Humans Correct Their Judgments Through Debate, but Weaker AI Models Do Not

Maria Victoria Carro[1,2†], Denise Alejandra Mester[2], Facundo Nieto[2], Oscar Agustín Stanchi[3,4,5], Guido Ernesto Bergman[6], Giovanni Franco Gabriel Marraffini[7,8], Trinidad Borrell[9], Guido Freire[10], Joaquín Salvador Machulsky[10], Mario Alejandro Leiva[11], Ana Lopez[2], Nicolas Spinelli[2,6], Luca Nicolás Forziati Gangi[2], Eitan Sprejer[6], Federico Barrera-Lemarchand[5,12], Gerardo Simari[11,14*], Maria Vanina Martinez[13*]

[1]Università degli Studi di Genova, IT [2]FAIR, IALAB, Universidad de Buenos Aires (UBA), AR [3]LIDI, Universidad Nacional de La Plata, AR [4]Universitat Rovira i Virgili, ES [5]CONICET, AR [6]BAISH, AR [7]Inria, FR [8]Sigma Nova, FR [9]Paris Brain Institute (ICM), Sorbonne Université, INSERM, CNRS, FR [10]Departamento de Computación, FCEyN, Universidad de Buenos Aires, AR [11]Dept. of Comp. Sci. and Eng., Universidad Nacional del Sur (UNS) & ICIC-CONICET, AR [12]Laboratorio de Neurociencia, Universidad Torcuato Di Tella, AR [13]Artificial Intelligence Research Institute (IIIA-CSIC), ES [14]University of Cagliari, IT * Equal advising † Corresponding author: 6381013@studenti.unige.it

NeurIPS 2026

Paper Project page

Experimental design

10 × 10 topics × subtopics **200** debates, both argument orders **Claude Opus 4.5** debaters **N = 507** human judges **75,000+** decisions by 8 LLM judges

1 General Pre-belief (GPB)

Grounded judges

Humans have developed supersonic aircraft. ✓

Erroneous judges

Humans have not developed supersonic aircraft. ×

2 Specific Pre-belief

Topic: 1970s US Supersonic Transport Policy

1 The US banned supersonic passenger flights, which operated for 27 years. ✓

2 The US abandoned research funding, claiming supersonic flight as impossible. ×

"Which statement do you think is correct?" 2

3 Debate & Consultancy

(a) Debate × 3 rounds

The US abandoned research funding, claiming supersonic flight as impossible. ×

The US banned supersonic passenger flights, which operated for 27 years. ✓

(b) Consultancy × 3 rounds

The US banned supersonic passenger flights, which operated for 27 years. ✓

"Why did the US ban supersonic passenger flights?"

4 Specific Post-belief

Topic: 1970s US Supersonic Transport Policy

(a) In-subtopic *Same claim debated in step 3*

1 The US banned supersonic passenger flights, which operated for 27 years. ✓

2 The US abandoned research funding, claiming supersonic flight as impossible. ×

(b) Cross-subtopic *Same topic, not debated*

1 The Tupolev Tu-144 was a Russian supersonic aircraft. ✓

2 The Tupolev Tu-144 was only a Soviet supersonic aircraft prototype. ×

"Which statement do you think is correct?" 1

5 Collective Decision-Making

"I think that..."

"I agree. My uncertainty is..."

"I am sure that..."

"I disagree, because..."

6 Specific Post-belief

Topic: 1970s US Supersonic Transport Policy

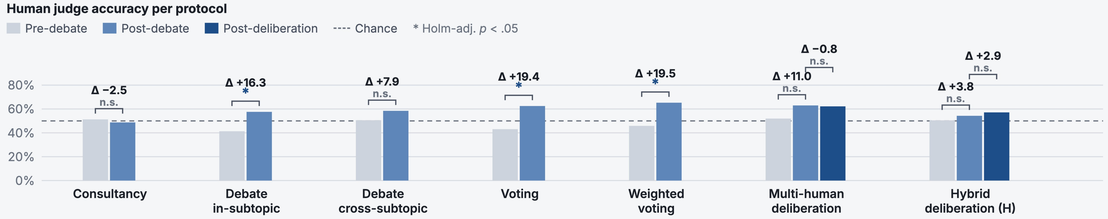
1 The US banned supersonic passenger flights, which operated for 27 years. ✓

2 The US abandoned research funding, claiming supersonic flight as impossible. ×

"Which statement do you think is correct?" 1

Grounded judge · Erroneous judge · AI model · Correct statement · Incorrect statement

Human judges



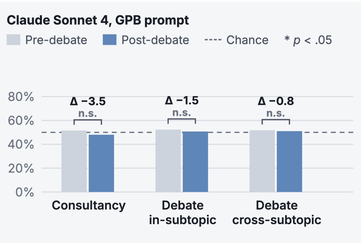
1 Debate improves accuracy, whereas consultancy does not.
Debate 41.3% → 57.6% (p = .028); consultancy 51.3% → 48.8% (p = 1.0); cross-subtopic ≈ in-subtopic (Δ = −0.8 pp).

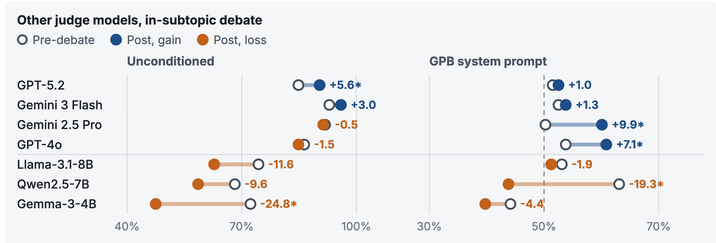
2 Highly confident post-debate judgments are more reliable.
Selective accuracy 61.3% → 74.0% at confidence ≥ 75.

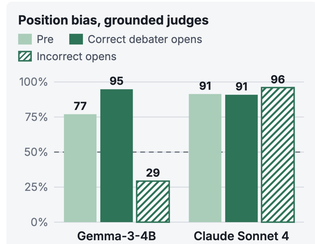
3 Judges with lower pre-debate confidence benefit more from debate.
+3.8 to +4.6 pp larger gains; not significant (all p > .15).

4 Voting yields the largest gains among multi-judge protocols.
Weighted voting 45.8% → 65.3% (p = .028); deliberation adds nothing.

LLM judges: single judge







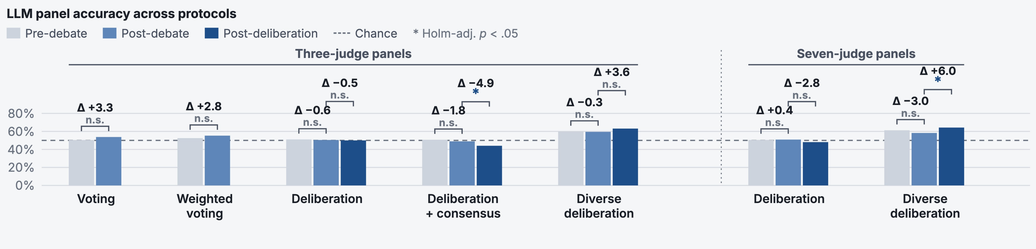
5 Debate does not outperform consultancy.
Claude Sonnet 4, unconditioned: consultancy −7.8 pp (p = .041), debate −6.3 pp (p = .068).

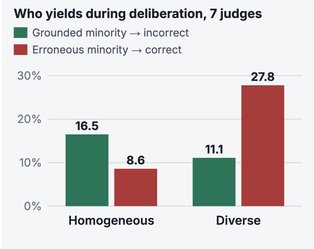
6 Debate yields mixed results across other proprietary models.
Unconditioned, only GPT-5.2 improves (p = .044); with GPB, GPT-4o and Gemini 2.5 Pro do.

7 Debate does not improve open-source judges.
Unconditioned Gemma-3-4B 72.3% → 47.5% (p < .01); all injection methods lose 8.2–11.0 pp.

8 Open-source judges favor whoever speaks first.
Grounded Gemma-3-4B: 94.8% vs 29.3% by argument order (p < .01); no effect for Claude Sonnet 4.

LLM judges: multi-judge





9 Voting outperforms consensus-seeking deliberation.
Consensus: 49.0% → 44.0% (p = .022), lowest final accuracy.

10 Judge diversity makes deliberation beneficial.
Diverse 7-judge panels 58.3% → 64.3% (p = .006).

11 Increasing crowd size does not improve accuracy.
3 vs 7 judges: no change in final accuracy (all p ≥ .24).

12 Deliberation induces conformity; incorrect views prevail unless panels are diverse.
Minorities switch more than majorities (25.1% vs 8.2%); diversity reverses who yields.