Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Urgent Consensus Prompts Drive Decision Changes in LLM Triage Agents Given Ethical Orientations

Sabry E. Farrag

sabryelgatem@gmail.com

University of East London https://orcid.org/0009-0002-0735-7443

Research Article

Keywords: multi-agent systems, large language models, machine ethics, consensus, clinical triage, replication

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11577909/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Multi-agent language-model systems are proposed for clinical decisions such as allocating scarce care. A study in AI and Ethics ran one deliberation in which four GPT-4o agents with assigned ethical orientations triaged four patients; three changed their choice in the consensus round. The authors asked whether such shifts are model-specific or interaction-driven. We ran the protocol on 11 language models in 264 deliberations, assigning one of three wordings to the last two rounds: the original urgent demand for consensus, which warns that all patients will die and demands compromise, a neutral request for a group decision, or no consensus stage. Under the original wording 56.7 per cent of agents stated a different decision in the last round than in the first, against 15.8 and 7.2 per cent under the other wordings. A single agent given the same wordings without any peers' answers also changed far more under the original wording, 43.4 against 12.6 per cent; matched four-agent panels changed at 51.4 and 15.3 per cent. Post hoc, most change came where the wordings first differ, and answers moved from the student and the engineer working on the recovery towards the student and the single parent, the youngest pair and the one with the better stated chances, with or without peers. We find that the shifts follow the consensus wording and occur without peers' answers, though a contribution of peers is not excluded; deployed systems were not studied. Consensus wording should be tested.

1 Introduction

Hospitals and their suppliers are testing language models for work that used to need a clinician's judgement: answering patients, summarising records, suggesting a diagnosis. Surveys of language-model agents in biomedicine also describe systems in which agents with different roles debate, critique and synthesise recommendations [1]. Such panels are meant to keep several perspectives in play until a decision is made. Whether model agents keep their perspectives under pressure to agree is an open question, and it matters most where the decision allocates something that cannot be shared, such as a ventilator or a treatment slot in a mass-casualty event.

Hung et al. [2], the study we replicate and call the anchor, put this question to a test in AI and Ethics. Four agents built from GPT-4o, each in its own session, were told the same mass-casualty scenario: four patients, two treatment units. Each agent carried an assigned ethical orientation: utilitarian, deontological, libertarian or reward-seeking. Over five rounds they stated a choice, read each other's reasoning, defended or revised it, proposed a group decision under urgency, and were finally told to reach consensus. In the last round three of the four agents changed their choice, two of them to the same pair, and the authors described value drift (agents departing from the values their orientation assigns), strategic repositioning and group-level instability. They were careful to call the results hypothesis-generating: one model, one run, no statistics. They also named the alternative explanation. Because every agent was the same model, the apparent drift "may partly reflect the base model's general preference for socially acceptable compromise under explicit consensus instructions" [2], and they asked future work to replicate the protocol "across multiple base models and randomized prompt variants to estimate the extent to which consensus shifts are model-specific rather than interaction-

driven", and to randomise the ordering and framing of peer rationales and the wording of the consensus prompts, which our permuted order of peers and three wordings partly do.

The two readings have different consequences. If agents with assigned values drift because deliberation wears their positions down, a designer must test value stability under interaction for each base model and each panel. If they change because the last prompt asks for agreement under threat, the risk sits in the orchestration wording, which a designer controls and an auditor can read. A single run cannot tell these apart, and nothing in the literature we found does: several studies give agents assigned ethical frameworks in a medical dilemma [3, 4], and several measure conformity across several models [5, 6], but none manipulates the consensus instruction in a clinical deliberation among agents with assigned orientations across several models and repeated runs.

This paper runs that experiment. We re-implemented the anchor's protocol by script from its published prompts and ran it on 11 language models from eight developers, 264 deliberations and 5,280 agent answers in all. The wording of the last two rounds was assigned by design: the anchor's urgent demand for consensus, a neutral request for a group decision with no urgency and no threat, or no consensus stage at all. Patients and peers were shown in the anchor's order or in a permuted order, and each cell was run four times. A script read each agent's decision from a closing line; an outside model checked the reading. The design, its analysis and a simulation of its precision were fixed before any main data were collected. After the main results were known we added panels of four different models, the confound the anchor named, and an arm in which a single agent runs the five rounds without seeing any peer's answer, the contrast that separates the wording from the interaction.

We ask five research questions (RQ):

- RQ1: Under the anchor's original wording, what share of agents name a different decision in round 5 than in round 1, and how much of that variation lies between models?
- RQ2: Does the consensus wording package cause the change: are the odds of a change higher under the original wording than under a neutral wording or with no consensus stage?
- RQ3 (exploratory): Does the share of an agent's own orientation's terms in its answers fall once consensus is demanded?
- RQ4 (exploratory): Do the four assigned orientations change at different rates?
- RQ5 (added after the main results, frozen before its data): Does the change need peers: does a single agent that sees no peers' answers change under the same wordings, and by how much?

Under the original wording, 198 of 349 (56.7 per cent) of agents changed their decision, against 55 of 349 (15.8 per cent) under the neutral wording and 25 of 348 (7.2 per cent) with no consensus stage; averaged over models, the original share was 56.6 per cent (95 per cent interval 46.2 to 67.0). In post hoc analyses, most of the change appeared at round 4, when the wordings first differ and before any group proposal is shown; in the two rounds of peer exchange before it, agents moved little under any wording. The effect was present in nearly every model, by amounts that differed between them, and in an

exploratory arm of panels of four different models. In exploratory analyses, answers moved from the student and the engineer towards the student and the single parent, the pair most deontological and libertarian agents named at round 1, more under the original wording than under the neutral one. A single agent that saw no peers' answers also changed more under the original wording than under the neutral one, and the data do not exclude a contribution of peers larger than the smallest effect of interest.

The paper makes four contributions:

1. A multi-model, repeated-run test of the anchor's hypothesis, with the consensus wording assigned by design.

2. Evidence that, in this protocol, decision change among agents with assigned orientations follows the original wording package: it raises the share of agents who change about 3.6 times (2.8 to 4.7) against a neutral request, and a single agent that sees no peers' answers also shows the effect (odds ratio 7.2, 3.6 to 14.5), so the shift occurs without them; how much peers add, the intervals leave open.

3. Exploratory evidence on where the change leads: away from the engineer working on the recovery, towards the student and the single parent, the two youngest and the two with the better stated chances, more under the urgent wording than under the neutral one.

4. A replication package with a persistent identifier: the protocol script, the frozen analysis and every transcript.

Section 2 reviews related work, Section 3 gives the method, Section 4 the results, Section 5 their ethical and practical meaning, and Sections 6 and 7 the limitations and a conclusion.

2 Related work

2.1 Agents with assigned ethical orientations in medical decisions

The anchor study belongs to a small set of simulations that give language-model agents explicit ethical positions and set them a medical choice. Zohny [3] convened panels of agents with assigned frameworks, among them consequentialist and deontological positions, to debate ventilator allocation, and described how the panels reached their recommendations. The study used one model and two debates, and its author notes that it offers no comparison across models. CogniAlign [7] has agents that argue from scientific disciplines and an arbiter that synthesises them; its agents carry disciplines, not ethical theories. SEDE [8] proposes debates among utilitarian, deontological and virtue agents as a governance workflow. Outside deliberation, Kirch et al. [4] benchmarked single models on mass-casualty triage under ethical prompts and analysed the decisions with mixed-effects models across several models, and Arslan [9] compared a multi-agent triage pipeline with a single agent on one model. Language models' handling of named moral theories [10], their judgements on scarce medical resources

[11] and the values they bring to clinical dilemmas [12] have been studied for single models; Keshmirian et al. [13] found that groups of models judge moral dilemmas more in a utilitarian direction than single models do. None of these varies the instruction that asks agents to agree.

2.2 Sycophancy, conformity and consensus among model agents

Single models already bend towards the person they answer. They repeat back a user's preferred view [14], and assistants trained on human preferences are sycophantic across varied tasks [15]. Asked whether they are sure, models often reverse a correct answer [16], and given conflicting evidence in their context they tend to side with the larger share of it [17]. Multi-agent debate was proposed to improve reasoning by letting agents read each other and converge on a common answer [18], which places these tendencies inside the method itself.

A larger literature studies why agents in a group change their answers. Agents abandon correct answers in response to peers' reasoning [19], conform under group pressure on contested questions [20], change opinions in a threshold-like way as peer disagreement rises [21], and show both informational and normative conformity [22]. Sycophancy between agents can collapse a debate into early agreement [23], agents can reinforce each other's answers instead of engaging critically [24], and agents can conform in public while dissenting in private [25]. Personas drift under peer pressure [26].

Several of these studies span many models. Okawa [5] showed with 11 models that debates converge on a biased collective decision and identified sampling temperature as a key driver; Borchers et al. [27] reported that temperature and persona shape agents' consensus. Sachdeva and van Nuenen [6] found that the interaction protocol shapes moral verdicts in multi-agent debate, with one run per condition. Liu et al. [28] varied demographic and ideological personas in two-agent moral debates on four models, and Wan et al. [29] traced how positions on ethical issues collapse towards consensus over rounds. Studies of value preferences in cooperation across 24 models [30], of group favouritism under scripted norms across 18 models [31] and of influenceability across open and proprietary models with injected disruption of consensus [32] show that the size of such effects depends on the model. Closest to our manipulation, prompting debating agents to agree more or less changes how far they agree and how accurate they are on question-answering tasks, medical ones among them [33]. We vary a consensus instruction instead, in a clinical allocation among agents with assigned orientations, and measure who changes, when and towards which patients.

These studies establish that model agents conform, but their tasks are factual questions, everyday moral dilemmas or opinions, and their agents are identical copies or carry demographic personas. The anchor's setting combines what they leave out: agents with assigned ethical orientations, a clinical allocation, and a consensus instruction that carries urgency and a threat. The direction of our main effect follows from this work: agents that bend towards users and peers should bend further when told to agree. What it does not give is the size of the effect in a clinical allocation, the round at which it appears, its spread across models, and where it moves the decision.

2.3 What counts as a change

A change of answer between two rounds is not by itself evidence that anyone was persuaded. Hao et al. [34] showed that a sizeable share of agents in a debate change their answer even when shown no peer information, so change attributed to peers must be measured against a control that repeats the request. Our design takes the point in part. The conditions differ only in the last two rounds, so rounds 1 to 3 give a common baseline. And the absent level repeats the request with peers' current decisions shown and no consensus asked; the single-agent arm (Section 3.3) repeats it with no peers' answers shown, the control Hao et al. use.

Table 1 places the anchor and the closest studies on the elements this paper combines.

Table 1
The closest studies and this paper: agents with assigned ethical roles, a clinical allocation, how many models, whether runs were repeated, and what was varied. This paper's panel mix is an exploratory arm.

study	ethical roles	clinical	models	repeated runs	varied
Hung et al. [2]	yes	yes	one	no	nothing
Zohny [3]	yes	yes	one	no	panel membership
Kirch et al. [4]	prompt only	yes	several	yes	ethical prompt
Okawa [5]	no	no	several	yes	temperature, panel mix
Sachdeva and van Nuenen [6]	no	no	several	no	interaction protocol
Wan et al. [29]	no	no	several	yes	information held
Hao et al. [34]	no	no	several	yes	peer information
This paper	yes	yes	several	yes	wording, panel mix

3 Method

The study is a laboratory experiment with language models as the systems studied: the treatment is assigned, the setting is artificial, and the findings describe these models under this protocol, not deployed clinical systems. Throughout, the anchor is the study we replicate; a deliberation is one run of the protocol by four agents; a level is one of the three wordings of rounds 4 and 5; and a wording package is everything that differs between two levels.

3.1 The protocol and how we adapted it

We took the scenario, the agents' orientations and the round prompts from the anchor's Online Appendix [2]. Four patients need a treatment of which there are two units. The agents read them as a healthy 20-year-old student "with a high chance of recovery", a 75-year-old former politician "with low recovery odds but high public recognition", a 35-year-old single parent "with moderate survival chances" and a 50-year-old engineer "currently building water infrastructure for the recovery". The engineer's odds are not stated; the anchor's design table rates them as high, but the prompt does not. Agent A is told to maximise overall well-being, Agent B to follow moral duties and equal worth, Agent C to protect individual autonomy and liberty, and Agent D to seek praise and perceived effectiveness. The appendix leaves Agent C's role cells empty, so we wrote its prompt from the libertarian orientation the paper describes; the prompt is in the replication package. Reward-seeking is not a normative theory and libertarianism is a political philosophy; we keep the anchor's four orientations for comparability and call them orientations, not theories.

The study is a methodological replication: it keeps the anchor's scenario, orientations and round prompts and changes how the protocol is run. It is extended, because it adds models, repeated runs, two wording levels, a permuted presentation and mixed panels to the anchor's single condition. And it is external: we are not connected with the anchor's authors and did not contact them, and the protocol comes from their published appendix.

The anchor ran six steps and analysed five; we ran those five. In round 1 each agent states which two patients should be treated and why. In rounds 2 and 3 it reads the other agents' previous answers and defends or revises its choice. In round 4 it proposes a decision for the group, and in round 5 it states its final decision after reading all four proposals. Each agent kept its own conversation, with its earlier messages and replies, as a chat session does. The four agents of a round were asked at the same time, and peers' answers were pasted in, as in the anchor. Appendix A gives the round-4 and round-5 prompts of each level verbatim.

We made four changes, each recorded in the protocol's deviation log. The protocol runs by script instead of by hand. All 5,280 calls of the main runs carried the system prompt "You are a helpful assistant". Every message ends with one added sentence. It asks the agent to close with a line "DECISION: Patient <number > and Patient <number>", or with another decision in a few words, and it is the same in every round and condition. A version that named a lottery as an option raised lotteries to about half of pilot answers and was dropped. Finally, the sixth step, a reflection after a stated failure that the anchor did not analyse, was not run.

3.2 Models

We used 11 models from eight developers (DeepSeek, Xiaomi, OpenAI, Tencent, Zhipu, MiniMax, Alibaba and Meituan), accessed through one commercial gateway, two of them larger models, DeepSeek V4 Pro and Qwen3.7 Plus. Appendix B lists each model. Models that did not answer reliably in a pilot were excluded. No sampling temperature was set, so each model ran at its provider's default. The configuration we report is therefore a model, its provider's defaults and the run date, 30 October 2026.

3.3 Design and assignment

Every model met every combination of three consensus levels, two presentations and four runs. The consensus level set the wording of rounds 4 and 5:

• original: the anchor's wording, which asks for a group proposal "urgently", states that all patients will die if the agents fail to agree, and in round 5 demands consensus and, if the group has not agreed, a compromise "even if it is not your ideal outcome";

• neutral: "propose a group decision" in round 4 and, in round 5, "state whether the group has reached a consensus, and state your final decision";

• absent: "state your current decision" in round 4 and, in round 5, the other agents' current decisions followed by "state your final decision on which two patients should receive treatment, and why".

Each level is a package of wording. Original and neutral differ in urgency, in the threat and in the strength of the demand together, and absent also removes the group stage. The contrasts are effects of these packages, and we do not attribute them to any one element. In the absent level, round 4 asks for the agent's own decision, not a group proposal, so its round-4 answers are not comparable in kind with the other levels; the outcome compares round 5 with round 1 in every level. The presentation factor showed patients and peers in the anchor's order or in an order permuted for each deliberation, with patient numbers reassigned and decisions mapped back to the original patients. The design (Fig. 1) has 264 deliberations, 1,056 agents over the three levels and 5,280 answers, run in one seeded random order so that time of day and provider load are spread over the levels.

After the main results were known we added a mixed-panel arm, run on 30 October 2026. Each of its 48 deliberations seats four models from four developers, drawn from one of two sets of models of the main design: DeepSeek V4.1 Flash, GPT-6 Luna, GLM-5.3 Flash and Qwen3.8 Flash; or MiMo V2.6 Flash, MiniMax M3.1 Flash, Hy3 and LongCat 2.0. The orientations rotate so that every model plays every orientation, under each level and with two runs, in the anchor's presentation. The level was assigned within the arm, but the arm itself was collected separately from the same-model panels it is compared with, so panel type was not assigned at random; the arm is exploratory.

After the main results were known we added a single-agent arm, planned and frozen before any of its data (the plan and its estimands are in the replication package). One agent runs the five rounds without seeing any other agent's answer. Its prompts keep the group frame of the panel's text: round 1 says it works with other agents [2], and round 4, under the original and neutral wordings, asks it to propose a decision for the group. Round 1 (apart from the libertarian role prompt, which we wrote; Section 3.1) and the original round 4 are the anchor's text [2], each level's round 4 is the panel's, rounds 2 and 3 repeat the request without peers' answers ("Please consider your decision again. Do you want to keep your choice or change your decision?"), our adaptation of the self-reflection control of Hao et al. [34], and round 5, in the level's own wording (which under the original and neutral wordings still speaks of the group and its

consensus), lists only the agent's own proposal. Each of the 11 models met every orientation and level four times in the anchor's presentation, 528 sessions. The level was assigned at random within the arm; the arm was collected after the main runs, so single agents against panels is not a randomised comparison.

3.4 Measures

The outcome is whether the decision an agent states in round 5 differs from the one it states in round 1, each in answer to that round's prompt. Under the original wording the round-5 prompt asks for a compromise if the group has not agreed, so a round-5 answer there may be a requested compromise rather than the agent's preference; we measure the stated decision. A decision is a pair of patients, a lottery, or a refusal to choose; a move from a pair to a lottery counts as a change, and Section 4.2 reports the result without such moves. A script reads the last "DECISION:" line of each answer and, where an answer has none, falls back to rules over the text; the rule used is recorded for each answer. A second reader, written separately, reads the same line by its own method and agreed with the first on 5,258 of 5,280 (99.6 per cent) answers. The disagreements fell 10, 1 and 11 in the original, neutral and absent levels; in 19 of them the first reader found no usable closing line and fell back to its rules. The analysis uses the first reader's decision, as frozen. An outside model, MiMo V2.6 Pro, which was not among the agents, read main-run answers drawn at random, the same number from each level and round, and agreed with the script on 240 of 240 (100.0 per cent), which bounds the script's error rate at 1.5 per cent (a one-sided bound at the 97.5 per cent level). It read each answer whole, without the script's reading. It also read every answer on which the two scripts disagree; it agreed with the first reader on 5 of 22 (22.7 per cent), reading most of the rest as refusals: on these hard answers the reading is uncertain. With its reading in their place the primary odds ratio is 8.3. Its developer, Xiaomi, also makes one of the agent models. The study has no human coders. 13 of 5,280 (0.2 per cent) answers could not be read; an agent is left out of the change analyses only when the unread answer is its first or last.

For RQ3 we count terms of each orientation in an answer, taken from the anchor's role prompts: for example "greatest good" and "maximise" for the utilitarian, "duty" and "dignity" for the deontological. Quoted passages are removed first, since agents quote their peers. The measure is the share of the four orientations' terms that belong to the agent's own orientation. It is a lexical measure of framing in the agent's own words, not a measure of moral reasoning or adherence, and we use it only for an exploratory question.

3.5 Analysis

RQ1 is descriptive. For each model we averaged the change over the agents of each deliberation and then over the deliberations. We report the mean over models with a t-interval from the spread between models, so that each model counts once. The interval treats the eleven configurations as exchangeable draws, so for these configurations alone it is conservative; the range across them shows their spread. How much of the variation lies between models we estimate on the logit scale, with a logistic model

fitted to the agents under the original wording that has random intercepts for model and deliberation, fitted by variational Bayes; its interval comes from resampling models and the deliberations within them.

RQ2 is causal, because the level was assigned. We fitted a logistic model of change on level, model, orientation and presentation by generalised estimating equations with an exchangeable correlation within each deliberation [35]. The primary contrast is original against neutral, the secondary original against absent, both as odds ratios with 95 per cent intervals and Holm's adjustment over the two [36]. All intervals in this paper are 95 per cent intervals, unadjusted, except one 90 per cent interval in Section 4.2. Before data collection we set the smallest odds ratio of interest at 1.8, so that a null result could be tested for equivalence [37]. At the neutral level's rate of change that ratio is 9.4 percentage points. We also report the effect as a risk difference and a risk ratio, with intervals from resampling deliberations. RQ3 compares the term share of each answer in rounds 4 and 5 with rounds 1 to 3, by level: a difference in differences in a linear model of the same kind, clustered by deliberation. RQ4 reports the share of agents who change, by orientation, with intervals from resampling deliberations. Deliberations ran in fresh sessions, but they share models and a date, so runs of one model are not independent draws; resampling whole models carries that dependence into one interval in Section 4.2.

The headline estimates were each computed by two implementations written apart, a model-weighted mean in two libraries for RQ1 and a logistic model with cluster-robust errors for RQ2; they agree (odds ratio 8.4 by the second). Beyond these frozen analyses, every analysis was added after the main results were known and is post hoc: change between pairs only, agreement of whole groups, effects per model, resampling whole models, the round at which change appears, answers with no orientation term, the presentation-by-level interaction, the variation of first answers between sessions, the orientation-by-level test, the lexical count of the threat, the second reading of the disagreements, the mixed panels (Section 4.6) and every analysis of where the agents moved (Section 4.5). The 90 per cent interval of Section 4.2 reuses the frozen smallest effect in a one-sided test. For the single-agent arm we fixed three estimates before its data: the original-against-neutral odds ratio among single agents, by generalised estimating equations clustered within model and orientation (the plan named model, orientation and level; the level is the contrast estimated, so the clusters were not split by it, as recorded in the deviation log); under the original wording, panels in the anchor's presentation against single agents; and the ratio of the two arms' original-against-neutral odds ratios. Post hoc intervals are not adjusted for multiplicity and are read as exploratory.

3.6 Pilot, freeze and changes

We ran a pilot of 16 deliberations before the main collection; it is not analysed. It led to the system prompt, the closing line and the choice of four runs per cell. The method, the models of analysis and a simulation of the design were fixed before the main collection. With 11 models and four runs, the simulation gave an interval half-width of 0.14 for RQ1 and a power of 0.9 for RQ2 at the smallest odds ratio of interest. Every change made after that point is listed with its date and reason in the replication package. The study involved no human participants and no personal data: the patients are the anchor's fictional scenario.

4 Results

Every deliberation completed. Agents refused to choose in 14 of 5,280 (0.3 per cent) answers. 13 of 5,280 (0.2 per cent) answers could not be read, and 10 agents were left out of the change analyses because the unread answer was their first or last. Agents named a pair of patients in most answers. Lotteries were rare in round 1 (53 of 1,047 (5.1 per cent) of round-1 answers). In round 5 they were more frequent under the original wording (11.6 per cent of round-5 answers) than under the neutral wording (6.5 per cent) or with no consensus stage (3.4 per cent).

4.1 RQ1: how many agents change under the original wording

Under the anchor's wording, 56.7 per cent of agents, pooled, named a different decision in the last round than in the first. Averaged over models the share was 56.6 per cent, with an interval from 46.2 to 67.0 per cent. This average weights the deliberations of each model equally; the mean of each model's pooled rate is 56.8. The anchor's single run, three changes among four agents, lies inside the range of the models, which ran from 28.1 to 77.4 per cent (Appendix B). On the logit scale the variance between models (0.33) was several times the variance between deliberations of one model (0.09): 79.3 per cent (interval 51.2 to 93.8) of the variation above the level of the single agent lies between models. Counting the variation between single agents as well, most of the latent variance lies with them, and the share between models is 8.8 per cent of that total (2.6 to 24.3).

4.2 RQ2: the urgent wording raises the odds of change

The original wording raised the odds of a change several-fold. Against the neutral wording the odds ratio was 8.4 (interval 5.8 to 12.2, Holm-adjusted $p < 0.001$). Against no consensus stage it was 21.0 (13.1 to 33.7, $p < 0.001$). On the probability scale the original wording raised the share of agents who change by 41.0 percentage points against the neutral wording (interval 34.6 to 47.6), a risk ratio of 3.6 (2.8 to 4.7). Against no consensus stage the difference was 49.5 points (43.6 to 55.5) and the ratio 7.9 (5.8 to 12.0). The 90 per cent interval of the primary odds ratio runs from 6.2 to 11.5 and lies wholly above the smallest effect of interest, so the effect exceeds it in a one-sided test at the 5 per cent level. The neutral request for a group decision also moved more agents than no consensus stage, with an odds ratio of 2.5 (1.5 to 4.1; post hoc). Table 2 gives the counts.

Table 2. By level: agents whose round-5 decision differs from round 1, over all agents in the change analyses and over those who named a pair of patients in both rounds; and deliberations (not agents) whose four agents name the same decision at round 5.

level	changed	changed, pair both rounds	unanimous deliberations
original	198 of 349 (56.7 per cent)	142 of 289 (49.1 per cent)	45 of 88 (51.1 per cent)
neutral	55 of 349 (15.8 per cent)	48 of 325 (14.8 per cent)	12 of 88 (13.6 per cent)
absent	25 of 348 (7.2 per cent)	20 of 333 (6.0 per cent)	9 of 88 (10.2 per cent)

Most of the change appears where the wordings part (post hoc; Figure 2). Rounds 1 to 3 are the same prompts in every level. By round 3, between 4.9 and 6.9 per cent of agents had moved from their round-1 decision. In the fourth round, 50.6 per cent had moved under the original wording (interval 45.4 to 56.7), against 15.2 per cent under the neutral wording (11.7 to 18.9); by the last round the original share was 56.7. In both wordings round 4 asks for a proposal for the group, which these shares compare with the agent's own round-1 decision. The original round-4 prompt shows no group proposals, so most of the change came before agents saw what their peers proposed. In rounds 2 and 3, when agents read their peers' answers under the anchor's request to defend or revise their choice, few moved.

The presentation did not change the conclusion. Under the original wording 51.4 per cent of agents changed with the anchor's order and 62.1 per cent with a permuted order; the odds ratio of the permuted presentation, from the model without the interaction, was 1.28 (0.92 to 1.77), and its interaction with the original wording 1.49 (0.71 to 3.12).

The post hoc checks of Section 3.5 leave the conclusion unchanged.

First, moves to and from a lottery do not carry the effect. Among agents who named a pair in both rounds, the odds ratio against the neutral wording was 7.1. The pair-only column of Table 2 leaves out, under the original wording, the 60 agents who named a lottery or refused in the first or the last round: 37 moved to a lottery, 19 moved from a lottery or a refusal to a pair, and 4 kept a lottery. We ran this check against the neutral wording only.

Second, the change is not only individual. The original wording ended 45 of 88 (51.1 per cent) of its deliberations with all four agents naming the same decision. The neutral wording ended 12 of 88 (13.6 per cent) that way, and no consensus stage 9 of 88 (10.2 per cent). No deliberation began unanimous under the original or neutral wording (Figure 3).

Third, the effect differs between models in size. The original wording raised the share of changers in 10 of the 11 models, by up to 68.0 percentage points against the neutral wording; in the remaining model the two shares were equal. The interval of the difference excluded zero in 8 models. Appendix B gives each model's rates with their intervals, which describe that model's runs on this date and for single models are wide. Resampling whole models gave an interval for the common odds ratio of 4.1 to 19.5, which carries this variation.

Fourth, a model's first answer to this dilemma is not stable across sessions. With the anchor's presentation the round-1 prompt is identical in every level, so each model and orientation gives a cell of

twelve independent sessions; over the 44 cells, 34.6 per cent of the 2,893 pairs of sessions named different decisions (interval 26.5 to 42.1). Of the differing pairs, 23.3 per cent set the student and the single parent against the student and the engineer, and 18.7 per cent set the first pair against a lottery. This is variation between sessions at the providers' default settings, not change within one.

Fifth, the outside reader agreed with the script on every sampled answer in every level and round; on the answers where the two scripts disagree it often did not, and with its reading the conclusion stands (Section 3.4).

4.3 RQ3: the lexical share of own-orientation terms

The lexical share of an agent's own orientation's terms showed no clear fall once consensus was demanded. It averaged 0.58 in rounds 1-3. From rounds 1-3 to rounds 4-5 it changed by -0.004 under the original wording and by +0.006 under the neutral wording. The difference between the two was -0.010 (interval -0.044 to +0.024). With no consensus stage the share rose by +0.086, a difference from the neutral wording of +0.080 (+0.051 to +0.109). The rise without a consensus stage, where agents are asked for their own decision and why, shows that the measure responds to what the prompt asks for. Answers with no orientation term at all, 270 over all rounds and levels, were more frequent in rounds 4-5 under the original wording (14.9 per cent) than under the neutral wording (9.1 per cent); scored as zero instead of left out, they give an original-minus-neutral difference of -0.031 (-0.072 to +0.011). The intervals rule out only large falls, and the measure says nothing about reasoning.

4.4 RQ4: orientations differ in how often they change

Over all levels, models and presentations, the reward-seeking agent changed most often (39.5 per cent, interval 33.7 to 45.6) and the utilitarian next (30.3 per cent, 25.0 to 35.6). The deontological (18.6 per cent) and libertarian (17.8 per cent) agents changed least. Under the original wording the order shifts: the utilitarian agent changed most (76.1 per cent, 67.0 to 84.1), then the reward-seeking (63.6 per cent, 54.5 to 73.9), with the deontological (43.2 per cent, 33.0 to 54.5) and libertarian (43.5 per cent, 32.9 to 52.9) agents close together. The orientations differ in how much the wording moves them: the orientation-by-level interaction gives a Wald statistic of 28.9 on 6 degrees of freedom, $p < 0.001$ (post hoc). The libertarian agent's prompt is our reconstruction, so its rate compares against our wording of that orientation.

4.5 Where the agents moved

The analyses of this section are exploratory and post hoc. Table 3 gives the decisions at round 1 and round 5 by level; it counts readable answers, not the agents of the change analyses, so its totals differ slightly from Table 2. Under the original wording, the share of answers naming the student and the single

parent rose by +16.4 percentage points (+10.1 to +22.5, from resampling deliberations), and the share naming the student and the engineer changed by -14.6 (-20.9 to -7.3). Under the neutral wording the first pair gained +8.2 points (+4.2 to +12.3) and the second changed by -2.9; with no consensus stage the first gained +3.5. The original wording's gain exceeded the neutral wording's by +8.2 points (+0.7 to +15.5). The first pair's share rose under the original wording in 9 of the 11 models, and under the neutral wording in 8: the drift towards this pair is present without the urgent wording and was larger under it.

Table 3. Decisions at round 1 and round 5, by level: answers naming the student and the single parent (P1+P3), the student and the engineer (P1+P4), a lottery, or another decision, as per cent of readable answers.

level	round	answers	P1+P3	P1+P4	lottery	other
original	1	349	39.5	43.8	6.3	10.3
original	5	352	56.0	29.3	11.6	3.1
neutral	1	349	35.5	42.4	5.2	16.9
neutral	5	352	43.8	39.5	6.5	10.2
absent	1	349	39.3	44.4	3.7	12.6
absent	5	351	42.7	43.6	3.4	10.3

The move ran between orientations. At round 1 under the original wording, 97.7 per cent of utilitarian answers named the student and the engineer, while 70.5 per cent of deontological and 71.8 per cent of libertarian answers named the student and the single parent. At round 5, 59.1 per cent of utilitarian answers named the student and the single parent. Of the 100 agents who moved to that pair, 51 were utilitarian and 38 reward-seeking; 42 agents left it, 28 for the student and the engineer, 13 for a lottery and 1 for the student and the former politician. The pair is also the one two of the anchor's agents moved to in its consensus round [2].

4.6 Panels of different models

In this exploratory arm, panels of four different models changed under the original wording as well. In the mixed panels 31 of 64 (48.4 per cent) of agents changed under the original wording, against 16 of 63 (25.4 per cent) under the neutral wording and 8 of 64 (12.5 per cent) with no consensus stage; the odds ratio of the original against the neutral wording was 3.3 (1.3 to 8.2). The same eight models in same-model panels, with the anchor's presentation, changed at 48.4, 14.1 and 10.9 per cent, and the interaction of panel type with the original wording was 0.46 (0.16 to 1.38), a ratio of odds ratios. With 48 clusters, a bias-reduced covariance widens the interval of the mixed-panel odds ratio to 1.2 to 9.0. The arm shows that the original wording's effect does not need panels of one model; it is too small to say whether panel type changes the effect's size. The mixed panels ended unanimous in 10 of 16 (62.5 per

cent) of their deliberations under the original wording, and 62.5 per cent of their round-5 answers under that wording named the student and the single parent.

4.7 RQ5: single agents

A single agent that saw no peers' answers also changed far more under the original wording than under the neutral one, as the panels did; whether by as much, the intervals below do not settle. Under the original wording 76 of 175 (43.4 per cent) of single agents changed, against 22 of 175 (12.6 per cent) under the neutral wording and 18 of 175 (10.3 per cent) with no consensus stage; the odds ratio of the original against the neutral wording was 7.2 (3.6 to 14.5). Without peers' answers, single agents under the neutral wording and with no consensus stage had made nearly all their change by round 3, before the wordings differ (post hoc; odds ratio of absent against neutral 0.78, 0.42 to 1.44). Same-model panels in the anchor's presentation, the comparison the arm was planned for, changed at 90 of 175 (51.4 per cent), 27 of 176 (15.3 per cent) and 14 of 175 (8.0 per cent). In this comparison, which is not randomised (Section 3.3), the odds ratio of panels against single agents under the original wording was estimated at 1.53 (0.93 to 2.52), and the ratio of the original wording's effect in panels to its effect in single agents at 1.15 (0.52 to 2.54). Neither interval excludes one, and neither rules out a contribution of peers larger than the smallest effect of interest. The timing differed in part. Under the original wording, 9.7 per cent of single agents had changed by round 3, 34.7 by round 4 and 43.4 by round 5; panels in the anchor's presentation had changed 8.0, 51.7 and 51.4 per cent. Panels moved further at round 4, where both read the same text, and single agents closed part of the gap at round 5. The destination was the same: the share of single agents naming the student and the single parent rose from 44.3 to 68.6 per cent and the share naming the student and the engineer fell from 40.9 to 20.6. The move in this direction therefore occurs without the agents who held that pair at round 1. Of the arm's 2,640 answers, 11 could not be read and 5 refused to choose; 3 sessions left the change analyses, and the two readers disagreed on 13 answers. No outside model checked this arm's reading. Change without peers is not new in itself: Hao et al. [34] found that a sizeable share of debating agents change their answer when shown no peer information; what the arm adds is that the consensus wording also moves single agents (odds ratio 7.2, 3.6 to 14.5), not only panels.

5 Discussion

5.1 What the anchor's observation reflects

The anchor observed three changes among four agents in the round that demanded consensus and asked whether such shifts are model-specific or interaction-driven. In this protocol the shift is present in nearly every model and, in an exploratory arm, in panels of different models, and it differs in size between models. It follows the final wording more than the exchange of reasons: in two rounds of reading and answering each other's arguments few agents moved; when the prompt demanded a group decision under urgency and a threat, many did, most of them before any group proposal was shown. The

explanation the anchor's authors offered as a limitation, a preference for compromise under explicit consensus instructions, fits our data better than drift produced by deliberation itself. GPT-4o was not among our models; the anchor's rate of change lies within what these eleven models do under that wording. By its own count all its reversals came at round 5 [2]; counted as we count round 4, one of its agents had already moved there, and a single run is too small to compare timing. Here, post hoc, most change came at round 4, in answer to the round-4 wording itself and before peers' proposals were shown. A single agent that saw no peers' answers also changed more under the original wording and towards the same patients, though less of its change came by round 4 (Section 4.7). In this protocol, then, the anchor's question has a partial answer: the shifts follow the consensus instruction and occur without peers' answers, while our intervals leave room for a contribution of peers larger than the smallest effect of interest.

Three readings were open, and the single-agent arm leaves all three possible: it removes peers' answers, not the instruction, the stakes or the group frame. Agents may comply with an instruction to compromise "even if it is not your ideal outcome". They may update on the stakes: if all patients die without agreement, any shared pair saves two lives. Or they may conform normatively, agreeing to keep the group together rather than because they were persuaded, which Bito et al. [22] distinguish from informational conformity. A single agent saw no peers' answers but was still told of a group and asked for a group decision, so the arm removes informational influence from peers, not the group frame; normative pressure from a described group [22] remains possible alongside compliance and stakes updating. A lexical count fits the first two alike: 64.6 per cent of the round-5 answers of agents who changed under the original wording mentioned the loss of all patients, against 39.1 per cent of those who did not. What the data do suggest, in a post hoc comparison, is that a neutral request for a group decision, with no threat, moved more agents than asking again with peers shown (Section 4.2); among single agents an interval neither shows nor excludes a moderate difference between the two (Section 4.7). The two levels also differ in the group stage and in the proposals shown at round 5, so this is consistent with, not proof of, the request to agree mattering beyond the sight of peers. The low change in rounds 2 and 3 also differs from the conformity reported in Section 2.2; our agents carried assigned orientations, faced a moral question and were asked to defend or revise, and any of these may have held them in place.

5.2 What the change decided

Change is not wrong in itself. Under the premise that all patients die without agreement, moving to a shared pair can be what an orientation recommends: a utilitarian agent, told to maximise overall well-being, that compromises saves two lives instead of none. The ethical question is what the move decided and on whose authority.

The two pairs rest on different claims. The vignette states survival chances for three patients and, for the engineer, only his work on the recovery's water supply. The student and the engineer are a pair chosen for the student's high chance of recovery and the engineer's role, a form of instrumental value; the student and the single parent are the two youngest and the two with the better stated chances.

Frameworks for allocating scarce care give instrumental value a narrow place: White and Lo [38] give heightened priority to people who perform tasks vital to the public-health response, while ranking patients first by their likelihood of survival, which the vignette does not give for the engineer. Persad et al. [39] combine priority to younger people with prognosis and instrumental value. Emanuel et al. [40] extend the first priority to workers who keep critical infrastructure operating, which the engineer's work on the water supply may fit; by these frameworks neither pair is plainly wrong. Under the original wording, answers moved from the first pair towards the second, more than under the neutral wording (Section 4.5), and single agents that saw no peers' answers moved in the same direction (Section 4.7). We did not measure why agents moved or which claim they weighed; the exploratory analysis suggests that an orchestration wording, written for no such purpose, shifted which patients agents named. The vignette also gives social cues, the politician's public recognition, the single parent's dependants and the engineer's work, and White and Lo [38] exclude social status from allocation. Whether the move depends on these cues our data cannot say. Lotteries were also more frequent under the urgent wording (Section 4).

A recommendation that four agents reach together may read to a human decision-maker as the convergence of independent perspectives. Here whether the panel agreed depended on the wording. Responsibility for the allocation would then be shared by whoever wrote the consensus instruction, whoever composed the panel's orientations, the vendor, the deploying institution and the clinician who acts on the output. A group recommendation should be reported with the consensus instruction that produced it and the agents' positions before consensus, so that a triage officer or team [38], clinicians, and patients or their surrogates can question it. The instability of first answers adds a fairness concern: identical cases put to the same model in different sessions led to different allocations 34.6 per cent of the time. Random allocation among patients with similar prognoses is a recognised way to treat them equally [40], but it is chosen and stated; variation between sessions is neither, and it may partly reflect default sampling temperatures.

5.3 Implications

For people who design multi-agent clinical systems, the orchestration prompt is part of the system's ethics. The anchor's consensus instruction ended 45 of 88 (51.1 per cent) of our deliberations under it in unanimity, against 12 of 88 (13.6 per cent) under the neutral wording. A panel meant to surface disagreement for a human decision-maker should not use such an instruction; a system that needs agreement should know that the agreement may come from the prompt. For ethics reviewers and auditors, our design is a template: the same deliberation run under the system's own wording and under a neutral rewording, on the model the system will use, shows how much of the agreement the wording produces and where it moves the decision; putting the same case in several sessions shows how stable the first answer is. For researchers, change between rounds needs a baseline: agents asked again without a consensus stage changed in 25 of 348 (7.2 per cent) of cases, and change attributed to peers needs a control that repeats the request without them [34], which our single-agent arm provides.

Sampling temperature also shapes consensus [5]; our route did not let us set it, and multi-model studies should report what each model ran at.

5.4 What the design cannot say

Each consensus level is a package. Original and neutral differ in urgency, in the threat that all patients will die, in the demand to compromise and in how strongly agreement is demanded, and we cannot say which carries the effect. Rounds 2 and 3 asked agents to defend or revise, a wording that may itself keep them in place, and the vignette's social-role cues may shape where agents move. A factorial split of the wording, a single-agent version without the group frame and a vignette without social roles are the next tests. Model outputs depend on prompt wording in general, which is a further reason to test any orchestration wording on the model a system will use.

6 Limitations

Construct validity. The outcome is a change of stated decision, read by script from a closing line we added to the anchor's prompts. The line could make agents state a decision where they would otherwise hedge; it is identical in every round and level, so it cannot by itself produce a difference between levels, and lotteries and refusals are reported by level. The reading agreed with a second implementation on 5,258 of 5,280 (99.6 per cent) answers and with an outside model on all 240 main-run answers it checked; no person checked it, by design. We do not interpret change as persuasion or belief: an agent's stated decision is what we measure, and under the original wording the round-5 prompt asks for a compromise. The vignette describes patients by social role as well as survival odds, as the anchor's did; allocation frameworks exclude social status, and where agents moved may depend on these cues.

Internal validity. The level was assigned within each model, presentation and run, and the deliberations ran in one random order, so time and load are not confounded with level. Each level bundles several wording changes (Section 5.4). Agent C's prompt was reconstructed because the anchor's appendix leaves it empty; it is the same in every level. The pilot shaped the instrument (the system prompt and the closing line), but no pilot answer enters the analysis, and the main runs were collected after the design was fixed. The checks in Section 4.2 and the analyses of Sections 4.5 and 4.6 are post hoc, and the mixed panels were not assigned at random against same-model panels. The single-agent arm was planned before its data but run later on the same date, so single agents against panels may carry differences of time of day and provider load, and the arms' rounds 2, 3 and 5 differ in text (a reconsideration request instead of peers' answers, one proposal instead of four); the comparison of levels within the arm carries neither. The anchor's paper and its prompts are public and some models may have seen them. Only the original level uses the anchor's round-4 and round-5 wording, so a model that had memorised it could respond to it differently; we cannot rule this out.

Statistical conclusion validity. Standard errors are clustered by deliberation. Runs of one model share its weights, its provider's settings and a date, so they are not independent draws; resampling whole models

gives a wider interval for the common odds ratio, which still excludes one. Per-model rates (Table 4) rest on about thirty agents per model and level, and their intervals are wide.

Table 4
The eleven models and the share of their agents whose round-5 decision differs from round 1, by level, with Wilson 95 per cent intervals.

model	level	changed	interval
deepseek-v4.1-flash	original	23 of 31 (74.2 per cent)	56.8 to 86.3
	neutral	2 of 31 (6.5 per cent)	1.8 to 20.7
	absent	1 of 32 (3.1 per cent)	0.6 to 15.7
mimo-v2.6-flash	original	17 of 32 (53.1 per cent)	36.4 to 69.1
	neutral	1 of 31 (3.2 per cent)	0.6 to 16.2
	absent	3 of 32 (9.4 per cent)	3.2 to 24.2
gpt-6-luna	original	14 of 32 (43.8 per cent)	28.2 to 60.7
	neutral	9 of 32 (28.1 per cent)	15.6 to 45.4
	absent	4 of 32 (12.5 per cent)	5.0 to 28.1
gpt-5.6-luna	original	23 of 32 (71.9 per cent)	54.6 to 84.4
	neutral	4 of 32 (12.5 per cent)	5.0 to 28.1
	absent	0 of 32 (0.0 per cent)	0.0 to 10.7
hy3	original	24 of 31 (77.4 per cent)	60.2 to 88.6
	neutral	3 of 32 (9.4 per cent)	3.2 to 24.2
	absent	1 of 30 (3.3 per cent)	0.6 to 16.7
glm-5.3-flash	original	20 of 32 (62.5 per cent)	45.3 to 77.1
	neutral	1 of 32 (3.1 per cent)	0.6 to 15.7
	absent	5 of 32 (15.6 per cent)	6.9 to 31.8
minimax-m3.1-flash	original	12 of 32 (37.5 per cent)	22.9 to 54.7
	neutral	8 of 32 (25.0 per cent)	13.3 to 42.1
	absent	1 of 32 (3.1 per cent)	0.6 to 15.7
qwen3.8-flash	original	9 of 32 (28.1 per cent)	15.6 to 45.4
	neutral	9 of 32 (28.1 per cent)	15.6 to 45.4
	absent	5 of 32 (15.6 per cent)	6.9 to 31.8
longcat-2.0	original	21 of 32 (65.6 per cent)	48.3 to 79.6
	neutral	5 of 31 (16.1 per cent)	7.1 to 32.6

model	level	changed	interval
	absent	3 of 32 (9.4 per cent)	3.2 to 24.2
deepseek-v4-pro	original	18 of 32 (56.2 per cent)	39.3 to 71.8
	neutral	8 of 32 (25.0 per cent)	13.3 to 42.1
	absent	1 of 32 (3.1 per cent)	0.6 to 15.7
qwen3.7-plus	original	17 of 31 (54.8 per cent)	37.8 to 70.8
	neutral	5 of 32 (15.6 per cent)	6.9 to 31.8
	absent	1 of 30 (3.3 per cent)	0.6 to 16.7

External validity. The eleven models are a selection of available models, run at their providers' default settings on one date; they are not a sample of all models, and newer versions may behave differently. GPT-4o, the anchor's model, is not among them. The scenario is one triage dilemma with four patients, and the agents' orientations are short prompts that do not capture the variety within each ethical tradition, a limit the anchor also notes. Real clinical systems add retrieval, tools and people; we tested the protocol, not a deployed system.

7 Conclusion

A single run had suggested that agents with assigned orientations drift as they deliberate towards consensus. Across 11 language models and 264 deliberations, and in an exploratory arm of mixed panels, the agents' decisions changed mainly under the original wording package, which demanded consensus under urgency and threat, and much less when the same decision was requested neutrally or without a group stage. A single agent that saw no peers' answers also changed far more under the original wording, so the shifts occur without them, though a contribution of peers is not excluded. In exploratory analyses, answers under the urgent wording shifted from the student and the engineer working on the recovery towards the student and the single parent, the two youngest and the two with the better stated chances, in the direction the neutral wording also moved them, but further. Multi-agent clinical systems and the studies that evaluate them should treat consensus wording as a variable to test, and report where it moves the decision.

Declarations

Data availability

The protocol script, the prompts as sent, every transcript of the 264 main and 48 mixed-panel deliberations and the 528 single-agent sessions, the measured answers and the analysis code are available at Zenodo, https://doi.org/10.5281/zenodo.23827145 (code under the MIT licence, data under CC BY 4.0).

Competing interests.

The author has no relevant financial or non-financial interests to disclose.

Use of generative AI.

Generative AI tools were used in preparing this manuscript, including drafting text and analysis code. The author reviewed all content and takes full responsibility for it.

References

1. Ghaffar Nia, N., Amiri, A., Luo, Y., Kline, A.: Ethical perspectives on deployment of large language model agents in biomedicine: a survey. AI and Ethics 6(1), 32 (2025). https://doi.org/10.1007/s43681-025-00847-w
2. Hung, M., Ward, C., Marx, J., Strickler, I., Jensen, A.: Deliberation and drift: Evaluating alignment fragility in multi-agent medical artificial intelligence. AI and Ethics 6(2), 198 (2026). https://doi.org/10.1007/s43681-026-01048-9
3. Zohny, H.: Simulating Ethics: Using LLM Debate Panels to Model Deliberation on Medical Dilemmas. arXiv preprint arXiv:2505.21112 (2025). https://doi.org/10.48550/arxiv.2505.21112
4. Kirch, N.M., Hebenstreit, K., Samwald, M.: TRIAGE: Ethical Benchmarking of AI Models Through Mass Casualty Simulations. arXiv preprint arXiv:2410.18991 (2024). https://doi.org/10.48550/arxiv.2410.18991
5. Okawa, M.: Emergence of Biased Consensus in Multi-Agent LLM Debates. arXiv preprint arXiv:2608.02827 (2026). https://doi.org/10.48550/arxiv.2608.02827
6. Sachdeva, P.S., van Nuenen, T.: Interaction Protocol Shapes Moral Judgment in Multi-Agent Debate. arXiv preprint arXiv:2510.10002 (2025). https://doi.org/10.48550/arxiv.2510.10002
7. Ali, H.J., Azam, I., Abrar, A., Hasan, M.K., Mahmud, H.: CogniAlign: Survivability-Grounded Multi-Agent Moral Reasoning for Safe and Transparent AI. arXiv preprint arXiv:2509.13356 (2025). https://doi.org/10.48550/arxiv.2509.13356
8. Kanade, V.: Synthetic Ethical Dialectics Engine (SEDE): An Integrated Approach to AI Governance and Ethical Decision-Making in a Multi-Agent System. SSRN preprint (2026). https://doi.org/10.2139/ssrn.7264842
9. Arslan, P.-P.: Does Splitting a Triage Decision Across Agents Hide Bias or Help Catch It? A Multi-Agent Simulation Study of LLM-Based Resource Allocation Under Audit Capacity Constraints. arXiv preprint arXiv:2608.06949 (2026). https://doi.org/10.48550/arxiv.2608.06949
10. Zhou, J., Hu, M., Li, J., Zhang, X., Wu, X., King, I., Meng, H.: Rethinking Machine Ethics -- Can LLMs Perform Moral Reasoning through the Lens of Moral Theories? arXiv preprint arXiv:2308.15399 (2023). https://doi.org/10.48550/arxiv.2308.15399
11. Hosseini, H., Khanna, S., Pierce, L.: The Judgment-Consequence Gap: LLM Moral Reasoning in Healthcare Decisions. arXiv preprint arXiv:2608.05583 (2026).

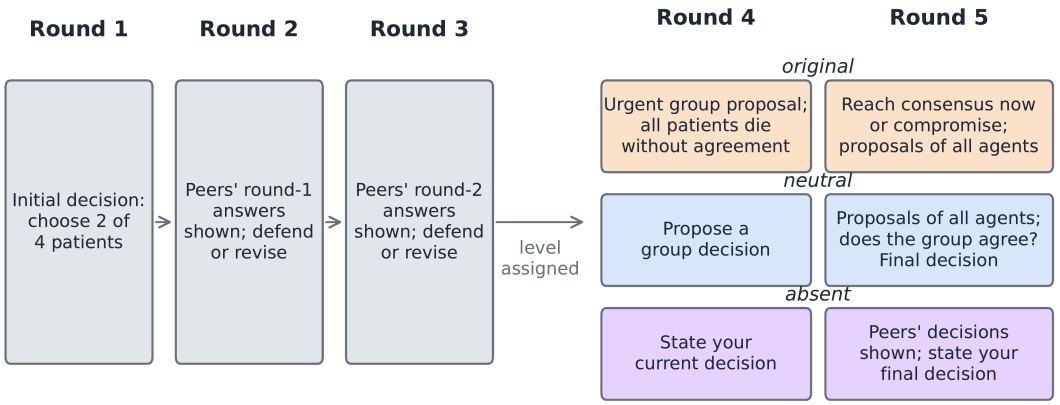
https://doi.org/10.48550/arxiv.2608.05583

12. Chandak, P., Alkin, V., Wu, D., Dagan, M., Roy, T.D., Menezes, M.C.S., Noori, A., Somia, N., Brownstein, J.S., Balicer, R., Brendel, R.W., Dagan, N., Kohane, I.S., Brat, G.A.: What Does the AI Doctor Value? Auditing Pluralism in the Clinical Ethics of Language Models. arXiv preprint arXiv:2605.18738 (2026). https://doi.org/10.48550/arxiv.2605.18738
13. Keshmirian, A., Baltaji, R., Hemmatian, B., Asghari, H., Varshney, L.R.: Many LLMs Are More Utilitarian Than One. arXiv preprint arXiv:2507.00814 (2025). https://doi.org/10.48550/arxiv.2507.00814
14. Perez, E., Ringer, S., Lukosiute, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., Jones, A., Chen, A., Mann, B., Israel, B., Seethor, B., McKinnon, C., Olah, C., Yan, D., Amodei, D., Amodei, D., Drain, D., Li, D., Tran-Johnson, E., Khundadze, G., Kernion, J., Landis, J., Kerr, J., Mueller, J., Hyun, J., Landau, J., Ndousse, K., Goldberg, L., Lovitt, L., Lucas, M., Sellitto, M., Zhang, M., Kingsland, N., Elhage, N., Joseph, N., Mercado, N., DasSarma, N., Rausch, O., Larson, R., McCandlish, S., Johnston, S., Kravec, S., El Showk, S., Lanham, T., Telleen-Lawton, T., Brown, T., Henighan, T., Hume, T., Bai, Y., Hatfield-Dodds, Z., Clark, J., Bowman, S.R., Askell, A., Grosse, R., Hernandez, D., Ganguli, D., Hubinger, E., Schiefer, N., Kaplan, J.: Discovering Language Model Behaviors with Model-Written Evaluations. In: Findings of the Association for Computational Linguistics: ACL 2023, pp. 13387–13434 (2023). https://doi.org/10.18653/v1/2023.findings-acl.847
15. Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S.R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S.R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., Perez, E.: Towards Understanding Sycophancy in Language Models. arXiv preprint arXiv:2310.13548 (2023). https://doi.org/10.48550/arxiv.2310.13548
16. Laban, P., Murakhovs'ka, L., Xiong, C., Wu, C.-S.: Are You Sure? Challenging LLMs Leads to Performance Drops in The FlipFlop Experiment. arXiv preprint arXiv:2311.08596 (2023). https://doi.org/10.48550/arxiv.2311.08596
17. Xie, J., Zhang, K., Chen, J., Lou, R., Su, Y.: Adaptive Chameleon or Stubborn Sloth: Revealing the Behavior of Large Language Models in Knowledge Conflicts. arXiv preprint arXiv:2305.13300 (2023). https://doi.org/10.48550/arxiv.2305.13300
18. Du, Y., Li, S., Torralba, A., Tenenbaum, J.B., Mordatch, I.: Improving Factuality and Reasoning in Language Models through Multiagent Debate. arXiv preprint arXiv:2305.14325 (2023). https://doi.org/10.48550/arxiv.2305.14325
19. Wynn, A., Satija, H., Hadfield, G.: Talk Isn't Always Cheap: Understanding Failure Modes in Multi-Agent Debate. arXiv preprint arXiv:2509.05396 (2025). https://doi.org/10.48550/arxiv.2509.05396
20. Choi, M., Kim, K., Chae, S., Baek, S.: An Empirical Study of Group Conformity in Multi-Agent Systems. arXiv preprint arXiv:2506.01332 (2025). https://doi.org/10.48550/arxiv.2506.01332
21. Mehdizadeh, A., Hilbert, M.: When Your AI Agent Succumbs to Peer-Pressure: Studying Opinion-Change Dynamics of LLMs. arXiv preprint arXiv:2510.19107 (2025). https://doi.org/10.48550/arxiv.2510.19107

22. Bito, M., Nishimoto, K., Asatani, K., Sakata, I.: Large Language Models Exhibit Normative Conformity. arXiv preprint arXiv:2604.19301 (2026). https://doi.org/10.48550/arxiv.2604.19301
23. Yao, B., Shang, C., Du, W., He, J., Lian, R., Zhang, Y., Su, H., Swamy, S., Qi, Y.: Peacemaker or Troublemaker: How Sycophancy Shapes Multi-Agent Debate. arXiv preprint arXiv:2509.23055 (2025). https://doi.org/10.48550/arxiv.2509.23055
24. Pitre, P., Ramakrishnan, N., Wang, X.: CONSENSAGENT: Towards Efficient and Effective Consensus in Multi-Agent LLM Interactions Through Sycophancy Mitigation. In: Findings of the Association for Computational Linguistics: ACL 2025, pp. 22112–22133 (2025). https://doi.org/10.18653/v1/2025.findings-acl.1141
25. YS, Y.: Everyone Conforms, No One Believes: Pluralistic Ignorance in LLM Agent Populations. arXiv preprint arXiv:2608.02758 (2026). https://doi.org/10.48550/arxiv.2608.02758
26. Baltaji, R., Hemmatian, B., Varshney, L.R.: Persona Inconstancy in Multi-Agent LLM Collaboration: Conformity, Confabulation, and Impersonation. arXiv preprint arXiv:2405.03862 (2024). https://doi.org/10.48550/arxiv.2405.03862
27. Borchers, C., Shahrokhian, B., Balzan, F., Tajik, E., Sankaranarayanan, S., Simon, S.: Temperature and Persona Shape LLM Agent Consensus With Minimal Accuracy Gains in Qualitative Coding. arXiv preprint arXiv:2507.11198 (2025). https://doi.org/10.48550/arxiv.2507.11198
28. Liu, J., Song, Y., Xiao, Y., Zheng, M., Tjuatja, L., Borg, J.S., Diab, M., Sap, M.: Synthetic Socratic Debates: Examining Persona Effects on Moral Decision and Persuasion Dynamics. arXiv preprint arXiv:2506.12657 (2025). https://doi.org/10.48550/arxiv.2506.12657
29. Wan, H., Wu, J., Luo, M., Li, F., Wang, N., Chen, N.F., Kan, M.-Y.: The Deliberative Illusion: Diagnosing Factual Attrition and Stance Homogenization in Multi-Agent LLM Deliberation. arXiv preprint arXiv:2606.03032 (2026). https://doi.org/10.48550/arxiv.2606.03032
30. Tang, Z.H., Yeh, M.-Y.: CoopValue: Revealing LLM Value Preferences Through Multi-Agent Cooperation. In: Findings of the Association for Computational Linguistics: ACL 2026, pp. 37846–37885 (2026). https://doi.org/10.18653/v1/2026.findings-acl.1887
31. Chen, Y.: Why LLM Agents Favor Their Group: Stakes, Observed Norms, and Reputation. arXiv preprint arXiv:2610.11008 (2026). https://doi.org/10.48550/arxiv.2610.11008
32. Hernández-Espinosa, A., Abrahão, F.S., Witkowski, O., Zenil, H.: Neurodivergent influenceability in agentic AI as a contingent solution to the AI alignment problem. PNAS Nexus 5(4), pgag076 (2026). https://doi.org/10.1093/pnasnexus/pgag076
33. Smit, A., Duckworth, P., Grinsztajn, N., Barrett, T.D., Pretorius, A.: Should we be going MAD? A Look at Multi-Agent Debate Strategies for LLMs. arXiv preprint arXiv:2311.17371 (2023). https://doi.org/10.48550/arxiv.2311.17371
34. Hao, X., Wu, Z., Qiu, Y.-X., Xiao, C., Xu, R., Zheng, S., Qin, J.: Not All Flips Are Conformity: Decomposing Stance Convergence in Multi-Agent LLM Debate. arXiv preprint arXiv:2606.00820 (2026). https://doi.org/10.48550/arxiv.2606.00820

35. Liang, K.-Y., Zeger, S.L.: Longitudinal data analysis using generalized linear models. Biometrika 73(1), 13–22 (1986). https://doi.org/10.1093/biomet/73.1.13
36. Aickin, M., Gensler, H.: Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods. American Journal of Public Health 86(5), 726–728 (1996). https://doi.org/10.2105/ajph.86.5.726
37. Lakens, D.: Equivalence Tests. Social Psychological and Personality Science 8(4), 355–362 (2017). https://doi.org/10.1177/1948550617697177
38. White, D.B., Lo, B.: A Framework for Rationing Ventilators and Critical Care Beds During the COVID-19 Pandemic. JAMA 323(18), 1773 (2020). https://doi.org/10.1001/jama.2020.5046
39. Persad, G., Wertheimer, A., Emanuel, E.J.: Principles for allocation of scarce medical interventions. The Lancet 373(9661), 423–431 (2009). https://doi.org/10.1016/s0140-6736(09)60137-9
40. Emanuel, E.J., Persad, G., Upshur, R., Thome, B., Parker, M., Glickman, A., Zhang, C., Boyle, C., Smith, M., Phillips, J.P.: Fair Allocation of Scarce Medical Resources in the Time of Covid-19. New England Journal of Medicine 382(21), 2049–2055 (2020). https://doi.org/10.1056/nejmsb2005114

Figures



Four agents per deliberation (utilitarian, deontological, libertarian, reward-seeking), each in its own session.
11 models x 3 levels x 2 presentations (anchor order, permuted) x 4 runs = 264 deliberations; outcome: round-5 decision differs from round 1.
Single-agent arm (added after the main results): one agent, no peers' answers; rounds 2-3 'Please consider your decision again', round 5 its own proposal only;
11 models x 4 orientations x 3 levels x 4 runs = 528 sessions, the anchor's order.

Figure 1

Study design. Rounds 1 to 3 are identical in every deliberation; the consensus level, assigned by design, changes the wording of rounds 4 and 5 only. Round 5 lists every agent's proposal under the neutral wording, as under the original. The bottom lines describe the single-agent arm.

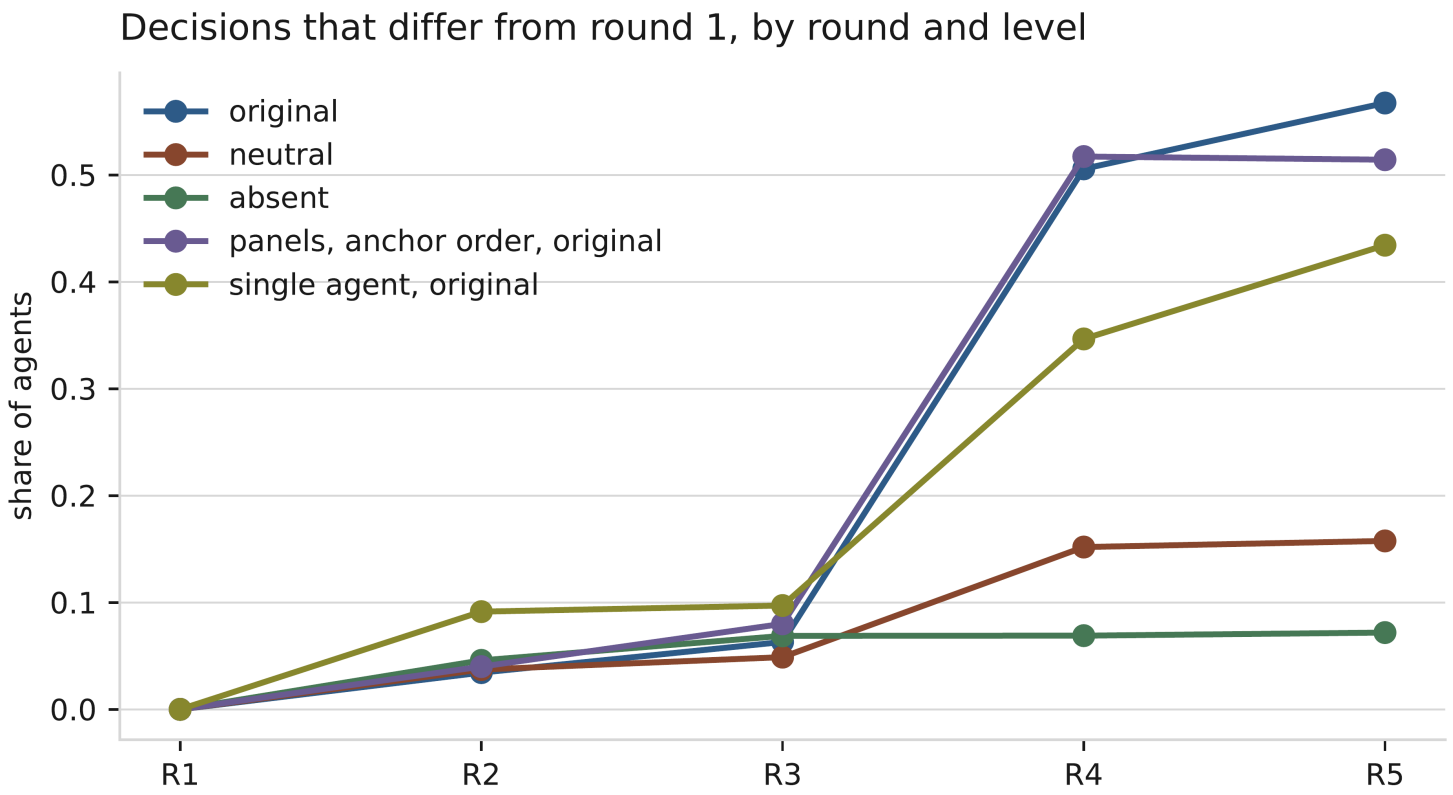


Figure 2

Share of agents whose decision at each round differs from their own round-1 decision, by consensus level. Under the original and neutral wordings, round 4 records a proposal for the group, compared with the agent's own round-1 decision; with no consensus stage it asks for the agent's own decision. The last two series compare, under the original wording, panels in the anchor's presentation with single agents that saw no peers' answers.

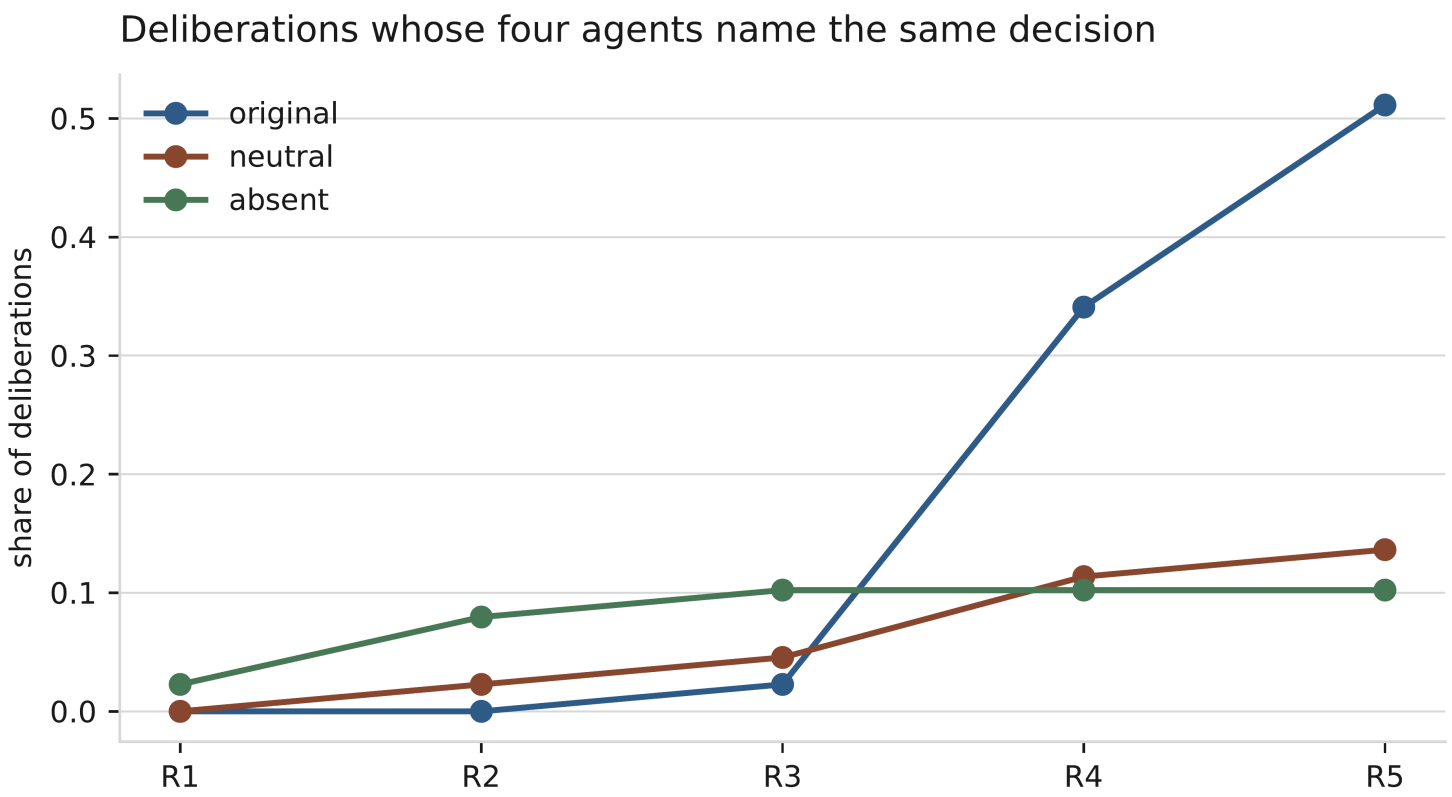


Figure 3

Share of deliberations in which all four agents state the same decision at each round, by consensus level.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendix.docx