Use of large language models in reporting guideline development: Reproduction of ACCORD voting items from systematic review findings

Ellen L. Hughes[1] William T. Gattrell,[2] Keith Goldman,[3] Niall Harrison,[4] Amy Price[5] Paul Blazey[6]

1. Camino Communications, Nottingham, UK; 2. Bristol Myers Squibb, Uxbridge, UK; 3. Global Scientific Publications, AbbVie, North Chicago, IL, USA; 4. OPEN Health Communications, London, UK; 5. Department of Community and Family Medicine (CFMED), Dartmouth-Hitchcock Clinics, Lebanon NH, USA; 6. School of Kinesiology, University of British Columbia, Vancouver, Canada.

16

BACKGROUND

- Consensus methods are widely used in biomedical research, but key methodological steps remain variably reported[1–7]
- Generation of voting items is essential to consensus processes; however, standardised, evidence-based methods are lacking and this step can be prone to subjective judgement and inefficiencies[2–4,8,9]
- Generative AI, particularly LLMs, has shown promise in supporting evidence synthesis tasks[10-12]
- ACCORD-AI evaluates whether LLMs can generate viable draft consensus voting statements from SLR outputs, and how these compare with draft human-generated ACCORD statements[7]

METHODS

Detailed methodology can be accessed via the QR code below

ACCORD SLR outputs[6]

3 independent chats per model; prompts aligned with Anthropic best practice[13]

Claude 3.5 Sonnet

Claude 3.7 Sonnet

DeepSeek R1

Gemini 1.5-pro-002

Gemini 2.0 pro experimental

Llama 3.2

ChatGPT-4.5

ChatGPT-4o

ChatGPT-o1

≤5 iterations to achieve adequate output quality

AI-generated voting statements were evaluated (2 independent assessments) for 1) content relevance, 2) repetition within generated output, 3) novelty vs original ACCORD 56-item longlist

CONCLUSIONS

- LLMs can reliably generate viable draft voting statements for a consensus process
- LLMs can reduce manual effort at the statement generation stage, suggesting they may be best used as pre-consensus accelerators to capture core evidence-based concepts. Choice of model should depend on desired balance of conservative vs novel concept generation
- Human input remains essential for contextual judgement, strategic omissions, interpretation, and final wording, to ensure an overall methodologically robust hybrid AI-human approach
- The models assessed represent a rapidly evolving landscape. Newer AI models designed to support iterative, human-supervised research processes may offer more advanced support for complex consensus processes, including refinement and synthesis across consensus rounds

RESULTS

Fig. 1. Mean number of LLM-generated statements by model

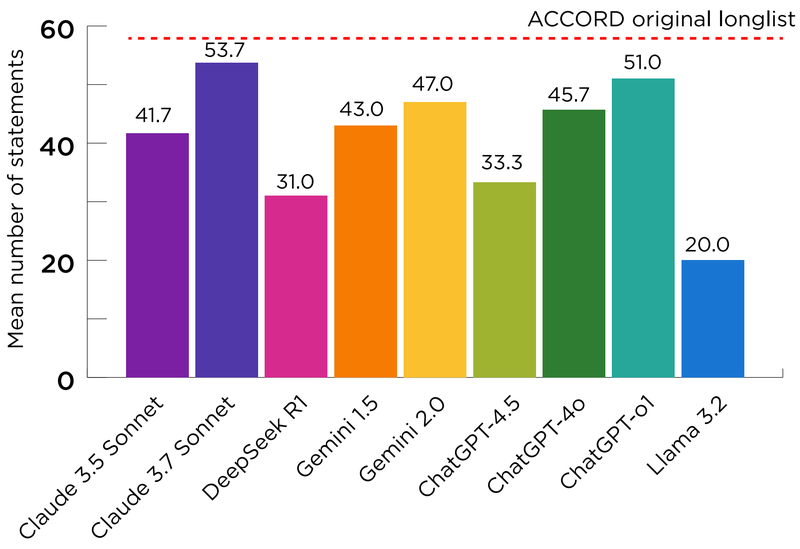
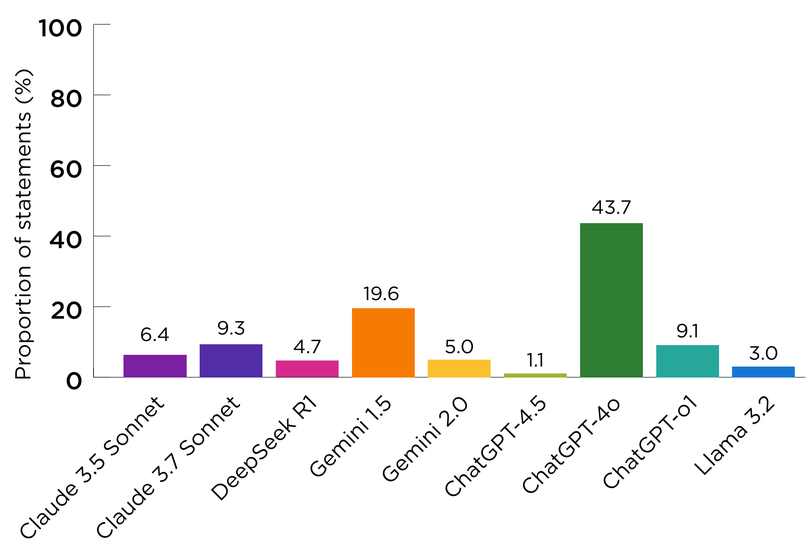


Fig. 2. Mean proportion of LLM-generated statements that were repeated



OUTPUT VOLUME AND EFFICIENCY

- Across nine LLMs, the number of voting statements generated varied substantially despite identical inputs (Fig. 1), indicating variance between LLM models. No LLM model met the 56 items of the original ACCORD longlist, but were generally sufficient to support an initial consensus round and achieved within ≤5 prompt iterations

QUALITATIVE PERFORMANCE

- Most models (7/9) produced a high (>90%) proportion of relevant statements
- Within-list repetition (Fig. 2) and generation of statements that were novel but not relevant (Fig. 3) were generally low but model-dependent. The proportion of statements that were both relevant and novel varied widely (Fig. 3)
- Comparison of the three best-performing models with the original ACCORD longlist showed variable overlap between models and across repeated runs (Fig. 4), indicating differences in how consistently LLMs reproduced established consensus concepts vs introducing novel statements

COMPARATIVE MODEL INSIGHTS

- No single model optimised all performance dimensions
- Differences in output volume and overlap with the original ACCORD longlist appeared to reflect underlying model behaviour rather than prompt instability, with some models favouring concise synthesis and others producing broader concept coverage
- Among those tested, we considered Claude Sonnet 3.7, Gemini-2.0 and ChatGPT-4.5 to show the most balanced performance, while DeepSeek R1 favoured greater novelty

Fig. 3. Mean proportion of novel LLM-generated statements by relevance

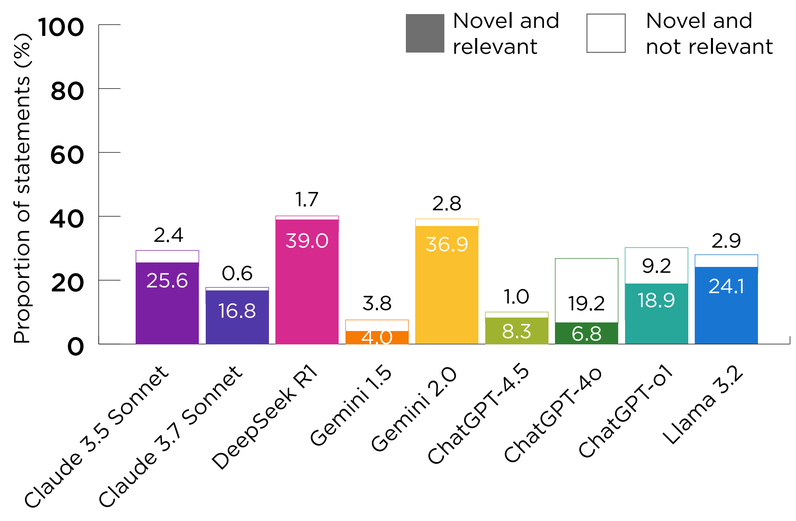
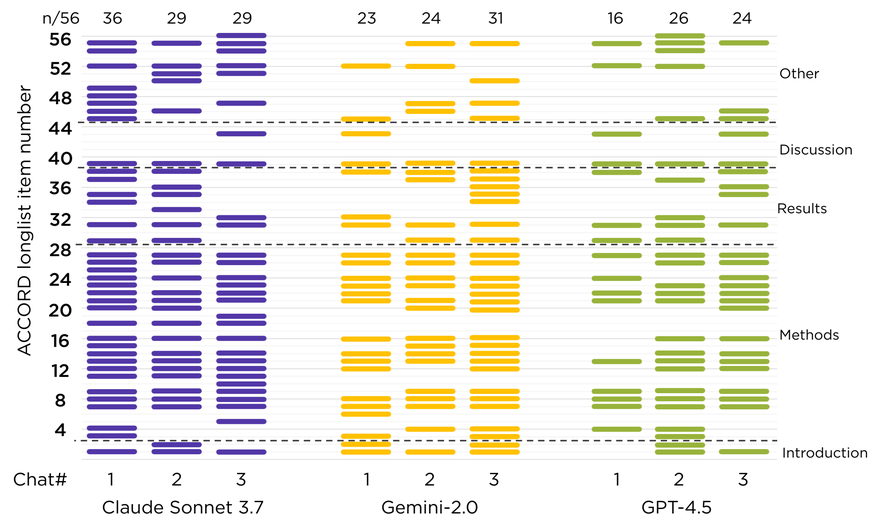


Fig. 4. Comparison of LLM-generated statements overlapping with the ACCORD 56-item longlist



ACCORD, ACcurate COnsensus Reporting Document; AI, artificial intelligence; LLM, large language model; SLR, systematic literature review.

REFERENCES

1. Murphy MK, et al. *Health Technol Assess* 1998;2:i–iv,1–88; **2.** Hasson F, et al. *J Adv Nurs* 2000;32:1008–15; **3.** Diamond IR, et al. *J Clin Epidemiol* 2014;67:401–9; **4.** Jünger S, et al. *Palliat Med* 2017;31:684–706; **5.** Hutchings A, et al. *Qual Saf Health Care* 2006;15:90–95; **6.** van Zuuren EJ, et al. *BMJ Open* 2022;12:e065154; **7.** Gattrell WT, et al. *PLoS Med* 2024;21:e1004326; **8.** Fink A, et al. *N Engl J Med* 1984;310:666–73; **9.** Keeney S, et al. *The Delphi Technique in Nursing and Health Research. Wiley-Blackwell*; 2011; **10.** van Dijk SHB, et al. *BMJ Open* 2023;13:e072254; **11.** Yao X, et al. *Cancer Epidemiol* 2024;88:102511; **12.** Fabiano N, et al. *JCPP Adv* 2024;4:e12234; **13.** Anthropic. Prompt Engineering Available at: https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/overview. Last accessed December 2025.

REFERENCES

Access to paid or professional LLM models was funded by Camino Communications.

DISCLOSURES

EH is an employee of Camino Communications. WTG was an independent medical communications professional at the time of this study and is currently an employee of Bristol Myers Squibb. KG is an employee of AbbVie. NH is an employee of OPEN Health Communications. AP and PB have no potential conflicts of interest to declare. No authors were reimbursed for participating in the initiative.

ACCORD
Accurate Consensus Reporting Document

Poster presented at the 2026 European Meeting of the International Society for Medical Publication Professionals, 26–28 January 2026, London, UK