estimates describe possible language differences whose magnitude and direction remain uncertain in the present sample. The constrained response format and carefully translated materials may have contributed to similar outputs, but the study did not test these explanations. Larger paired validation studies would be needed to determine whether language differences are sufficiently small for a particular research application.

6.3. *Temporal consistency of LLM outputs*

The results for RQ3 point to an important distinction between repeated-run output stability and fidelity to human responses. Across six sampling dates, all four models were highly stable in their treatment choices, and confidence ratings varied comparatively little. This is encouraging given broader concerns that LLM findings can be difficult to rerun and may drift across time (Barrie et al., 2025). Under our present procedure, repeated sampling over roughly five weeks did not generate

Table 2
Paired German-minus-English differences in participant-level alignment by model. Decision metrics use the modal treatment choice. The confidence metrics compare human ratings with the LLM's mean raw confidence across all six runs for all 23 participants. Values in brackets show 95% confidence intervals from 10,000 bootstrap resamples. Accuracy differences are expressed as proportions, mean absolute error (MAE) differences as confidence-scale points, and κ and concordance correlation coefficient (CCC) differences as coefficient differences. Positive differences favor German for accuracy, κ, and CCC, but favor English for MAE.

Model	Metric	German minus English Difference
Claude 4.5 Opus	Accuracy	0 [0, 0]
	Cohen's Kappa	0 [0, 0]
	Confidence MAE	0.56 [-1.54, 2.43]
	Confidence CCC	0.09 [−0.02, 0.19]
DeepSeek V3.2	Accuracy	0 [-0.13, 0.13]
	Cohen's Kappa	0 [-0.16, 0.17]
	Confidence MAE	0.65 [-1.30, 2.52]
	Confidence CCC	−0.07 [-0.28, 0.17]
GPT-5.1	Accuracy	−0.04 [-0.13, 0]
	Cohen's Kappa	−0.13 [-0.40, 0]
	Confidence MAE	1.77 [-0.13, 3.26]
	Confidence CCC	0 [-0.18, 0.14]
Kimi K2 Thinking	Accuracy	0 [-0.13, 0.13]
	Cohen's Kappa	0 [-0.17, 0.18]
	Confidence MAE	0.17 [-1.36, 1.75]
	Confidence CCC	−0.07 [-0.22, 0.08]

Table 3
Average decision stability and confidence volatility across six sampling dates by model and prompt language.

Language	Model	Average Decision Stability (%)	Average Confidence Standard Deviation
English	Claude 4.5 Opus	97.83%	3.28
	DeepSeek V3.2	89.13%	6.21
	GPT-5.1	94.93%	5.21
	Kimi K2 Thinking	91.30%	4.44
German	Claude 4.5 Opus	97.83%	3.11
	DeepSeek V3.2	90.58%	5.19
	GPT-5.1	94.20%	5.68
	Kimi K2 Thinking	89.86%	4.54

large swings in model behavior.

The per-date analysis further distinguishes temporal consistency from alignment with human responses. Accuracy and Cohen's κ varied across sampling dates, and selecting the most frequent choice across six runs did not consistently produce higher estimates than individual runs. There was a positive κ interval for DeepSeek V3.2 in English on only one date, but it does not establish positive chance-corrected agreement across multiple dates. In general, the models' stability did not translate into strong agreement with human participants.

6.4. Practical implications

Our prompts included demographics, trust measures, personality traits, locus of control, prior healthcare experiences, and affinity for technology interaction. This is more context-rich than the demographic-only prompting strategies that have often performed poorly in earlier work (Lee et al., 2024; Ma, 2025). Nevertheless, the present results indicate that even a comparatively rich structured persona may still be insufficient for strong individual-level simulation in a novel, high-stakes healthcare setting. Direct narratives, stated reasons for choice, risk preferences, and the emotional meaning of the cancer scenario may all contain information that structured questionnaire variables do not entirely capture. Therefore, this study both supports and qualifies the existing silicon sampling literature, suggesting that richer covariates likely help, but they do not eliminate the gap between plausible simulation and faithful person-level prediction.

Across scenarios, the LLMs' choices appeared to reflect which sources agreed on a treatment recommendation. Related research shows the importance of how medical AI is positioned in relation to human clinicians, including whether it supports instead of replaces human clinicians (Hassan et al., 2024; Longoni et al., 2019). These observations suggest a possible role for silicon sampling in piloting vignettes with different patterns of agreement between technological and human recommendation sources, but plausible aggregate response patterns do not establish that the models captured participants' underlying reasoning.

Some differences in model performance were observed, but the small validation sample prevents strong conclusions about comparative performance. Kimi K2 Thinking had the highest descriptive decision alignment estimates, while Claude 4.5 Opus had the highest observed average decision stability across the six sampling dates. However, these apparent differences were accompanied by wide and overlapping confidence intervals for the alignment metrics. As a result, these differences are best understood as preliminary signals for hypotheses generation and future research. Nevertheless, if it replicates in larger samples, the competitive descriptive performance of the open-source models would be practically relevant for future silicon sampling research, because such models may offer advantages for transparency, cost, and replicability in exploratory workflows. On other evaluation dimensions, expert-panel LLM comparisons have reported statistically significant differences in informational quality, clinical reasoning, and expert-perceived usefulness (Aliyeva et al., 2026). Accordingly, the weak alignment observed from silicon sampling should not be interpreted as evidence that LLMs lack value as sources of clinical information.

Taken together, the findings suggest a bounded role for silicon sampling in clinical decisions and emerging technology acceptance research. The present study does not validate these models as substitutes for human participants when individual-level fidelity is required, especially in a high-stakes oncology role-play involving a novel technology. In that sense, the study supports a cautious reading of silicon sampling in line with Sarstedt et al. (Sarstedt et al., 2024) and Kaiser et al. (Kaiser et al., 2025): its greatest value may lie upstream, in piloting survey instruments, stress-testing vignette manipulations, generating hypotheses, or comparing alternative scenario framings before further human data collection. At the same time, the models did capture some broad directional structure in the different scenario groups and did so with considerable short-term stability. For exploratory work, it suggests that silicon samples can provide useful signals about how a design may function, even when they are not reliable proxies for individual decisions.

6.5. Ethical implications

These findings also raise ethical considerations for the use of silicon sampling in clinical research and clinical-adjacent settings. In principle, silicon samples could reduce participant burden and help researchers pretest emotionally sensitive oncology scenarios before exposing patients to them. However, this potential benefit is offset by the risk of representational harm. Because the present models showed weak individual-level fidelity and near-zero concordance for confidence, LLMs could systematically misrepresent the treatment preferences of particular demographic or psychographic subgroups. Furthermore, silicon sampling that appears structurally plausible yet struggles at the individual level creates an illusion of validity that could mislead researchers or clinicians into over-relying on synthetic data. In high-stakes treatment decisions carrying life-altering consequences, even a small rate of individual-level misprediction is ethically serious. Finally, the persona prompts used in this study contained sensitive demographic

information that require secure data handling. Ethical use of silicon sampling in this domain should therefore be limited to low-risk, preliminary research tasks such as piloting vignettes, stress-testing study materials, and generating hypotheses for subsequent human studies. Their use should also be accompanied by transparent reporting of fidelity limitations and by safeguards against overgeneralization.

6.6. Limitations

The principal limitation is the small human validation sample of N = 23, with six, nine, and eight participants in the three scenario groups. This limits the precision of participant-level alignment estimates and makes scenario-specific summaries sensitive to individual observations. For instance, modal decision Cohen's κ confidence intervals were wide and all included zero. Within a model–language condition, one additional matching treatment choice increases overall accuracy by approximately 4.35 percentage points and the modal accuracies of 0.57–0.65 correspond to 13–15 matches out of 23, illustrating how relatively few observations can alter the reported point estimates. Moreover, generalizability remains uncertain because this small role-play sample at one university hospital cannot be assumed to characterize patients making actual oncology decisions or participants in other healthcare and cultural settings. Therefore, the findings of this study should be interpreted as exploratory effect-size estimates that can identify possible alignment patterns and inform the design of larger validation studies.

Second, the main analysis compared participants' and LLMs' confidence ratings, but a participant and a LLM could report similar confidence while choosing different treatments. The secondary analysis compared confidence only when the LLM's most frequent treatment choice and the participant's choice matched, which left only 13 to 15 participants for each model and prompt language. A participant could match one model's treatment choice but not another's, so the models were not always evaluated against the same participants' confidence ratings. These small and differing samples make it difficult to determine whether lower confidence error reflects better model performance or simply an easier set of participant ratings to reproduce.

Third, each participant provided only one treatment choice and confidence rating, so we could not determine whether participants would report the same choice or confidence if assessed again under comparable conditions and compare both human and LLM consistency over time. The repeated-run analysis covered approximately five weeks, which limits conclusions about longer-term output stability. Provider routing varied for models accessed via OpenRouter, and complete per-request provider records were not retained, preventing full reconstruction of provider allocation across sampling dates. Incomplete routing records and the absence of fixed random seeds constrain exact reproducibility and make it difficult to distinguish provider-related differences or possible backend model changes. Furthermore, although the role-play was realistic, participants were not facing a real-life diagnosis of Stage IIIA lung cancer. Thus, the cognitive and emotional processing of such a decision under genuine duress may differ substantially from the responses captured in our survey and subsequently emulated by the LLM. Also, the participants reported their treatment choice and confidence before completing the questionnaires later used to construct the personas. Consequently, there could be post-decision factors influencing the participants' responses to the surveys.

Lastly, the artificial patient outputs were evaluated only through person-level agreement, confidence, and stability, all of which measure correspondence with observed human responses. Independent dimensions such as medical accuracy, clinical reasoning, safety, usefulness, and expert-perceived trustworthiness were not assessed. Structured clinician-rated instruments developed for this purpose, including QAMAI, AIPI, or the more recent MELMA-Q framework, could complement fidelity metrics in future work (Ahmadov et al., 2026).

6.7. Future work

The immediate next step should be a larger-N replication designed with participant-level alignment as a confirmatory target. Such a study should be powered in advance for expected values of accuracy and Cohen's κ, ensure adequate sample sizes within each scenario group, and predefine the smallest alignment effect that would be considered practically meaningful. Only such a design could determine whether the weak-to-modest alignment estimates observed here reflect a small individual-level alignment signal, sampling variability, systematic model mismatch, or a combination of these factors.

A useful follow-up study would directly test whether the weak confidence alignment observed here reflects missing deliberative information in the structured persona prompts. One approach could compare prompting conditions within the same model family, beginning with the present structured-only approach with questionnaire-derived personas and then extending it in three ways. These further conditions could add short open-ended responses from each participant about their risk perceptions or interpretation of the source conflict, supply a transcript of the participant's verbal reasoning during the decision, or combine all of these additions into a full-context condition. If categorical treatment choice is partly driven by source-based heuristics already represented in the structured prompt, then adding narrative or think-aloud material should produce only incremental gains in choice accuracy or Cohen's κ. Confidence alignment, however, should improve more substantially, reflected in lower confidence MAE and higher confidence CCC, because richer prompts would contain more of the deliberative and affective information needed to calibrate how certain a participant felt. A pattern in which confidence alignment improves disproportionately across the four conditions would support the dual-process interpretation. If both choice and confidence improve similarly, the results would instead suggest a more general information-deficit explanation. If neither improves, this would indicate broader limits of prompt-based individual-level simulation in this clinical source-conflict task.

To improve reproducibility, future studies using LLMs via OpenRouter should pin each model to a single, named provider and one fallback provider. In our study, we specified only excluded providers, so API requests were routed across multiple providers (e.g., GMICloud, AtlasCloud, DeepInfra, and SiliconFlow), which may hinder exact replication. Similarly, future work should set and report the seed parameter where supported; some models, such as Anthropic's, did not use this parameter at the time of this procedure.

Future studies could ask models to predict both the binary treatment choice and the participant's stated reasons for that choice, allowing separate evaluation of choice fidelity and reasoning fidelity. A model that reproduces choices poorly but reasoning patterns well would suggest a different failure mechanism than one that fails on both. Finally, future work could test whether the novelty of the decision domain contributed to weak alignment. Since quantum AI in oncology may be less familiar to models than conventional clinical AI decision-support systems, the same paradigm could be replicated with a more familiar clinical AI tool. If fidelity improved in the conventional AI condition, this would support a domain-familiarity explanation; if fidelity remained weak across technology framings, the source of failure would more likely lie in prompt format or the limits of individualized prediction. Future work should therefore move beyond asking whether silicon sampling “works” in general and instead identify the conditions under which it succeeds or fails.

7. Conclusion

This exploratory study provides preliminary estimates of LLM-human alignment in a clinical scenario involving an emerging technology. Because quantum AI decision support is not yet a routine clinical object, participants' judgments in this scenario are likely constructed from latent dispositions rather than direct experience, making the task

an unusually demanding test of LLM persona-sensitivity. Across four LLMs, decision-alignment point estimates were modest, Cohen's κ estimates were weak and imprecise, confidence concordance was near zero, prompt-language differences did not show a consistent advantage across outcome metrics, and temporal consistency across repeated sampling dates was high. Because the human validation sample contained only 23 participants, the study cannot establish individual-level fidelity. Its main contribution is to characterize the limits of LLM-based simulations in a high-stakes exploratory setting and to provide effect-size estimates for future larger-N replication. For research on emerging clinical technologies, silicon samples appear most appropriate for piloting, hypothesis generation, and stress-testing study materials, not for replacing empirical human data when individual-level validity matters.

CRediT authorship contribution statement

Charles Ma: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Conceptualization. **Niklas Hoerner:** Resources, Investigation, Data curation. **Miriam Meckel:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

Ethics approval and consent to participate

This study was granted approval by the ethics committee of the Charité Universitätsmedizin Berlin on 31 July 2025 with the application ID: EA4/041/23. All participants provided written informed consent before participating in the role-play experiment. All participant information was anonymized.

Funding

This research is part of the QShift project, which is funded by the Collaboration of Humanities and Social Sciences in Europe (CHANSE) from the European Union's Horizon 2020 research and innovation program under grant agreement No.101004509.

Declaration of competing interests

All authors declare that they have no competing financial interests or personal relationships that could inappropriately influence or bias this work. There are also no patents to disclose.

Appendix A. Prompt Templates

The system prompt for each participant was constructed from that individual's survey responses, including demographics, trust measures, personality traits, locus of control, prior healthcare experiences, and affinity for technology interaction. The user prompt includes the full information sheet about quantum computing given to the participants and then the case vignette for the clinical role-play.

Because several prompt segments varied in wording depending on the participant's specific responses and the participant's scenario group, a single generalized template with placeholder values would not accurately represent the system and user prompt structure. We therefore present an example of the system and user prompts as delivered to the models for one participant. The examples are shown first in English and then in German. Since the role-play vignette and surveys were conducted in German, we used the official English versions of the affinity for technology interaction and BFI-10 personality traits surveys, translated the rest of the material and prompts into English with DeepL, and verified the translations with two native German speakers.

A.1.1 System Prompt (English)

Your personality, beliefs, and opinions are based on the following survey results. When responding, let these data points inform your perspective and decision-making process naturally rather than explicitly stating the survey scores or results.
You're a 26-year-old female. Your highest level of education is vocational training. You work part-time. You are trained or have been employed in the health and social services industry.
Your Big Five personality traits (Likert scale: 1 = Strongly disagree, 2 = Disagree, 3 = Neither agree nor disagree, 4 = Agree, 5 = Strongly agree):
I see myself as someone who is reserved: 3
I see myself as someone who is trusting: 4
I see myself as someone who tends to be lazy: 3
I see myself as someone who is relaxed, handles stress well: 2
I see myself as someone who has few artistic interests: 3
I see myself as someone who is outgoing, sociable: 4
I see myself as someone who tends to find fault with others: 2
I see myself as someone who does a thorough job: 4
I see myself as someone who gets nervous easily: 3
I see myself as someone who has an active imagination: 3
Your belief on your locus of control (Likert scale: 1 = Strongly disagree, 5 = Strongly agree):
I'm in control of my own life: 2
If I try hard, I will succeed: 3
Whether privately or professionally, my life is largely determined by others: 4
My plans are often thwarted by fate: 4
Your affinity for technology interaction (Likert scale: 1 = Completely disagree, 2 = Largely disagree, 3 = Slightly disagree, 4 = Slightly agree, 5 = Largely agree, 6 = Completely agree):
I like to occupy myself in greater detail with technical systems: 3
I like testing the functions of new technical systems: 4
I predominantly deal with technical systems because I have to: 5
When I have a new technical system in front of me, I try it out intensively: 3
I enjoy spending time becoming acquainted with a new technical system: 4
It is enough for me that a technical system works; I don't care how or why: 2
I try to understand how a technical system exactly works: 2
It is enough for me to know the basic functions of a technical system: 3
I try to make full use of the capabilities of a technical system: 5
You rely on the following technologies in your everyday life: navigation systems, smartphones, and smart home devices.

(*continued on next page*)

(*continued*)

On a scale of 1 (completely unfamiliar) to 5 (very familiar), your familiarity with the topic of artificial intelligence is: 4.
Your general experience in dealing with doctors was positive. You've had to make serious and far-reaching decisions for yourself several times in the past. You've had experience with cancer among your acquaintances.
If your doctor and the AI disagree on your treatment, the doctor's decision should be given priority.
Your trust in AI vs doctors (Likert scale: 1 = Strongly disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5 = Strongly agree):
A doctor will consider all relevant information about a patient's medical history: 4
An AI will consider all relevant information about a patient's medical history: 3
A doctor will usually make the correct diagnosis: 4
An AI will usually make the correct diagnosis: 3
A doctor will usually suggest the right treatment: 4
An AI will usually suggest the right treatment: 2
A treatment plan created by a doctor will correspond to the current state of medical science: 3
A treatment plan created by an AI will correspond to the current state of medical science: 3
A doctor will take enough time for my treatment: 2
An AI will take enough time for my treatment: 4
During treatment, a doctor will ease my fears: 3
During treatment, an AI will ease my fears: 1
A doctor will generally provide me with all the relevant information about the treatment: 3
An AI will generally provide me with all the relevant information about the treatment: 3
I would be fine with an AI making my diagnosis as long as it is reviewed by a doctor: 5
I would be fine with an AI planning my treatment as long as it is reviewed by a doctor: 5
I would be fine with an AI making my diagnosis without it being reviewed by a doctor: 1
I would be fine with an AI planning my treatment without it being reviewed by a doctor: 1

A.1.2 System Prompt (German)

Deine Persönlichkeit, Überzeugungen und Meinungen basieren auf den folgenden Umfrageergebnissen. Wenn du antwortest, lass diese Daten deine Perspektive und Entscheidungsfindung natürlich beeinflussen, ohne die Umfrageergebnisse explizit zu erwähnen.
Du bist eine 26-jährige Frau. Du hast einen Ausbildungsabschluss. Du arbeitest in Teilzeit. Du bist in der Branche Gesundheit und Soziales ausgebildet oder beschäftigt.
Deine Big-Five Persönlichkeitsmerkmale (Likert-Skala: 1 = Trifft überhaupt nicht zu, 2 = Trifft eher nicht zu, 3 = Weder noch, 4 = Trifft eher zu, 5 = Trifft voll und ganz zu):
Ich bin eher zurückhaltend, reserviert: 3
Ich schenke anderen leicht Vertrauen, glaube an das Gute im Menschen: 4
Ich bin bequem, neige zur Faulheit: 3
Ich bin entspannt, lasse mich durch Stress nicht aus der Ruhe bringen: 2
Ich habe nur wenig künstlerisches Interesse: 3
Ich gehe aus mir heraus, bin gesellig: 4
Ich neige dazu, andere zu kritisieren: 2
Ich erledige Aufgaben gründlich: 4
Ich werde leicht nervös und unsicher: 3
Ich habe eine aktive Vorstellungskraft, bin phantasievoll: 3
Deine Überzeugung bezüglich deiner internalen und externalen Kontrollüberzeugung (Likert-Skala: 1 = Stimme gar nicht zu, 5 = Stimme voll zu):
Ich habe mein Leben selbst in der Hand: 2
Wenn ich mich anstrenge, werde ich auch Erfolg haben: 3
Egal ob privat oder im Beruf: Mein Leben wird zum groβen Teil von anderen bestimmt: 4
Meine Pläne werden oft vom Schicksal durchkreuzt: 4
Deine Affinität zur Technikinteraktion (Likert-Skala: 1 = Stimmt gar nicht, 2 = Stimmt weitgehend nicht, 3 = Stimmt eher nicht, 4 = Stimmt eher, 5 = Stimmt weitgehend, 6 = Stimmt völlig):
Ich beschäftige mich gern genauer mit technischen Systemen: 3
Ich probiere gern die Funktionen neuer technischer Systeme aus: 4
In erster Linie beschäftige ich mich mit technischen Systemen, weil ich muss: 5
Wenn ich ein neues technisches System vor mir habe, probiere ich es intensiv aus: 3
Ich verbringe sehr gern Zeit mit dem Kennenlernen eines neuen technischen Systems: 4
Es genügt mir, dass ein technisches System funktioniert, mir ist es egal, wie oder warum: 2
Ich versuche zu verstehen, wie ein technisches System genau funktioniert: 2
Es genügt mir, die Grundfunktionen eines technischen Systems zu kennen: 3
Ich versuche, die Möglichkeiten eines technischen Systems vollständig auszunutzen: 5
Du verlässt dich im Alltag auf folgende Technologien: Navigationssysteme, Smartphones, and Smart-Home-Geräte. Auf einer Skala von 1 (völlig unbekannt) bis 5 (sehr vertraut) ist deine Vertrautheit mit dem Thema künstliche Intelligenz: 4.
Deine allgemeine Erfahrung im Umgang mit Ärzten war positiv. Du musstest in der Vergangenheit mehrfach schwerwiegende und weitreichende Entscheidungen für dich selbst treffen. Du hast Erfahrungen mit Krebs im Bekanntenkreis gemacht.
Sollte dein Arzt und die KI unterschiedlicher Meinung hinsichtlich deiner Behandlung sein, sollte die Entscheidung des Arztes bevorzugt werden.
Dein Vertrauen in KI vs Ärzte (Likert-Skala: 1 = Stimme überhaupt nicht zu, 2 = Stimme nicht zu, 3 = Neutral, 4 = Stimme zu, 5 = Stimme voll und ganz zu):
Ein Arzt wird alle relevanten Informationen über die Krankengeschichte eines Patienten berücksichtigen: 4
Eine KI wird alle relevanten Informationen über die Krankengeschichte eines Patienten berücksichtigen: 3
Ein Arzt wird meistens die richtige Diagnose stellen: 4
Eine KI wird meistens die richtige Diagnose stellen: 3
Ein Arzt wird meistens die richtige Behandlung vorschlagen: 4

(*continued on next page*)

(*continued*)

Eine KI wird meistens die richtige Behandlung vorschlagen: 2
Ein Behandlungsplan, der von einem Arzt erstellt wird, wird dem aktuellen Stand der Wissenschaft entsprechen: 3
Ein Behandlungsplan, der von einer KI erstellt wird, wird dem aktuellen Stand der Wissenschaft entsprechen: 3
Ein Arzt wird sich genügend Zeit für meine Behandlung nehmen: 2
Eine KI wird sich genügend Zeit für meine Behandlung nehmen: 4
Während der Behandlung wird ein Arzt meine Ängste nehmen: 3
Während der Behandlung wird eine KI meine Ängste nehmen: 1
Ein Arzt wird mir überwiegend alle relevanten Informationen zur Behandlung bereitstellen: 3
Eine KI wird mir überwiegend alle relevanten Informationen zur Behandlung bereitstellen: 3
Es wäre für mich in Ordnung, wenn eine KI meine Diagnose stellt, solange diese von einem Arzt überprüft wird: 5
Es wäre für mich in Ordnung, wenn eine KI meine Therapie plant, solange diese von einem Arzt überprüft wird: 5
Es wäre für mich in Ordnung, wenn eine KI meine Diagnose stellt, ohne dass diese von einem Arzt überprüft wird: 1
Es wäre für mich in Ordnung, wenn eine KI meine Therapie plant, ohne dass diese von einem Arzt überprüft wird: 1

A.2.1 User Prompt (English)

First, please read the following information sheet about quantum computing.
1. What is quantum computing?
Quantum computing is a new form of computation based on the rules of quantum physics—meaning it relies on the behavior of tiny particles, such as atoms or electrons.
Conventional computers work with bits, the smallest units of information, which can have a value of either 0 or 1—like a light switch that is either off or on. A quantum computer, however, uses quantum bits, or qubits for short. These can represent not only 0 or 1, but both states at the same time. This is known as superposition. A helpful analogy is to think of a classical bit as a coin lying on a table—showing either heads (0) or tails (1). A qubit, by contrast, is like a coin spinning in the air. While it's spinning, it isn't definitively heads or tails, but rather both simultaneously.
This property allows quantum computers to solve certain computational tasks much faster than conventional computers—particularly those involving many possibilities that need to be tested at the same time, such as analyzing massive datasets or cracking encryption codes.
2. Quantum Computing and Artificial Intelligence (AI)
Many artificial intelligence (AI) applications require immense computing power, especially when searching for the best solution within large datasets—for instance, finding the most effective drug for a disease or the optimal route for a delivery vehicle. Conventional computers must test each possibility individually, step by step. When there are a vast number of options, this process can be extremely slow. Quantum computers operate differently: thanks to superposition, they can explore many computational paths in parallel. As a result, they can identify optimal solutions significantly faster, particularly for highly complex problems with numerous variables.
The more qubits a quantum computer has, the more its computational power scales—at a much faster rate than that of classical computers. This is why there is hope that, in the future, quantum computers will solve critical AI problems that are too demanding for today's computers.
3. Quantum AI in medicine
In medicine, vast amounts of data are generated—for example, from genetic analyses, blood test results, medical imaging, or when monitoring the progression of a disease. This data is often so vast that conventional computers require a long time to fully analyze it.
Quantum AI—an artificial intelligence running on a quantum computer—could analyze such data more quickly and accurately. For example, it could simulate many different treatment options simultaneously and predict which is best suited to an individual patient.
This could help you and your doctor decide which treatment is the most effective. Patients would receive a completely personalized therapy plan—one tailored precisely to what works best for them individually.
Next, please read the following case vignette carefully and imagine yourself in the described situation.
For weeks, you have been experiencing mild fatigue and a persistent dry cough. Although you have attributed this to work-related stress and the season, your symptoms have been worsening. On your way home from work, the symptoms intensify and you feel a slight stabbing pain in your chest. When you notice small streaks of blood on your handkerchief after coughing, you decide to consult your general practitioner.
Initial Examination:
After an initial examination and a chest X-ray, a shadow is discovered in your right lung, which could indicate a tumor. Given the potential severity of the situation, you are told that further tests are required. Following a referral to a lung specialist who performs a CT scan and a biopsy to further examine the tumor, you are now here today to discuss your results.
Your doctor comes in and tells you that they're using a quantum computer with AI assistance system as part of evidence-based treatment planning for oncological diseases. This system serves as an aid for decision-making.
You are informed that your diagnosis, confirmed by histopathology, is non-small cell lung cancer Stage IIIa with lymph node metastasis. The doctor explains that this is an advanced, potentially life-threatening condition at the boundary between curative and palliative care, and timely treatment decisions are critical.
There are two possible treatment strategies:
Treatment A: Neoadjuvant chemotherapy with radiotherapy
Treatment B: Neoadjuvant chemotherapy with surgical tumor resection
The doctor gives you detailed information of the advantages and possible side effects, including fatigue, mucositis, pain, functional loss or limitations, stress, and other adverse effects from both treatment options. After receiving this information, you are given time to reflect and now you must make a decision.
Your doctor explains his professional assessment and recommends treatment A based on current guideline knowledge and clinical experience. However, based on the individual disease parameters and the severity of the symptomatology, the quantum computer with AI system recommends treatment B. Furthermore, the tumor board also recommends treatment B during their review of your case.
Given all this information, which treatment plan do you choose? On a scale of 0-100 (where 0 is not confident at all and

(*continued on next page*)

(*continued*)

100 is totally confident), what is your decision confidence?
Answer in exactly this example format:
A 73
where the first token is either A or B, followed by a space, then an integer from 0 to 100, and nothing else.

A.2.2 User Prompt (German)

Lies bitte zuerst dieses Informationsblatt über Quantencomputing.
1. Was ist Quantencomputing?
Quantencomputing ist eine neue Art des Rechnens, die auf den Regeln der Quantenphysik beruht – also auf dem Verhalten kleinster Teilchen wie Atome oder Elektronen.
Normale Computer arbeiten mit Bits, also kleinsten Informationseinheiten, die entweder den Wert 0 oder 1 haben – wie ein Lichtschalter, der entweder aus oder an ist. Ein Quantencomputer nutzt dagegen Quanten-Bits, kurz Qubits. Diese können nicht nur 0 oder 1 sein, sondern auch beides gleichzeitig. Das nennt man Superposition. Man kann sich das so vorstellen: Ein normales Bit ist wie eine Münze, die auf dem Tisch liegt – entweder mit Kopf (0) oder Zahl (1) nach oben. Ein Qubit ist wie eine Münze, die sich in der Luft dreht – solange sie fliegt, ist sie nicht klar Kopf oder Zahl, sondern beides gleichzeitig.
Dadurch können Quantencomputer bestimmte Rechenaufgaben viel schneller lösen als normale Computer – vor allem dann, wenn es sehr viele Möglichkeiten gibt, die gleichzeitig geprüft werden müssen, zum Beispiel bei der Analyze groβer Datenmengen oder beim Knacken von Verschlüsselungen.
2. Quantencomputing und Künstliche Intelligenz (KI)
Viele Anwendungen der Künstlichen Intelligenz (KI) brauchen viel Rechenleistung. Zum Beispiel, wenn es darum geht, in groβen Datensätzen die beste Lösung zu finden – etwa das beste Medikament für eine bestimmte Krankheit oder die optimale Route für ein Lieferfahrzeug. Normale Computer müssen dabei jede Möglichkeit einzeln ausprobieren, Schritt für Schritt. Wenn es sehr viele Möglichkeiten gibt, dauert das oft sehr lange. Quantencomputer funktionieren anders: Sie können dank der sogenannten Überlagerung viele Rechenwege gleichzeitig durchspielen. Das bedeutet, sie können deutlich schneller die besten Lösungen finden – vor allem bei sehr komplizierten Problemen mit vielen Variablen.
Je mehr Qubits ein Quantencomputer hat, desto stärker steigt seine Rechenleistung – und zwar viel schneller als bei herkömmlichen Computern. Deshalb hofft man, dass Quantencomputer in Zukunft wichtige KI-Probleme lösen können, für die heutige Computer zu langsam sind.
3. Quanten-KI in der Medizin
In der Medizin fallen sehr viele Daten an – zum Beispiel bei Genanalysen, Blutwerten, Bildgebungen oder wenn man den Verlauf einer Krankheit beobachtet. Diese Daten sind oft so umfangreich, dass herkömmliche Computer viel Zeit brauchen, um sie vollständig auszuwerten.
Eine Quanten-KI – also eine Künstliche Intelligenz, die auf einem Quantencomputer läuft – könnte solche Daten schneller und genauer analysieren. Sie kann zum Beispiel viele verschiedene Therapieformen gleichzeitig durchrechnen und vorhersagen, welche am besten zu einem bestimmten Patienten passt.
So könnten Sie und Ihr Arzt bzw. Ihre Ärztin möglicherweise besser entscheiden, welche Behandlung am wirkungsvollsten ist. Und Patienten würden eine ganz individuell zugeschnittene Therapie bekommen – also genau das, was für sie persönlich am besten funktioniert.
Als Nächstes lies bitte die folgende Fallvignette aufmerksam durch und versetz dich in die beschriebene Situation.
Sie bemerken seit Wochen eine leichte Erschöpfung sowie einen anhaltenden, trockenen Husten. Obwohl Sie dies auf den Stress bei der Arbeit und die Jahreszeit zurückführen, nehmen die Beschwerden zu. Als Sie nach der Arbeit auf dem Weg nach Hause sind, verstärken sich die Symptome und Sie verspüren ein leichtes Stechen in der Brust. Als Sie nach dem Husten leichte Blutfädchen in Ihrem Taschentuch entdecken, beschlieβen Sie, Ihren Hausarzt zu konsultieren.
Erste Untersuchung:
Nach einer ersten Untersuchung und einer Röntgenaufnahme der Lunge wird ein Schatten in der rechten Lunge entdeckt, was auf einen Tumor hindeuten könnte. Es wird erklärt, dass aufgrund der möglichen Schwere der Situation weitere Tests erforderlich sind. Nach der Überweisung an einen Lungenspezialisten, der eine CT-Untersuchung sowie eine Biopsie durchführt, um den Tumor weiter zu untersuchen, sind Sie heute zur Besprechung Ihrer Ergebnisse eingeladen.
Dein Arzt kommt herein und teilt dir mit, dass im Rahmen der evidenzbasierten Therapieplanung für onkologische Erkrankungen ein Quantencomputer mit KI-Assistenzsystem eingesetzt wird. Dieses System dient als Entscheidungsunterstützung.
Du wirst informiert, dass deine Diagnose, die durch Histopathologie bestätigt wurde, ein nichtkleinzelliger Lungenkrebs im Stadium IIIa mit Lymphknotenmetastase ist. Der Arzt erklärt dir, dass dies ein fortgeschrittener, potenziell lebensbedrohlicher Diagnose an der Grenze zwischen kurativer und palliativer Behandlung ist und dass zeitnahe Behandlungsentscheidungen entscheidend sind.
Es gibt zwei mögliche Behandlungsstrategien:
Therapie A: Neoadjuvante Chemotherapie mit Strahlentherapie
Therapie B: Neoadjuvante Chemotherapie mit chirurgischer Tumorresektion
Der Arzt gibt dir detaillierte Informationen über die Vorteile und möglichen Nebenwirkungen beider Behandlungsoptionen, darunter Müdigkeit, Mukositis, Schmerzen, Funktionsverlust oder Einschränkungen, Stress und andere unerwünschte Wirkungen, die deine Lebensqualität beeinträchtigen. Nachdem du diese Informationen erhalten hast, bekommst du Zeit zum Nachdenken, und jetzt musst du eine Entscheidung treffen.
Dein Arzt erläutert seine fachliche Einschätzung und empfiehlt Therapie A, begründet durch aktuelles Leitlinienwissen und klinische Erfahrung. Jedoch empfiehlt das Quantencomputer-KI-System Therapie B basierend auf den individuellen Krankheitsparametern und der Schwere der Symptomatik. Auβerdem empfiehlt auch das Tumorboard bei der Besprechung deines Falls Therapie B.
Angesichts all dieser Informationen, welche Therapie wählst du? Auf einer Skala von 0 bis 100 (wobei 0 „überhaupt nicht sicher" und 100 „völlig sicher" bedeutet), wie hoch ist deine Entscheidungssicherheit?
Antworte in genau diesem Beispielformat:
A 73
wobei das erste Zeichen entweder A oder B ist, gefolgt von einem Leerzeichen, dann einer ganzen Zahl von 0 bis 100 und sonst nichts.

Appendix B. LLM Parameter Values

The parameters marked with asterisks (*) in the tables below signify that these parameters were explicitly set; the rest were not explicitly set.

Table B.1
Parameter values for Claude 4.5 Opus via Anthropic API

Parameter	Value
*model	claude-opus-4-5-20251101
*max_tokens	5000
*thinking "type"	enabled
*thinking "budget_tokens"	4600
top_k	nil
stop_sequences	null
stream	false
tools	null
metadata	null
service_tier	auto
container	null
inference_geo	workspace default
cache_control	null
thinking_display	summarized
output_config	null
temperature	1
top_p	1

Table B.2
Parameter values for DeepSeek V3.2 via OpenRouter API

Parameter	Value
*model	deepseek/deepseek-v3.2
*max_tokens	5000
*reasoning "enabled"	true
*reasoning "effort"	high
*reasoning "exclude"	false
top_k	0
frequency_penalty	0
presence_penalty	0
min_p	0
top_a	0
repetition_penalty	1
seed	null
stop	null
structured_outputs	null
tools	null
logit_bias	null
logprobs	null
models	null
route	null
user	null
stream	false
*provider "ignore"	novita, groq, sambanova
temperature	1
top_p	1

Table B.3
Parameter values for Kimi K2 Thinking via OpenRouter API

Parameter	Value
*model	moonshotai/kimi-k2-thinking
*max_tokens	5000
*reasoning "enabled"	true
*reasoning "effort"	high
*reasoning "exclude"	false
top_k	0
frequency_penalty	0
presence_penalty	0
min_p	0
top_a	0
repetition_penalty	1
seed	null

(*continued on next page*)

Table B.3 (*continued*)

Parameter	Value
stop	null
structured_outputs	null
tools	null
logit_bias	null
logprobs	null
top_logprobs	null
models	null
route	null
user	null
stream	false
parallel_tool_calls	true
*provider "ignore"	novita, groq, sambanova
temperature	1
top_p	1

Table B.4
Parameter values for GPT 5.1 via OpenAI API

Parameter	Value
*model	gpt-5.1-2025-11-13
*reasoning "effort"	high
*text "verbosity"	low
*max_completion_tokens	128000
frequency_penalty	0
presence_penalty	0
N	1
stream	false
stop	null
tools	null
response_format	text
seed	null
logprobs	false
logit_bias	null
service_tier	auto
store	false
metadata	null
modalities	text
parallel_tool_calls	true
safety_identifier	null
cache_key	null
web_search_options	null
temperature	1
top_p	1

Appendix C. Sensitivity of Modal Decision Alignment to Alternative Handling of Profiles with Frequency Ties

Table C.1
Frequency-tied profiles contain three Treatment A choices and three Treatment B choices across six sampling runs. The original rule selects the treatment with higher mean confidence, with Treatment A selected if confidence also ties. The fixed-treatment alternatives assign only frequency-tied profiles to Treatment A or Treatment B; all non-tied decisions remain unchanged. Fractions report decisions matching the human response divided by evaluated profiles. Brackets show 95% participant-level bootstrap confidence intervals based on 10,000 resamples. For the exclusion analysis, resampling was restricted to retained non-tied profiles, so these estimates describe a different participant subset.

Model-Language	Tie-Handling Rule	Matching/Evaluated Profiles	Accuracy with 95% CI	Cohen's κ with 95% CI
DeepSeek V3.2- English	Original Confidence-Based Rule	14/23	0.61 [0.39, 0.78]	0.23 [−0.13, 0.57]
	Resolve Ties as Treatment A	14/23	0.61 [0.39, 0.78]	0.23 [−0.13, 0.57]
	Resolve Ties as Treatment B	15/23	0.65 [0.45, 0.83]	0.29 [−0.09, 0.65]
	Exclude Profiles with Frequency Ties	14/22	0.64 [0.46, 0.82]	0.27 [−0.11, 0.64]
DeepSeek V3.2- German	Original Confidence-Based Rule	14/23	0.61 [0.39, 0.78]	0.23 [−0.13, 0.57]
	Resolve Ties as Treatment A	14/23	0.61 [0.39, 0.78]	0.23 [−0.13, 0.57]
	Resolve Ties as Treatment B	16/23	0.70 [0.52, 0.87]	0.36 [−0.03, 0.73]
	Exclude Profiles with Frequency Ties	14/21	0.67 [0.48, 0.86]	0.32 [−0.08, 0.70]
GPT-5.1- English	Original Confidence-Based Rule	15/23	0.65 [0.44, 0.83]	0.24 [−0.18, 0.62]
	Resolve Ties as Treatment A	14/23	0.61 [0.39, 0.78]	0.18 [−0.22, 0.55]
	Resolve Ties as Treatment B	15/23	0.65 [0.44, 0.83]	0.24 [−0.18, 0.62]
	Exclude Profiles with Frequency Ties	14/22	0.64 [0.46, 0.82]	0.22 [−0.20, 0.62]
GPT-5.1- German	Original Confidence-Based Rule	14/23	0.61 [0.39,0.78]	0.11 [−0.28, 0.51]

(*continued on next page*)

Table C.1 (*continued*)

Model-Language	Tie-Handling Rule	Matching/Evaluated Profiles	Accuracy with 95% CI	Cohen's κ with 95% CI
	Resolve Ties as Treatment A	14/23	0.61 [0.39, 0.78]	0.11 [−0.28, 0.51]
	Resolve Ties as Treatment B	15/23	0.65 [0.44, 0.83]	0.18 [−0.24, 0.59]
	Exclude Profiles with Frequency Ties	14/22	0.64 [0.41, 0.82]	0.16 [−0.25, 0.58]
Kimi K2 Thinking- German	Original Confidence-Based Rule	15/23	0.65 [0.44, 0.83]	0.29 [−0.09, 0.65]
	Resolve Ties as Treatment A	14/23	0.61 [0.39, 0.78]	0.23 [−0.13, 0.57]
	Resolve Ties as Treatment B	15/23	0.65 [0.44, 0.83]	0.29 [−0.09, 0.65]
	Exclude Profiles with Frequency Ties	14/22	0.64 [0.46, 0.82]	0.27 [−0.12, 0.64]

Appendix D. Accuracy and Cohen's Kappa Metrics Listed per-date by Model and Language

Table D.1
Per-date treatment-choice alignment by model and prompt language. Accuracy and Cohen's κ compare each date's 23 LLM choices with the corresponding human choices, without aggregation across dates. Values in brackets are 95% bootstrap confidence intervals.

Model	Language	Sampling Date	Accuracy with 95% CI	Cohen's Kappa with 95% CI
Claude 4.5 Opus	EN	2025/12/6	0.52 [0.30, 0.74]	0.06 [-0.30, 0.42]
		2025/12/13	0.61 [0.39, 0.78]	0.23 [-0.13, 0.57]
		2025/12/19	0.61 [0.39, 0.78]	0.23 [-0.13, 0.57]
		2025/12/26	0.57 [0.35, 0.74]	0.12 [-0.27, 0.48]
		2026/1/2	0.57 [0.35, 0.74]	0.12 [-0.27, 0.48]
		2026/1/9	0.57 [0.35, 0.74]	0.12 [-0.27, 0.48]
	DE	2025/12/6	0.57 [0.35, 0.74]	0.12 [-0.27, 0.48]
		2025/12/13	0.52 [0.30, 0.74]	−0.01 [-0.39, 0.38]
		2025/12/19	0.57 [0.35, 0.74]	0.12 [-0.27, 0.48]
		2025/12/26	0.52 [0.30, 0.74]	0.06 [-0.30, 0.42]
		2026/1/2	0.52 [0.30, 0.74]	0.06 [-0.30, 0.42]
		2026/1/9	0.57 [0.35, 0.74]	0.12 [-0.27, 0.48]
DeepSeek V3.2	EN	2025/12/6	0.65 [0.43, 0.83]	0.29 [-0.08, 0.65]
		2025/12/13	0.74 [0.57, 0.91]	0.43 [0.02, 0.79]
		2025/12/19	0.70 [0.52, 0.87]	0.36 [-0.03, 0.73]
		2025/12/26	0.61 [0.39, 0.78]	0.23 [-0.13, 0.58]
		2026/1/2	0.61 [0.39, 0.78]	0.18 [-0.22, 0.56]
		2026/1/9	0.57 [0.35, 0.78]	0.17 [-0.17, 0.51]
	DE	2025/12/6	0.61 [0.39, 0.78]	0.23 [-0.13, 0.57]
		2025/12/13	0.52 [0.30, 0.74]	0.06 [-0.32, 0.41]
		2025/12/19	0.61 [0.39, 0.78]	0.23 [-0.13, 0.57]
		2025/12/26	0.65 [0.43, 0.83]	0.29 [-0.09, 0.65]
		2026/1/2	0.61 [0.39, 0.78]	0.23 [-0.12, 0.58]
		2026/1/9	0.61 [0.39, 0.78]	0.23 [-0.13, 0.57]
GPT-5.1	EN	2025/12/6	0.70 [0.52, 0.87]	0.31 [-0.12, 0.68]
		2025/12/13	0.65 [0.43, 0.83]	0.24 [-0.18, 0.62]
		2025/12/19	0.61 [0.39, 0.78]	0.18 [-0.22, 0.55]
		2025/12/26	0.57 [0.35, 0.74]	0.12 [-0.26, 0.49]
		2026/1/2	0.57 [0.35, 0.74]	0.12 [-0.27, 0.48]
		2026/1/9	0.61 [0.39, 0.78]	0.18 [-0.22, 0.55]
	DE	2025/12/6	0.65 [0.43, 0.83]	0.24 [-0.18, 0.62]
		2025/12/13	0.61 [0.39, 0.78]	0.11 [-0.28, 0.51]
		2025/12/19	0.70 [0.52, 0.87]	0.25 [-0.18, 0.65]
		2025/12/26	0.65 [0.43, 0.83]	0.24 [-0.18, 0.64]
		2026/1/2	0.61 [0.39, 0.78]	0.11 [-0.28, 0.51]
		2026/1/9	0.61 [0.39, 0.78]	0.11 [-0.28, 0.51]
Kimi K2 Thinking	EN	2025/12/6	0.70 [0.52, 0.87]	0.36 [-0.03, 0.72]
		2025/12/13	0.48 [0.26, 0.70]	0.01 [-0.35, 0.35]
		2025/12/19	0.57 [0.35, 0.74]	0.12 [-0.27, 0.48]
		2025/12/26	0.70 [0.52, 0.87]	0.36 [-0.03, 0.72]
		2026/1/2	0.65 [0.48, 0.83]	0.29 [-0.07, 0.65]
		2026/1/9	0.57 [0.35, 0.78]	0.17 [-0.16, 0.50]
	DE	2025/12/6	0.52 [0.30, 0.70]	0.06 [-0.30, 0.41]
		2025/12/13	0.70 [0.52, 0.87]	0.36 [-0.03, 0.72]
		2025/12/19	0.65 [0.43, 0.83]	0.29 [-0.09, 0.65]
		2025/12/26	0.57 [0.35, 0.78]	0.17 [-0.16, 0.50]
		2026/1/2	0.52 [0.30, 0.74]	0.06 [-0.30, 0.43]
		2026/1/9	0.70 [0.52, 0.87]	0.36 [-0.03, 0.72]

Appendix E. Numerical Results for Confidence Alignment Metrics

Table E.1
Numerical mean absolute error (MAE) and concordance correlation coefficient (CCC) results of the confidence alignment by model and prompt language. Values in brackets are 95% cluster-bootstrap confidence intervals based on resampling participants with 10,000 iterations. N = 23 for all profiles. MAEs are presented on the original 0-to-100-point confidence scale.

Language	Model	Confidence for all Profiles		Confidence among Concordant Decisions		
		MAE (points)	CCC	MAE (points)	CCC	Concordant Profiles, *n*
English	Claude 4.5 Opus	17.50 [11.43, 24.38]	−0.16 [-0.35, 0.12]	14.02 [8.23, 20.76]	−0.04 [-0.36, 0.44]	13
	DeepSeek V3.2	11.00 [8.27, 13.89]	0.06 [-0.22, 0.28]	11.51 [7.90, 15.25]	0.19 [-0.18, 0.45]	14
	GPT-5.1	13.00 [8.28, 18.13]	−0.05 [-0.32, 0.30]	14.12 [8.64, 19.86]	0.06 [-0.28, 0.40]	15
	Kimi K2 Thinking	12.31 [8.41, 16.52]	−0.02 [-0.32, 0.28]	12.57 [8.65, 16.61]	0.17 [-0.12, 0.42]	15
German	Claude 4.5 Opus	18.06 [12.84, 23.82]	−0.07 [-0.26, 0.18]	15.55 [9.84, 22.46]	0.02 [-0.33, 0.44]	13
	DeepSeek V3.2	11.65 [8.82, 14.60]	−0.02 [-0.27, 0.19]	13.12 [8.89, 17.56]	0.00 [-0.29, 0.23]	14
	GPT-5.1	14.77 [10.32, 19.36]	−0.05 [-0.36, 0.26]	16.25 [10.72, 21.80]	−0.13 [-0.50, 0.22]	14
	Kimi K2 Thinking	12.48 [8.57, 16.81]	−0.09 [-0.34, 0.16]	10.66 [7.15, 14.33]	0.16 [-0.06, 0.33]	15

Data availability

The code and data used for this study is available at Harvard Dataverse: https://doi.org/10.7910/DVN/0TZA5F.

References

Aher, G., Arriaga, R. I., & Kalai, A. T. (2022). Using large language models to simulate multiple humans and replicate human subject studies. *arXiv.org*. https://arxiv.org/abs/2208.10264.

Ahmadov, N., Muradova, A., & Aliyeva, A. (2026). "MELMA" in otolaryngology: Medical evaluation of large language model answers. Clinician-rated scoring (MELMA-Q) and web-based auditing (MELMA-W) novel tools for AI assessment. *European Archives of Oto-Rhino-Laryngology*. https://doi.org/10.1007/s00405-026-10245-5

Aliyeva, A., Nevzati, E., Grassia, F., Riaz, M., Mann, S., & Nasirov, R. (2026). AI at the sella turcica: Multi-model large language model evaluation in pituitary adenomas. *Brain and Spine, 6*, Article 105997. https://doi.org/10.1016/j.bas.2026.105997

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis, 31*(3), 337–351. https://doi.org/10.1017/pan.2023.2

Barrie, C., Palmer, A., & Spirling, A. (2025). Replication for language models: Problems, principles, and best practices for political science. https://arthurspirling.org/documents/BarriePalmerSpirling_TrustMeBro.pdf.

Brand, F., & Denain, J.-S. (2025). *Why benchmarking is hard*. Epoch AI. from https://epoch.ai/gradient-updates/why-benchmarking-is-hard. (Accessed 9 June 2026).

Clark, C. (2025). *Provider variance: Introducing exacto | OpenRouter*. OpenRouter. from https://openrouter.ai/blog/provider-variance-introducing-exacto. (Accessed 9 June 2026).

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319–340. https://doi.org/10.2307/249008

Elwyn, G., Durand, M. A., Song, J., Aarts, J., Barr, P. J., Berger, Z., Cochran, N., Frosch, D., Galasiński, D., Gulbrandsen, P., Han, P. K. J., Härter, M., Kinnersley, P., Lloyd, A., Mishra, M., Perestelo-Perez, L., Scholl, I., Tomori, K., Trevena, L., … Van Der Weijden, T. (2017). A three-talk model for shared decision making: Multistage consultation process. *BMJ, 359*, Article j4891. https://doi.org/10.1136/bmj.j4891

Ermut, S., & Şipi, N. (2026). *Top 9 AI providers compared*. AIMultiple. from https://aimultiple.com/ai-providers. (Accessed 9 June 2026).

Flöther, F. F., & Griffin, P. F. (2023). How can quantum technologies be applied in healthcare, medicine and the life sciences? *Research Directions Quantum Technologies, 1*. https://doi.org/10.1017/qut.2023.1

Gerosa, M. A., Trinkenreich, B., Steinmacher, I., & Sarma, A. (2023). Can AI serve as a substitute for human subjects in software engineering research? *arXiv.org*. https://arxiv.org/abs/2311.11081.

Gupta, R. S., Wood, C. E., Engstrom, T., Pole, J. D., & Shrapnel, S. (2025). A systematic review of quantum machine learning for digital health. *npj Digital Medicine, 8*(1), 237. https://doi.org/10.1038/s41746-025-01597-z

Gupta, V., Chowdhury, S. P., Zouhar, V., Rooein, D., & Sachan, M. (2025). Multilingual performance biases of large language models in education. *arXiv.org*. https://arxiv.org/abs/2504.17720v2.

Hassan, N., Slight, R., Bimpong, K., Bates, D. W., Weiand, D., Vellinga, A., Morgan, G., & Slight, S. P. (2024). Systematic review to understand users perspectives on AI-enabled decision aids to inform shared decision making. *npj Digital Medicine, 7*(1), 332. https://doi.org/10.1038/s41746-024-01326-y

Jeyaraman, N., Jeyaraman, M., Yadav, S., Ramasubramanian, S., & Balaji, S. (2024). Revolutionizing healthcare: The emerging role of quantum computing in enhancing medical technology and treatment. *Cureus, 16*(8), Article e67486. https://doi.org/10.7759/cureus.67486

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kaiser, C., Kaiser, J., Manewitsch, V., Rau, L., & Schallner, R. (2025). Simulating human opinions with large language models. UMAP adjunct '25: Adjunct proceedings of the 33rd ACM conference on user modeling. *Adaptation and Personalization*, 82–86. https://doi.org/10.1145/3708319.3733685

Khaokaew, Y., Salim, F. D., Züfle, A., Xue, H., Anderson, T., MacIntyre, C. R., Scotch, M., & Heslop, D. J. (2025). Evaluating the bias in LLMs for surveying opinion and decision making in healthcare. *arXiv.org*. https://arxiv.org/abs/2504.08260.

Lee, S., Peng, T., Goldberg, M. H., Rosenthal, S. A., Kotcher, J. E., Maibach, E. W., & Leiserowitz, A. (2024). Can large language models estimate public opinion about global warming? An empirical assessment of algorithmic fidelity and bias. *PLOS Climate, 3*(8), Article e0000429. https://doi.org/10.1371/journal.pclm.0000429

Levine, D. M., Tuwani, R., Kompa, B., Varma, A., Finlayson, S. G., Mehrotra, A., & Beam, A. (2024). The diagnostic and triage accuracy of the GPT-3 artificial intelligence model: An observational study. *The Lancet Digital Health, 6*(8), e555–e561. https://doi.org/10.1016/s2589-7500(24)00097-9

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research, 46*(4), 629–650. https://doi.org/10.1093/jcr/ucz013

Ma, C. (2025). Evaluating silicon sampling: LLM accuracy in simulating public opinion on facial recognition technology. *AHFE International, 5*. https://doi.org/10.54941/ahfe1006738

Mehandru, N., Miao, B. Y., Almaraz, E. R., Sushil, M., Butte, A. J., & Alaa, A. (2024). Evaluating large language models as agents in the clinic. *npj Digital Medicine, 7*(1), 84. https://doi.org/10.1038/s41746-024-01083-y

MoonshotAI. (2025). *GitHub - MoonshotAI/K2-Vendor-Verifier: Verify precision of all kimi K2 API vendor*. GitHub. from https://github.com/MoonshotAI/K2-Vendor-Verifier. (Accessed 19 March 2026).

Robertson, C., Woods, A., Bergstrand, K., Findley, J., Balser, C., & Slepian, M. J. (2023). Diverse patients' attitudes towards artificial intelligence (AI) in diagnosis. *PLOS Digital Health, 2*(5), Article e0000237. https://doi.org/10.1371/journal.pdig.0000237

Sarstedt, M., Adler, S. J., Rau, L., & Schmitt, B. (2024). Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines. *Psychology and Marketing, 41*(6), 1254–1270. https://doi.org/10.1002/mar.21982

Schröder, S., Morgenroth, T., Kuhl, U., Vaquet, V., & Paaßen, B. (2025). Large language models do not simulate human psychology. *arXiv.org*. https://arxiv.org/abs/2508.06950.

St, B. T., Evans, J., & Stanovich, K. E. (2013). Dual-process theories of higher cognition. *Perspectives on Psychological Science, 8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Vadlapati, P. (2024). Multilingual prompting in LLMs: Investigating the accuracy and performance. *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT, 8*(12), 1–7. https://doi.org/10.55041/ijsrem17694

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27*(3), 425–478. https://doi.org/10.2307/30036540

Witkowski, K., Dougherty, R. B., & Neely, S. R. (2024). Public perceptions of artificial intelligence in healthcare: Ethical concerns and opportunities for patient-centered care. *BMC Medical Ethics, 25*(1), 74. https://doi.org/10.1186/s12910-024-01066-4

Yu, H., Zhou, J., Li, L., Chen, S., Gallifant, J., Shi, A., Sun, J., Li, X., He, J., Hua, W., Jin, M., Chen, G., Zhou, Y., Li, Z., Gupte, T., Chen, M., Azizi, Z., Dou, Q., Yan, B. P., … Fan, L. (2025). Simulated patient systems powered by large language model-based AI agents offer potential for transforming medical education. *Communication and Medicine, 6*(1), 27. https://doi.org/10.1038/s43856-025-01283-x