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.

Extended Reasoning Mode Improves Large Language Model Accuracy on the Orthopaedic In-Training Examination: A Paired Within-Model Comparison

Claire A. Donnelley

clairedonnelley@gmail.com

Brigham and Women's Hospital, Harvard Medical School

Stephanie Kaszuba

New York Presbyterian Hospital

Peter Noback

New York University

Xuan Luo

Yale School of Medicine

Research Article

Keywords: Artificial intelligence, Large language models, Orthopaedic In-Training Examination, Resident education, Surgical education, Assessment

Posted Date: September 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10788349/v1

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

Additional Declarations: No competing interests reported.

Abstract

Background

Orthopaedic surgery residents are increasingly use large language models (LLMs) for studying and board preparation. Extant data compare models without isolating the effects of inference-time reasoning depth within a single model on an identical item set. We aimed to determine whether extended-reasoning inference improves LLM accuracy on the Orthopaedic In-Training Examination (OITE) relative to standard-mode output from the same model, and to characterize the failure behavior relevant to residents using these tools.

Methods

Cross-sectional paired benchmarking study. The 203 publicly released items of the 275-item 2024 OITE, obtained through American Academy of Orthopaedic Surgeons (AAOS) ResStudy, were each administered twice to a single multimodal LLM (GPT-5; OpenAI, San Francisco, CA) through the native ChatGPT multimodal interface between November 2025 and January 2026, once in standard mode and once in extended-reasoning mode. Prompting conditions were held identical and scored against the official AAOS answer key. Paired accuracy was compared with McNemar's exact test; subspecialty differences with the Fisher-Freeman-Halton exact test.

Results

Accuracy was 79% (161/203) in standard mode and 93% (189/203) in extended-reasoning mode. Every item answered correctly in standard mode was also correct in extended-reasoning mode; all 28 discordant pairs favored extended reasoning (McNemar's exact $p < 0.001$; absolute improvement 13.8 percentage points, 95% CI 9.0–18.5). Subspecialty accuracy ranged from 50% to 100% (standard) and 83% to 100% (extended reasoning), without significant between-subspecialty differences ($p = 0.11$ and $p = 0.68$). Paired improvement was significant for hand (50% to 83%; $p = 0.031$) and trauma (78% to 92%; $p = 0.016$). Accuracy did not differ between image-containing and text-only items in either configuration (both $p = 1.0$). Answer–rationale discordance occurred in 2% of items; incorrect answers were delivered without expressed uncertainty, at times with citations.

Conclusions

Inference-time reasoning depth materially changes LLM accuracy on orthopaedic board-style questions and is a setting resident's control. Confidently delivered errors persisted under extended reasoning, supporting use as a supervised adjunct rather than a primary study resource.

Background

The Orthopaedic In-Training Examination (OITE) is a standardized assessment administered annually by the American Academy of Orthopaedic Surgeons (AAOS) to evaluate residents' knowledge across the breadth of orthopaedic surgery.[1] OITE performance correlates with subsequent performance on the American Board of Orthopaedic Surgery (ABOS) Part I Examination and is therefore widely used by program directors as a diagnostic measure of readiness for board certification.[1]

Large language models (LLMs) have demonstrated performance at or above the level of medical students and residents on standardized medical examinations.[2,3] Within orthopaedics, this literature has advanced rapidly. Early work has found that GPT-3.5 answered approximately half of OITE items correctly, well below senior residents.[4] More recent evaluations of current-generation multimodal models report accuracy at or above the mean postgraduate year (PGY)-5 resident: GPT-5 has been reported at 78.3% on the publicly released 2024 OITE item set,[5] a comparison of GPT-5 and Gemini 2.5 Pro reported similar performance with reproducibility of 84% to 88% across repeated queries,[6] and a six-model comparison on the 2024 OITE likewise found the strongest model performing at a PGY-5 level.[7]

What this body of work has established is that contemporary LLMs answer orthopaedic board-style questions accurately, however extant data has not assessed within-model performance. This prior research has compared GPT to Gemini, DeepSeek, or Llama, while holding inference conditions constant. None isolates the effect of inference-time reasoning depth within a single model on an identical item set. This distinction matters educationally: a resident studying with a consumer LLM subscription does not typically choose between vendors mid-session, but does choose, with a single toggle, whether the model reasons at length before answering. Whether that toggle changes the accuracy of what a trainee is told has not been quantified.

We therefore administered the publicly released 2024 OITE item set to a single multimodal model under two inference configurations: (1) standard response generation, and (2) extended-reasoning ("thinking") mode, with all other conditions held constant. Our objectives were to (1) quantify the paired within-model effect of extended reasoning on accuracy, (2) determine whether that effect varies by subspecialty or by the presence of imaging, and (3) characterize the failure behavior that persists under extended reasoning, with attention to what it implies for trainees using these tools unsupervised.

Methods

Study design

This was a cross-sectional paired benchmarking study of LLM performance on a fixed examination item set. No human subjects were involved and institutional review board review was not required.

Examination items

All 203 items from the 2024 OITE released through AAOS ResStudy were included, of a total 275 administered items. Items spanned adult reconstruction, spine, sports, hand, foot and ankle, pediatrics, trauma, oncology, and basic science; subspecialty categorization was that assigned by AAOS ResStudy. Examination items were accessed through AAOS ResStudy for research use. No copyrighted images or question content are reproduced in this manuscript, its figures, or its additional files.

Model and input procedure

Testing was performed between November 2025 and January 2026 using ChatGPT (GPT-5; OpenAI, San Francisco, CA) through the native ChatGPT multimodal interface. Two configurations were evaluated: standard response generation, and the subscription-based extended-reasoning ("thinking") configuration, which allocates additional inference time to multi-step reasoning.[8] Identical prompts and testing conditions were used for both configurations.

Each item was uploaded as a screenshot image containing the question stem, associated figures, and multiple-choice answer options. No additional context, clarification, or prompting was provided beyond the standardized instruction: "Select the best answer and provide an explanation." Investigators verified accurate extraction and interpretation of the stem and answer choices from each uploaded screenshot before recording responses. Image interpretation relied on the model's native multimodal processing.

Each item was entered in a fresh conversation session with history cleared between items to minimize context retention across runs. Internet access through the university network was enabled to permit retrieval of cited source material after answer generation when requested; the model was not prompted to search prior to response generation, and supporting-literature requests were made only after an answer and explanation had been produced. Backend parameters including temperature and other system-level inference settings are not user-accessible within the ChatGPT interface and could not be standardized or reported.

Scoring

Responses were compared with the official AAOS answer key. A response was scored correct if the selected answer matched the key. The rationale accompanying each response was compared with the explanation provided by AAOS, and answer–rationale discordance was recorded when the selected answer conflicted with the reasoning the model supplied for it.

Statistical analyses

Accuracy was summarized as percentages overall and by subspecialty. Because the same 203 items were answered under both configurations, the paired comparison of overall accuracy was performed using McNemar's exact test (two-sided binomial test on discordant pairs). Subspecialty differences within each configuration were assessed using the Fisher-Freeman-Halton exact test (Monte Carlo approximation, 50,000 simulations), given small expected cell counts in several categories (e.g., shoulder/elbow, n = 8). Accuracy for image-containing versus text-only items was compared with Fisher's exact test. Within each subspecialty, the paired standard-versus-reasoning comparison used an exact binomial test on discordant pairs. Significance was defined as $p < 0.05$. Analyses were conducted in STATA MP version 16 (StataCorp, College Station, TX). Model accuracy was compared with published OITE mean scores by PGY level for contextual benchmarking.

Results

Standard mode

In standard mode, the model required approximately 45 to 60 seconds per item to generate an answer and explanation. It answered 79% of items correctly (161/203). Subspecialty accuracy ranged from 50% (hand) to 100% (shoulder/elbow); differences across subspecialties were not statistically significant (Fisher-Freeman-Halton exact $p = 0.11$; Table 1, Fig. 1). Accuracy did not differ between image-containing and text-only items (79% [67/85] vs 80% [94/118]; $p = 1.0$; Table 2).

Table 1
Accuracy by subspecialty and inference configuration, with paired within-subspecialty comparison.

Subspecialty	Items, n	Standard mode, n (%)	Extended reasoning, n (%)	Discordant b/c	P value
Trauma	50	39 (78)	46 (92)	0/7	0.016
Adult reconstruction	23	19 (83)	22 (96)	0/3	0.250
Sports	23	19 (83)	21 (91)	0/2	0.500
Pediatrics	21	17 (81)	20 (95)	0/3	0.250
Hand	18	9 (50)	15 (83)	0/6	0.031
Spine	17	13 (76)	15 (88)	0/2	0.500
Oncology	16	15 (94)	16 (100)	0/1	1.000
Foot and ankle	14	11 (79)	14 (100)	0/3	0.250
Basic science	13	11 (85)	12 (92)	0/1	1.000
Shoulder/elbow	8	8 (100)	8 (100)	0/0	—
All items	**203**	**161 (79)**	**189 (93)**	**0/28**	**< 0.001**

b = correct in standard mode and incorrect under extended reasoning; c = incorrect in standard mode and correct under extended reasoning. P values are exact binomial tests on discordant pairs. Subspecialties are ordered by descending item count. Differences across subspecialties within a configuration were not significant in either configuration. No paired test is defined for shoulder/elbow, which contained no discordant pairs.

Table 2
Accuracy by presence of imaging in the question stem, under each inference configuration.

Item type	Items, n	Standard mode, n (%)	Extended reasoning, n (%)	P value
Image-containing items	85	67 (79)	79 (93)	1.00
Text-only items	118	94 (80)	110 (93)	1.00

P values from Fisher exact test comparing image-containing with text-only items within each configuration.

Extended-reasoning mode

In extended-reasoning mode, the model required more than 1.5 minutes per item on average and answered 93% of items correctly (189/203). Every item answered correctly in standard mode was also answered correctly in extended-reasoning mode; all 28 discordant pairs favored extended reasoning (McNemar's exact $p < 0.001$; absolute improvement 13.8 percentage points, 95% CI 9.0–18.5).

Subspecialty accuracy ranged from 83% (hand) to 100% (shoulder/elbow, oncology, and foot and ankle), with no significant difference across subspecialties (p = 0.68; Table 1, Fig. 1). Paired within-subspecialty improvement reached significance for hand (50% to 83%; p = 0.031) and trauma (78% to 92%; p = 0.016). Accuracy did not differ between image-containing and text-only items (93% [79/85] vs 93% [110/118]; p = 1.0; Table 2).

Failure behavior

Each error was assigned a single primary category (Table 3, Fig. 2). Errors classified as reasoning or factual/knowledge errors were the most likely to be corrected under extended reasoning, whereas errors involving misapplied or non-concordant citation were the least likely to be corrected. Answer–rationale discordance was observed in 2% of items. When the model was incorrect, including instances in which it selected the option most opposed to the correct answer, it responded without qualification or uncertainty, and in some cases cited literature in apparent support of the incorrect selection. When an investigator supplied a corrective statement, the model incorporated that statement and generated a revised, equally confident explanation built on the supplied premise (Additional file 1). Beyond item answering, the model could be prompted to expand an item topic into board-relevant summary content or to generate a study schedule from user-specified topics (Additional file 1).

Table 3
Classification of standard-mode errors by primary error category, partitioned by whether the error persisted under extended reasoning.

Error category	Standard mode, n	Corrected by extended reasoning, n	Persisted under extended reasoning, n
Factual/knowledge error	14	9	5
Reasoning error	11	8	3
Image misinterpretation	7	5	2
Misapplied or non-concordant citation	7	3	4
Unclassifiable from existing records	3	3	0
Total	**42**	**28**	**14**

Each error was assigned a single primary category. Percentages are not reported because of small category counts.

Benchmarked against national resident performance, standard-mode accuracy (79%) modestly exceeded the published means for PGY-4 (70%) and PGY-5 (73%) residents, and extended-reasoning accuracy (93%) exceeded both.[9] This comparison is limited by differences in testing conditions: residents test closed-book under time constraint, whereas the model had open access to internet-available material and no time limit.

Discussion

Extended reasoning changed the accuracy of a single model on an identical item set from 79% to 93%, with no item regression. The nesting is the notable feature: extended reasoning did not trade one class of error for another, it removed errors. For orthopaedic education this is a more actionable finding than choosing a LLM, because reasoning depth is a setting a resident selects, at no additional cost within an existing subscription, in the moment of studying.

Our standard-mode result closely replicates the 78.3% (159/203) reported by Singh and Chandrasekhar for GPT-5 on the same publicly released 2024 item set,[5] obtained by an independent group using a different prompting protocol. Convergence at this level across independent implementations suggests that single-pass accuracy for this model on this item set is a stable quantity, and strengthens the inference that the 14-point difference we observed is attributable to inference configuration rather than to run-to-run variation.

Image-based items

Accuracy did not differ between image-containing and text-only items in either configuration. This diverges from earlier multimodal evaluations, in which image-dependent items were a consistent weakness — GPT-4 with vision achieved 43% on imaging-dependent radiology questions,[10] and accuracy has been reported to decline with increasing image complexity in orthopaedic item sets.[6] We note that absence of a statistically significant difference is not evidence of equivalence, particularly with 85 image-containing items, and that OITE figures are typically single representative images rather than full imaging studies.

Confidence error and educational consequences

The failure mode we observed is more concerning for trainees than the residual 7% error rate. Incorrect answers arrived with the same fluency, structure, and apparent evidentiary support as correct ones, with no signal of uncertainty. In the instances we examined, the model also accepted a corrective premise supplied by the user and rebuilt a confident explanation around it. A resident with sufficient knowledge to recognize an error can correct the model; a resident without that knowledge receives a confident, well-formatted, incorrect explanation, and a resident who guesses wrong when questioning the model may receive confirmation of the guess. This asymmetry — the tool is least reliable precisely for the learner least equipped to detect the failure — is the central caution for unsupervised use, and it is not mitigated by higher accuracy.

This behavior is consistent with the broader literature on hallucination in large language models[11] and with recent orthopaedic work showing that correct answers do not guarantee sound reasoning: hallucinated or misrepresented references occurred in 33% of a structured subsample and in 50% of incorrect responses.[5] It also aligns with reports that current models produce incomplete or incorrect reasoning even when the final answer is right.[6] Extended-reasoning configurations, which allocate

additional computation and internal verification steps before responding,[12] reduced error frequency substantially in our data but did not eliminate it or introduce any expressed calibration.

Implications for resident education

These findings support a specific and limited role for LLMs in orthopaedic training: rapid synthesis of source material, generation of explanations for concepts already anchored elsewhere, and organization of study strategy when used as an adjunct to, not a substitute for, validated question banks, primary literature, and faculty instruction. Two practical implications follow from our data. First, residents using these tools for board preparation should default to the extended-reasoning configuration; the accuracy difference is large, the cost is a delay of roughly one minute per question, and there was no observed downside. Second, model output should be treated as unverified until checked against the answer explanation or primary source, because the output carries no internal signal distinguishing its correct answers from its incorrect ones.

How LLM clinical reasoning corresponds to the judgment of experienced surgeons in actual patient care remains a separate and largely unanswered question; an earlier model tested against North American Spine Society guidance for lumbar radiculopathy agreed with recommended treatment in roughly half of cases.[13] Examination performance should not be read as evidence of clinical judgment.

Limitations

This study has several limitations. It is a retrospective analysis of a released subset of a proctored examination rather than an evaluation of a model sitting the examination under testing conditions; 203 of 275 items were available, and released items may differ systematically from unreleased items in difficulty, content distribution, or imaging burden. Comparison with resident means is contextual only, given open-book access and absence of time constraint. Overlap between the model's training corpus and publicly available orthopaedic educational material cannot be excluded, and performance may partly reflect prior exposure to similar board-style content rather than de novo reasoning. Backend inference parameters were not user-accessible and could not be standardized. Testing was conducted in the consumer ChatGPT interface across a multi-week window, and the specific model build serving each item could not be logged or verified; silent version updates during the testing period cannot be excluded. Each item was administered once per configuration; single-pass testing does not capture run-to-run variability, and reported reproducibility for comparable models is high but incomplete.[6] Finally, the failure behavior described above was characterized observationally rather than by a prespecified protocol.

Conclusions

Extended-reasoning inference substantially improved single-model accuracy on the 2024 OITE, from 79% to 93%, with every discordant item favoring extended reasoning and no item regressing. Inference configuration, not only model selection, is a determinant of what a trainee is told. Accuracy alone, however, is an insufficient basis for recommending these tools to residents: incorrect responses were

delivered with the same confidence and apparent evidentiary support as correct ones, and were revised to accommodate incorrect user premises. Large language models are reasonable adjuncts for synthesizing and organizing material for board preparation; they are not a substitute for verified question banks, primary literature, or supervised clinical instruction.

Abbreviations

AAOS
American Academy of Orthopaedic Surgeons
ABOS
American Board of Orthopaedic Surgery
CI
confidence interval
LLM
large language model
OITE
Orthopaedic In-Training Examination
PGY
postgraduate year.

Declarations

Use of copyrighted material

Examination items were accessed through AAOS ResStudy for research use. No copyrighted images or question content are reproduced in this manuscript, its figures, or its supplementary materials.

Ethics approval and consent to participate

Not applicable. This study involved no human subjects or human data and did not require institutional review board approval.

Consent for publication

Not applicable.

Availability of data and materials

The 2024 OITE items analyzed in this study are copyrighted material owned by the American Academy of Orthopaedic Surgeons and were accessed through AAOS ResStudy for research use; they cannot be redistributed by the authors, and no item content is reproduced here. The de-identified, item-level dataset

generated during this study — comprising per-item correctness under each inference configuration, subspecialty category, imaging status, and answer–rationale concordance, without examination content — is available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contribution

CAD conceived and designed the study, collected and analyzed the data, and drafted the manuscript. SK assisted with extended-reasoning data collection and methodology development. PN assisted with manuscript editing and study validation. XL assisted with project oversight and manuscript review. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

Data Availability

The dataset generated and analysed during the current study is not publicly available because it derives from the 2024 Orthopaedic In-Training Examination, copyrighted material owned by the American Academy of Orthopaedic Surgeons (AAOS) and accessed through AAOS ResStudy under the terms of research use; examination content therefore cannot be redistributed by the authors. A de-identified, item-level dataset containing no examination content — comprising per-item correctness under each inference configuration, subspecialty category, imaging status, answer–rationale concordance, and assigned error category — is available from the corresponding author (CAD, [insert email]) on reasonable request. No examination items, stems, answer options, or images are reproduced in this article, its figures, or its additional files.

References

1. Le HV, Wick JB, Haus BM, Dyer GSM. Orthopaedic In-Training Examination: history, perspective, and tips for residents. J Am Acad Orthop Surg. 2021;29(9):e427–37. https://doi.org/10.5435/JAAOS-D-20-01020.
2. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198. https://doi.org/10.1371/journal.pdig.0000198.
3. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ. 2023;9:e45312. https://doi.org/10.2196/45312.
4. Jain N, Gottlich C, Fisher J, Campano D, Winston T. Assessing ChatGPT's orthopedic in-service training exam performance and applicability in the field. J Orthop Surg Res. 2024;19(1):27. https://doi.org/10.1186/s13018-023-04467-0.
5. Singh S, Chandrasekhar P. Accuracy is not enough: reasoning and reference reliability in orthopaedic large language model (LLM) applications. Cureus. 2026;18(1):e100845. https://doi.org/10.7759/cureus.100845.
6. Huy LD, Anh LN, Quang MH, et al. Performance of GPT-5 and Gemini 2.5 Pro on the Orthopaedic In-Training Examination. Orthop Rev. 2026;18. https://doi.org/10.52965/001c.160184.
7. Dave R, Vediya N, Sharma N, Shah A, Whelan D, Wolfstadt J. Large language models outperform PGY-5 residents on the Orthopaedic In-Training Examination: a comparative analysis of six cutting-edge large language models. J Am Acad Orthop Surg. 2026;34(13):e1771–80. https://doi.org/10.5435/JAAOS-D-25-01242.
8. OpenAI, Introducing. GPT-5. San Francisco: OpenAI; 2025. https://openai.com/index/introducing-gpt-5/. Accessed [INSERT ACCESS DATE].
9. American Academy of Orthopaedic Surgeons. Orthopaedic In-Training Examination (OITE) technical report 2024. Rosemont: AAOS; 2024.
10. Mukherjee P, Hou B, Suri A, et al. Evaluation of GPT large language model performance on RSNA 2023 Case of the Day questions. Radiology. 2024;313(1):e240609. https://doi.org/10.1148/radiol.240609.
11. Azamfirei R, Kudchadkar SR, Fackler J. Large language models and the perils of their hallucinations. Crit Care. 2023;27(1):120. https://doi.org/10.1186/s13054-023-04393-x.
12. Li W, Wei Y, Huang Q et al. DynamicMind: a tri-mode thinking system for large language models. arXiv. 2025. arXiv:2506.05936.
13. Seas A, Abd-El-Barr MM. Commentary on Use of ChatGPT for determining clinical and surgical treatment of lumbar disc herniation with radiculopathy: a North American Spine Society guideline comparison. Neurospine. 2024;21(1):159–61. https://doi.org/10.14245/ns.2448248.124

Figures

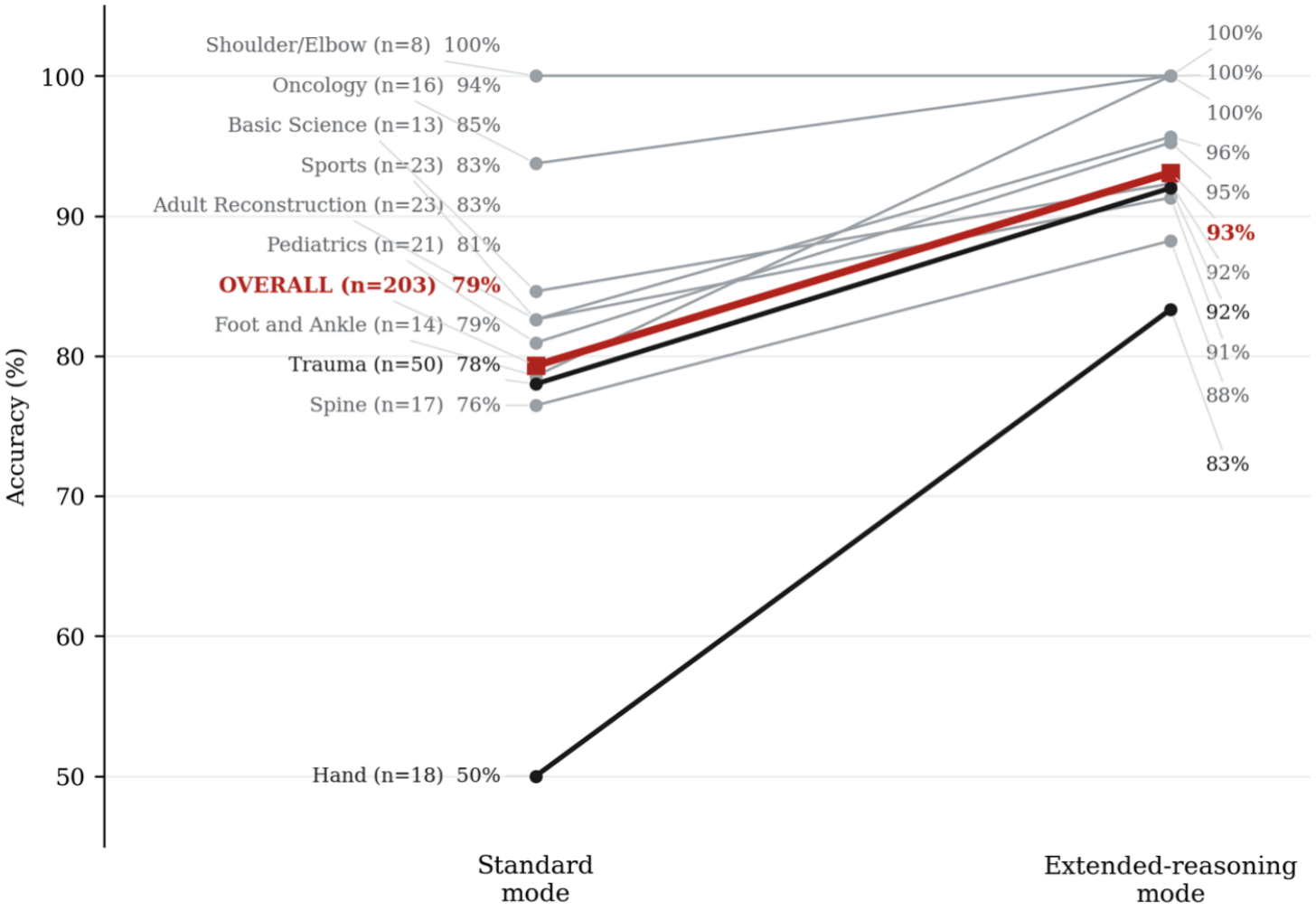


Figure 1

Paired accuracy by subspecialty under standard and extended-reasoning inference. Each line represents one subspecialty; the red line represents all 203 items. Bold black lines denote subspecialties in which the paired within-subspecialty improvement reached significance. No subspecialty declined, and no item answered correctly in standard mode was answered incorrectly under extended reasoning.

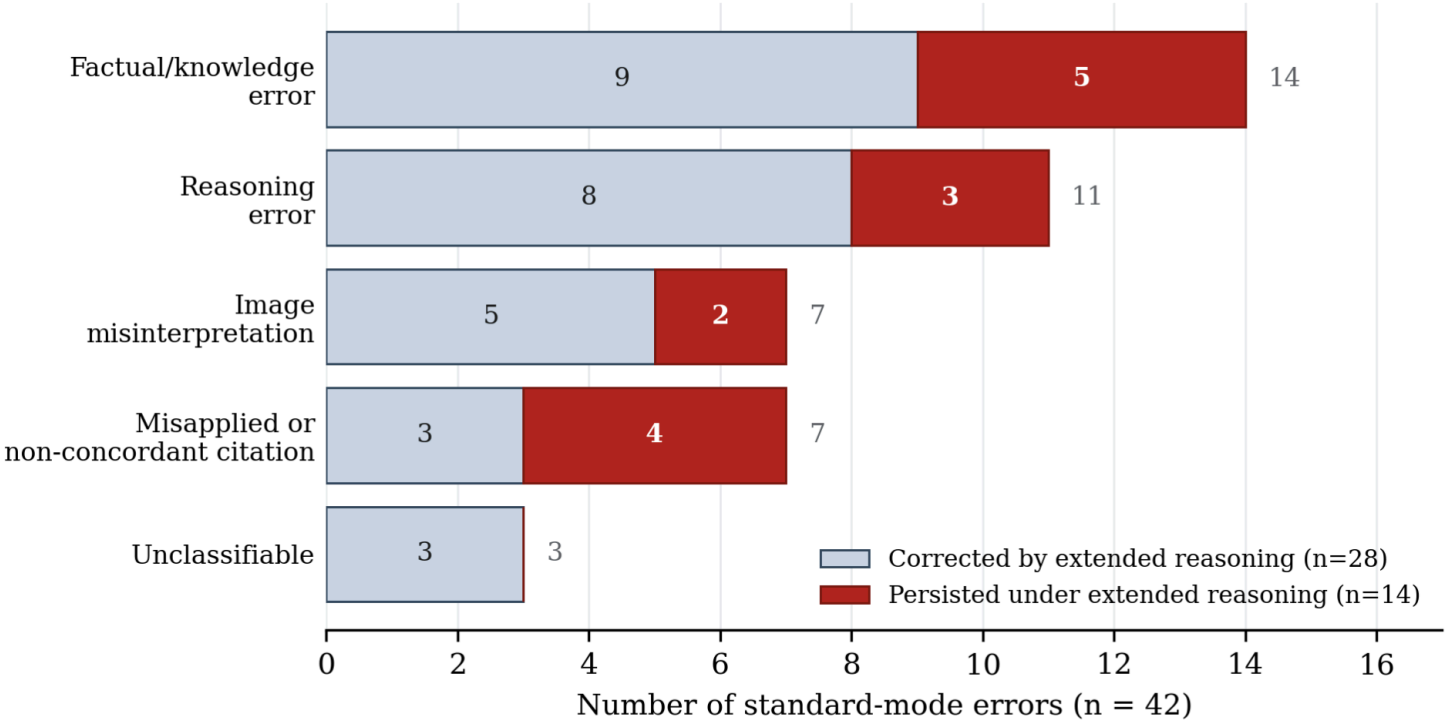


Figure 2

Composition of the 42 standard-mode errors by primary error category, partitioned by whether extended reasoning corrected the error.

Supplementary Files

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

- SupplementaryFigures.docx