

Accuracy of AI-Generated Procedural Descriptions in Peripheral Nerve Surgery: Implications for Surgical Training

Brandon Chai

The University of British Columbia Faculty of Medicine <https://orcid.org/0000-0002-6758-7913>

Megan M. Yim

myim@student.ubc.ca

The University of British Columbia Faculty of Medicine <https://orcid.org/0009-0000-3415-9827>

Damon Ranjbar

The University of British Columbia Faculty of Medicine

David T. Tang

The University of British Columbia Faculty of Medicine

Research Article

Keywords: Artificial intelligence, ChatGPT, large language models, surgical education, peripheral nerve surgery, resident training

Posted Date: September 25th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10198678/v1>

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

[Read Full License](#)

Abstract

Purpose

Self-directed learning is increasingly important in surgical education. Large language models may support this, however, their accuracy, reliability, and sensitivity to model generation and prompting strategy remain unclear. This study evaluated AI-generated procedural descriptions in peripheral nerve surgery using GPT-5.2, and tested robustness using a newer model, GPT-5.5.

Methods

Fifteen peripheral nerve procedures were randomly selected from the Washington University in St. Louis Surgical Education Website as reference standards and coded into steps. GPT-5.2 outputs were evaluated for accuracy and other response characteristics using a structured prompt. Regression assessed accuracy against complexity, word count, steps, and step density. For five procedures, GPT-5.5 was evaluated using structured and simple prompts.

Results

Mean GPT-5.2 accuracy was 81.2% (SD 11.5%) across 90 responses, ranging from 68.5% to 91.7%. Accuracy was positively associated with word count and negatively associated with step density, though neither remained independent predictors. Of 195 missed steps, most involved judgement-based maneuvers (86/195), major steps (56/195), or neurovascular identification (41/195). GPT-5.5 accuracy (81.4%, SD 16.2%) resembled GPT-5.2's but was more variable, decreasing to 40.5% (SD 40.2%) under a simple prompt, shifting omissions towards major steps.

Conclusion

GPT-5.2 and GPT-5.5 generated accurate but variable responses, inconsistently capturing judgement-dependent content. Simple prompting substantially reduced accuracy and shifted omissions toward major steps, indicating dependence on prompting strategy. AI-generated descriptions may aid orientation, but should not replace expert guidance for safety-critical and judgement-required steps.

Background

Surgical training is well recognized for intensity and rigor, requiring substantial knowledge acquisition and skills in a limited timeframe [1]. As such, efficient learning strategies are essential for surgical learners to meet their learning needs while simultaneously participating in their clinical responsibilities. This has become particularly important since the shift towards competency based medical education (CBME), as CBME residents responding to a recent national Canadian survey have reported being more

proactive in their self-directed learning and expressed concerns over whether CBME will accommodate individual learning needs and variability in learning pace [2]. These sentiments highlight how effective and efficient strategies for self-directed surgical education are critical to support the shift towards CBME.

The integration of large language models (LLMs) using artificial intelligence (AI) such as ChatGPT (OpenAI, San Francisco, CA, USA) to enhance medical education is rapidly evolving [3–7]. GPT-5.2 is one of the most recent multimodal language models of ChatGPT, which demonstrates powerful information integration abilities and can search the latest published literature to provide up-to-date information in response to search queries [8]. Recent studies have demonstrated variable accuracy and reliability of responses to medical queries, although generally report accurate responses limited by the complexity of the query [9–11]. A meta-analysis by Wei et al (2024) suggested that the variety between studies' methodology contributes to the variable accuracy in ChatGPT's ability to address medical queries [12]. As the capabilities of OpenAI's ChatGPT are continually advancing, its role in enhancing medical education will only increase with greater response accuracy and comprehensiveness.

Prompt engineering is an emerging skill which involves the strategic design and optimization of prompts to guide the LLM responses towards a relevant output [13, 14]. This skill requires an understanding of how LLMs operate, and involves carefully crafting phrases, wording, and context of prompts to optimize outputs; on the contrary, simple prompting fails to take full advantage of LLM capabilities, resulting in less relevant or accurate outputs [13, 14]. One prior study used prompt engineering to compare the accuracy in surgical steps for a Fisher cleft lip repair as written by a board-certified CMF plastic surgeon versus the output from ChatGPT 3.5 and found that it required significant prompting to produce organized, easily readable procedural steps; however, responses contained inaccuracies and were inferior to those written by the surgeon [15]. To the best of our knowledge, there are no studies that evaluate the capability of ChatGPT's newest model (GPT-5.2) for providing accurate descriptions of surgical steps for common plastic surgery procedures catered to the level of a medical student and plastic surgery resident. This study therefore aimed to apply a reproducible, step-based framework for evaluating AI-generated surgical procedural descriptions by assessing procedural accuracy, completeness, response reproducibility, and the nature and clinical relevance of omitted or inaccurate content. GPT-5.2 was used as the primary model case study. To assess the robustness of the framework to technological evolution and prompting strategy, a sensitivity analysis was performed using GPT-5.5, and GPT-5.5 responses generated using a structured, prompt-engineered query were compared with those generated using a simple, user-style prompt.

Methods

Study Design

The Washington University in St. Louis Surgical Education Website is a resource that offers video instruction from expert surgeons, tailored for plastic surgery residents and practicing surgeons alike [16].

This resource has a large repository of high-quality videos describing the surgical steps for many procedures, particularly in peripheral nerve surgery. A list of 15 random peripheral nerve procedures was created from the website (Online Resource 1). The latest and most advanced LLM released by OpenAI at the time of data collection (GPT-5.2 released in December 2025) was used for the primary analysis. Each procedure was inputted into ChatGPT using GPT-5.2 with a prompt designed for plastic surgery residents which can be found in Online Resource 2. This prompt used prompt engineering to optimize the output of the LLM. Prompt engineering was achieved by asking the LLM to describe the surgical procedure including specific details such as major procedural steps, technical considerations that may vary between surgeons, key landmarks, critical structures and instruments (Online Resource 2). There were no limitations on word count or reading level to provide a more accurate assessment of the performance of GPT-5.2 for providing comprehensive surgical education. A zero-shot mode was used, meaning that no example answers were provided to each LLM as a reference. Additionally, the chat session was refreshed before each question, preventing the LLM from using previous responses as a basis for subsequent responses. GPT-5.2 was accessed through the ChatGPT web interface in an incognito browser window. All settings, including web browsing/search tools, custom instructions, and memory, were left at their default configuration and were not manually enabled or disabled; decoding parameters such as temperatures were similarly left unmodified. This process was performed for 15 peripheral nerve procedures, with each procedure repeated 6 times to evaluate reliability. All 15 prompts were entered into ChatGPT-5.2 between December 18–24, 2025.

A reference standard was developed for each procedure using the textual information and YouTube videos available on the Washington University Surgical Education Website [17]. Each video was transcribed using YouTube's auto-generated captions, which were then manually reviewed and corrected for grammatical accuracy by two independent reviewers. Transcripts were coded into distinct broad procedural steps using an inductive approach, in which reviewers independently identified and labeled each step based on the narrative flow and surgical actions described. These steps were finalized through review by a board-certified plastic surgeon to ensure accuracy (Online Resource 3).

Data Collection

Reference Standard and GPT Response Characteristics

Descriptive information including total word count, number of paragraphs, section headers, and readability was evaluated for each GPT response. The number of procedural steps was defined a priori by the corresponding reference standard and therefore remained constant across repeated responses for a given procedure. To account for differences in procedural scope and response length, step density was calculated for each GPT response as the number of reference-standard procedural steps divided by the response word count and multiplied by 100, representing the number of reference-standard steps per 100 response words. Higher step density therefore reflected fewer words generated relative to the number of procedural steps required by the reference standard. Readability was evaluated using the Flesch-Kincaid Grade Level (FKGL), Gunning Fog Index (GFI) and Coleman-Liau Index (CLI) using a

validated online calculator (<https://readable.com/>) [18, 19]. These scores have been used in prior studies to assess readability [20, 21]. The FKGL, GFI and CLI all indicate the educational level required to understand the text. Both the FKGL and GFI are calculated using the average sentence length and number of syllables per word. The CLI uses the average number of letters per word and average number of sentences per 100 words to evaluate readability.

GPT Assessment

Two independent reviewers evaluated all 15 ChatGPT outputs against the reference standard. Each GPT-5.2 response was compared against the reference standard using two approaches. First, response accuracy was quantitatively evaluated by calculating the proportion of GPT-5.2 content that matches the reference standard. The second was a qualitative analysis, where the responses were evaluated for incorrect, missing or incomplete steps to identify recurring content deficiencies. Two reviewers independently coded each missed step into themes. Discrepancies in coding were resolved via a third author, and through multiple rounds of discussion, re-reading and iterative review of the coded steps. Responses generated using the newer model release and the simple user-style prompt were evaluated using these same accuracy and qualitative coding methods.

There is currently no standardized framework for evaluating the complexity of peripheral nerve procedures. To address this gap, a scoring approach was adapted from the Aristotle Basic Complexity score, which is a system widely used in other surgical subspecialties to quantify operative difficulty [22]. Given the distinct anatomical and technical considerations of peripheral nerve surgery, a set of six domains relevant to this context was selected and refined: (1) risk of morbidity, (2) technical difficulty, (3) typical operative time, (4) nerve reconstructive components (e.g., tendon or nerve transfer), (5) multi-tissue reconstructive components and (6) level of training required to perform the procedure independently (Online Resource 4). Each domain was scored on a scale from 1 (least complex) to 5 (most complex) by a board certified plastic surgeon. Mortality, a component of the original Aristotle score, was excluded due to its limited relevance in peripheral nerve surgery. A total complexity score was calculated by summing the six domain scores (maximum of 30 per procedure), and this composite was used in exploratory regression analyses to examine associations between complexity and GPT-5.2 performance.

Sensitivity Analysis

To evaluate whether these findings remain consistent across a newer model generation, the same protocol was repeated on July 6, 2026 using GPT-5.5, the free version of ChatGPT provided by default at the time. OpenAI's most advanced model, GPT-5.6, was available only under a paid subscription at that time and was not used, to remain consistent with the free-tier conditions of the primary analysis. GPT-5.5 was accessed under the same conditions as GPT-5.2 (ChatGPT web interface, incognito browser window, default settings not manually enabled or disabled). This protocol was repeated for a subset of 5

of the 15 procedures selected at random: Thoracic Outlet Decompression, Carpal Tunnel Release, Posterior Approach - Spinal Accessory to Suprascapular Nerve Transfer, Anterior Interosseous to Ulnar Motor Supercharge Nerve Transfer, and Radial Nerve Release at the Spinal Groove. Each of the 5 selected procedures was queried 6 times using GPT-5.5 using the same structured prompt as the primary analysis (Online Resource 2). This served as a limited sensitivity analysis to assess whether the accuracy and qualitative findings observed for GPT-5.2 are consistent in a more recent model release.

To determine whether the performance observed for GPT-5.2 reflects the underlying capability of the model itself or is dependent on the structured, prompt-engineered query used in the primary analysis, the same 5 procedures selected for the sensitivity analysis above were also queried using a simple, unstructured, user-style prompt (Online Resource 2), representative of how a typical medical student or resident might inquire ChatGPT without training in prompt engineering. Each of the 5 procedures were queried 6 times using GPT-5.5, accessed under the same conditions described above. Responses generated using the simple prompt were evaluated using the same accuracy and qualitative analysis methods as the primary structured-prompt responses, allowing for a direct comparison between the structured and simple prompting strategies.

Statistical Analysis

Descriptive statistics were used to summarize GPT-5.2 and GPT 5.5 response characteristics (e.g. word count) and accuracy. To evaluate procedural completeness, each GPT response was compared to the reference standard transcript. A step was considered matched if the action, sequence, or intent described by GPT was semantically equivalent to the corresponding step in the transcript. The percentage of matched steps was calculated for each procedure by dividing the number of matched steps by the total number of steps in the reference standard. For qualitative analysis, both procedural steps and free-text reviewer comments were analyzed thematically. Two of the reviewers independently coded omissions using an inductive thematic approach [23]. Codes were grouped into content themes and refined iteratively through discussion. Discrepancies in coding or interpretation were resolved by consensus or by consulting a third reviewer when needed.

Mixed-effects linear regression models were used to evaluate relationships between procedure complexity, response accuracy, response length, number of procedural steps, and step density. Accuracy was modeled as a function of procedural complexity to assess whether increasing procedural complexity was associated with differences in accuracy. Word count was similarly modeled as a function of procedural complexity to evaluate whether more complex procedures produced longer responses. Accuracy was also modeled as a function of word count, the number of reference-standard procedural steps, and step density to examine whether response length, procedural scope, and response length normalized to procedural scope were associated with response accuracy. In addition, a multivariable mixed-effects regression model was performed to evaluate the independent associations of procedural complexity, word count, number of procedural steps, and step density with accuracy. All models included a random intercept for the procedure to account for repeated runs within the same

procedure. Regression coefficients (β), 95% confidence intervals, and p-values were reported. All statistical analyses were conducted using R statistical software (version 4.5.0).

Ethics Approval

As per the 2022 Tri-Council Policy Statement (2nd edition) Ethics approval was not required for this study, as data collection exclusively used publicly available information [24].

Results

General Characteristics and Readability

A total of 90 GPT-5.2 responses were generated across 15 peripheral nerve procedures (6 responses per procedure). GPT-5.2 responses had a mean word count of 1292.0 (SD 496.7), compared with 2871 (SD 1441) for the reference standard. The reference standards contained between 6 and 15 procedural steps, with an overall mean of 10.93 steps. The corresponding reference-standard step count was assigned to each GPT-5.2 response and, when normalized to response length, yielded a mean step density of 0.95 (SD 0.41) reference-standard steps per 100 words. For comparison, the reference-standard transcripts had a mean step density of 0.45 (SD 0.21) steps per 100 words. In terms of readability, mean FKGL, GFI and CLI for GPT-5.2 responses were 11.63, 12.19, and 15.79, respectively. General characteristics of the reference standard and GPT-5.2 responses are summarized in Table 1.

Table 1
General characteristics and readability of reference standard transcripts and GPT-5.2 Responses

	Reference Standard (n = 15)		GPT5.2 Responses (n = 90)	
	Mean (SD)	Range (IQR)	Mean (SD)	Range (IQR)
General Characteristics				
Video Duration	21:44 ± 11:53	19:22 (16:43)	-	-
Word Count	2871 ± 1441	2875 (1969)	1292.0 ± 496.7	1114.0 (772.8)
Number of Paragraphs	-	-	92.7 ± 34.9	87.0 (38.0)
Presence of Section Headers	-	-	100.0% (0.0%)	100.0% (0.0%)
Reference-Standard Procedural Steps	10.93 ± 2.95	11(5)	-	-
Step Density (per 100 words)	0.45 ± 0.21	0.41(0.26)	0.95 ± 0.41	0.93 (0.66)
Readability				
Flesch-Kincaid Grade Level	-	-	11.63 ± 0.94	11.55 (1.31)
Gunning Fog Index	-	-	12.19 ± 1.40	11.98 (1.75)
Coleman-Liau Index	-	-	15.79 ± 1.30	15.60 (1.55)
Average Sentence Length	-	-	8.85 ± 2.27	8.90 (3.08)

*For GPT response, step density represents reference-standard procedural steps per 100 GPT-generated words.

Accuracy and Reliability of Responses

Mean accuracy of the GPT-5.2 responses ranged from 68.5% (SD 4.5%) to 91.7% (SD 11.7%) across six independent runs per procedure (Table 2). The overall mean accuracy was 81.2% (SD 11.5%). Procedure complexity scores ranged from 6–23 (maximum score of 30, see Online Resource 5). In univariable regression analyses, procedural complexity ($\beta = 0.21$, 95% CI -0.30 to 0.73; $p = 0.43$; Fig. 1a) and the number of reference-standard procedural steps ($\beta = 0.17$, 95% CI -1.02 to 1.36; $p = 0.79$; Fig. 1b) were not significantly associated with accuracy. Step density was negatively associated with accuracy ($\beta = -6.23$, 95% CI -12.11 to -0.16; $p = 0.041$; Fig. 1c), indicating lower accuracy among responses with more reference-standard steps per 100 response words. Word count was positively associated with accuracy ($\beta = 0.0064$, 95% CI 0.0019 to 0.0108; $p = 0.0057$; Fig. 1d). In the multivariable regression including procedural complexity, word count, number of reference-standard procedural steps, and step density, no variable remained independently associated with accuracy. The association with step density was

attenuated ($\beta = 11.79$, 95% CI -8.41 to 33.02; $p = 0.27$), while word count retained a borderline positive association ($\beta = 0.0143$, 95% CI -0.00017 to 0.0294; $p = 0.062$).

Table 2
Mean GPT-5.2 accuracy of peripheral nerve procedures

Procedure	GPT-5.2 Responses	
	Mean (SD)	Range (IQR)
Carpal Tunnel Release	69.4 ± 12.5	66.7 (12.5)
Guyon's Canal Release and Carpal Tunnel Release	85.2 ± 5.7	88.9 (8.3)
Radial Nerve Release at the Spiral Groove	79.2 ± 10.2	81.2 (12.5)
Posterior Interosseous Nerve Release	88.9 ± 20.2	100.0 (12.5)
Median Nerve Release in the Forearm (with Carpal Tunnel Release)	80.0 ± 6.3	80.0 (0.0)
Medial Antebrachial Cutaneous Nerve Graft Harvest	68.5 ± 4.5	66.7 (0.0)
Submuscular Ulnar Nerve Transposition	80.0 ± 4.2	80.0 (5.0)
Thoracodorsal to Long Thoracic Nerve Transfer	87.5 ± 7.0	91.7 (6.3)
Anterior Interosseous to Ulnar Motor Supercharge Nerve Transfer	89.7 ± 13.5	96.2 (19.2)
Double Fascicular Transfer: Median Fascicle to Biceps Branch and Ulnar Fascicle to Brachialis Branch	80.5 ± 4.3	83.3 (6.2)
Brachialis to Anterior Interosseous Nerve Transfer with Extended Forearm Incision	81.1 ± 5.0	80.0 (5.0)
Medial Triceps to Axillary Nerve Transfer	81.8 ± 15.2	81.8 (6.8)
Posterior Approach - Spinal Accessory to Suprascapular Nerve Transfer	91.7 ± 11.7	95.0 (10.0)
Thoracic Outlet Decompression	78.6 ± 12.8	75.0 (12.5)
Median (Flexor Carpi Radialis/ Flexor Digitorum Superficialis) to Radial (Posterior Interosseous nerve/ Extensor Carpi Radialis Brevis) Nerve Transfers (with Pronator Teres to Extensor Carpi Radialis Brevis Tendon Transfer)	75.5 ± 3.5	73.3 (5.0)
Overall Scores	81.2 ± 11.5	80.00 (15.60)

Sensitivity Analysis

For the five-procedure sensitivity analysis, GPT-5.5 responses generated using the structured prompt had an overall mean accuracy of 81.4% (SD 16.2%), similar to the 81.7% accuracy observed for GPT-5.2 across the same five procedures. In contrast, GPT-5.5 responses generated using the simple, unstructured prompt had an overall mean accuracy of 40.5% (SD 40.2%). Procedure-level mean accuracy ranged from 76.7% (SD 22.5%) to 91.7% (SD 2.9%) with the structured prompt and from 33.3% (SD 38.5%) to 50.0% (SD 39.4%) with the simple prompt (Online Resource 6).

GPT-5.5 structured-prompt responses had a mean word count of 2335.2 (SD 1152.1) and mean paragraph number of 277.3 (SD 77.5), compared with 699.2 (SD 286.6) words and 86.3 (SD 52.5) paragraphs for simple-prompt responses. Mean Flesch-Kincaid Grade Level, Gunning Fog Index, and Coleman-Liau Index were 12.11 (SD 0.86), 12.77 (SD 2.21), and 15.36 (SD 1.22), respectively, for structured-prompt responses, compared with 13.25 (SD 1.17), 14.92 (SD 1.33), and 17.38 (SD 1.61) for simple-prompt responses. Mean average sentence length was 7.69 (SD 4.23) and 8.54 (SD 3.89), respectively. When normalized to response length, mean step density was 0.53 (SD 0.27) reference-standard steps per 100 words for structured-prompt responses and 1.63 (SD 0.73) for simple-prompt responses. General characteristics and readability across GPT-5.2, GPT-5.5 structured, and GPT-5.5 simple responses are summarized in Table 3.

Table 3
General characteristics, readability, and overall accuracy of GPT-5.2 and GPT-5.5 responses.

	GPT-5.2 Responses (n = 90)		GPT-5.5 Structured Responses (n = 30)		GPT-5.5 Simple Responses (n = 30)	
	Mean (SD)	Range (IQR)	Mean (SD)	Range (IQR)	Mean (SD)	Range (IQR)
Accuracy						
Overall Accuracy (%)	81.2 ± 11.5	80.00 (15.60)	81.4 ± 16.2	83.3 (30.4)	40.5 ± 40.2	45.0 (82.1)
General Characteristics						
Word Count	1292.0 ± 496.7	1114.0 (772.8)	2335.2 ± 1152.1	1571.5 (2028.5)	699.2 ± 286.6	593.0 (266.0)
Paragraph Number	92.7 ± 34.9	87.0 (38.0)	277.3 ± 77.5	288.5 (111.3)	86.3 ± 52.5	82.5 (42.3)
Step Density (per 100 words)	0.95 ± 0.41	0.93 (0.66)	0.53 ± 0.27	0.46 (0.44)	1.63 ± 0.73	1.35 (0.93)
Readability						
Flesch-Kincaid Grade Level	11.63 ± 0.94	11.55 (1.31)	12.11 ± 0.86	12.36 (1.40)	13.25 ± 1.17	13.00 (1.55)
Gunning Fog Index	12.19 ± 1.40	11.98 (1.75)	12.77 ± 2.21	12.62 (3.30)	14.92 ± 1.33	15.14 (1.54)
Coleman-Liau Index	15.79 ± 1.30	15.60 (1.55)	15.36 ± 1.22	15.49 (2.04)	17.38 ± 1.61	17.40 (2.22)
Average Sentence Length	8.85 ± 2.27	8.90 (3.08)	7.69 ± 4.23	4.80 (6.79)	8.54 ± 3.89	6.80 (5.34)

Qualitative Analysis

Across all 90 runs, there were 195 (out of 990 total steps) missed, incomplete or inaccurate steps. Qualitative analysis of these 195 steps identified 4 main themes: (1) failure to describe judgement-based, nuanced maneuvers, (2) failure to correctly identify key neurovascular structures during the dissection, (3) missed major steps and (4) failure to identify confirmatory visual or tactile intraoperative clues.

The most common type of missed step involved judgement-based or nuanced maneuvers that are dependent on intra-operative decision making and surgeon preference (86/195). These types of steps were missing in 12/15 procedures. These included steps by which the reference standard described adjuncts or discretionary modifications to the core procedure. In these cases, GPT-5.2 adequately described the corresponding major step that is necessary to allow progression of the procedure, but failed to describe beneficial additional steps, conditional reasoning or nuance as per the reference

standard. Examples of common additional, non-core missed steps would include the addition of sensory nerve transfers in motor nerve transfer procedures (e.g. performing radial sensory end-to-side coaptation to median nerve during a median-to-radial motor nerve transfer [6/6 runs] or a radial sensory end-to-side coaptation to the sensory branch of axillary nerve during a medial triceps to axillary motor nerve transfer [5/6 runs]) or mentioning the optional use of a tourniquet (e.g. during a radial nerve release at the spiral groove [6/6 runs]). Other common examples of judgement-based maneuvers include optional or preferential techniques to improve exposure (e.g. step-lengthening the superficial head of pronator teres during a brachialis to anterior interosseous nerve exposure [6/6 runs] or opening Guyon's canal to identify the "V" junction between the transverse carpal ligament and hypothenar muscle units during a carpal tunnel release [6/6 runs]).

Major steps were defined as steps that are necessary to allow progression of the procedure and were absent in 12/15 procedures. A major step was coded as "missed" if it was not included in the GPT output, or the output was incorrect (e.g. using a 3 – 0 or 4 – 0 nylon suture for the coaptation during medial triceps to axillary nerve transfer). There were 25 distinct missed major steps from the reference standards across the 15 procedures, comprising 56 out of 195 total missed steps across all 6 runs. Most (15/25) missed major steps were sporadic, occurring in only 1 out of 6 runs for a given procedure. Four major steps were consistently missed across all runs (e.g. the creation of a proximal and distal fascial flap overlying the flexor-pronator mass for coverage of the transposed ulnar nerve during a submuscular transposition).

The third theme was the failure to identify key nerves to protect during the dissection (41/195). These steps were absent in (6/15) of the procedures. These steps represent critical safety maneuvers intended to prevent iatrogenic injury rather than the major components of the operation. Common examples included identifying and protecting the pronator teres motor branch of median nerve during the median to radial nerve transfer [6/6 runs] or the pronator and flexor carpi radialis branch of median nerve during a brachialis to anterior interosseous nerve transfer [5/6 runs].

Lastly, the least common type of missed procedural step included a description of confirmatory visual or tactile intraoperative clues that surgeons may use to verify the adequacy of a completed operative step. These only contributed 12 out of the 195 total missed steps and were present in 4/15 procedures. The most common examples were the visual confirmation of the "V" formed by the junction of transverse carpal ligament and hypothenar muscles during a Guyon's canal and carpal tunnel release [6/6 runs], or the yellow mid-palmar fat pad confirming the division of the distal transverse carpal ligament during a carpal tunnel release [4/6 runs].

Qualitative findings were further evaluated in the GPT-5.5 sensitivity subset. Using the same structured query as the primary analysis, GPT-5.5 demonstrated 55 missed, incomplete, or inaccurate step instances across 306 reference-standard step instances. Judgement-based or nuanced maneuvers remained the most common deficiency, accounting for 29/55 (52.7%) missed step instances and occurring across all five procedures. Major procedural steps accounted for 23/55 (41.8%) omissions and

occurred in 4/5 procedures, whereas failure to identify or protect neurovascular structures accounted for 3/55 (5.5%) and occurred in one procedure. No confirmatory intraoperative cues were missed. Overall, 18 distinct reference-standard steps were omitted at least once. Representative recurrent omissions included opening Guyon's canal to identify the "V" during carpal tunnel release (6/6 runs), incision planning for radial nerve release at the spiral groove (4/6 runs), and stimulation of the suprascapular nerve to confirm lack of recovery prior to nerve transfer (4/6 runs). These findings suggest that the qualitative pattern observed in the primary GPT-5.2 analysis remained broadly consistent in GPT-5.5 when the structured prompting strategy was retained, with omissions continuing to predominantly involve surgeon-dependent nuance rather than wholesale failure of core procedural sequencing.

In contrast, GPT-5.5 responses generated using the simple, non-engineered prompt demonstrated substantially broader omissions, with 185 missed, incomplete, or inaccurate step instances across the same 306 reference-standard step instances. Major procedural steps were the predominant deficiency, accounting for 110/185 (59.5%) missed step instances and occurring across all five procedures. Judgement-based or nuanced maneuvers accounted for 47/185 (25.4%), also occurring across all five procedures. Failures to identify or protect neurovascular structures accounted for 19/185 (10.3%) and confirmatory intraoperative cues accounted for 9/185 (4.9%), occurring in 2/5 procedures each. Omissions occurred at least once for all 51 distinct reference-standard steps across the five procedures. Examples of consistently missed content included the skin and subcutaneous incision during radial nerve release at the spiral groove (6/6 runs), entry into the appropriate fat plane to identify the spinal accessory nerve during posterior spinal accessory to suprascapular nerve transfer (6/6 runs), and mobilization of the supraclavicular nerve bundle during thoracic outlet decompression (6/6 runs). Compared with the structured GPT-5.5 responses, simple prompting shifted the dominant pattern of omission away from judgement-based nuance and toward absence of core procedural steps, resulting in substantially more widespread content gaps.

Discussion

This study applied a step-based framework to evaluate the accuracy, completeness, reproducibility, and qualitative content deficiencies of AI-generated procedural descriptions in peripheral nerve surgery. Using GPT-5.2 as the primary model case study, overall mean accuracy was 81.2% across 90 responses for fifteen procedures, although substantial variability was observed both within and across procedures. Qualitative analysis of 195 missed, incomplete, or inaccurate step instances identified four recurring themes: failure to describe judgement-based or nuanced maneuvers (86/195, 44.1%), missed major procedural steps (56/195, 28.7%), failure to identify key neurovascular structures to protect (41/195, 21.0%), and absence of confirmatory intraoperative cues (12/195, 6.2%). Application of the same framework to GPT-5.5 and to different prompting strategies further demonstrated that both model generation and prompt design can be evaluated using a common procedural benchmark. These findings support the utility of the framework for characterizing not only aggregate accuracy, but also response variability and the type and clinical importance of omitted procedural content.

GPT-5.2 demonstrated several notable strengths that suggest utility for specific educational contexts. The overall accuracy of over 80% suggests that the model can reliably capture core procedural frameworks across operations of varying complexities. These findings echo the broader literature suggesting that newer iterations of ChatGPT produce generally accurate responses to well-defined, structured medical queries [9–11]. The response format was consistently well-organized and readable, and GPT-5.2 responses were substantially shorter than the reference-standard transcripts. Accordingly, GPT-5.2 demonstrated a higher reference-standard step density than the reference material (0.95 vs. 0.45 steps per 100 words), reflecting fewer generated words relative to the procedural scope defined by the reference standard rather than greater procedural completeness. The combination of accuracy and accessibility suggests that GPT-5.2 can generate content that may be valuable for early learners, such as medical students or junior residents, seeking preliminary orientation to unfamiliar procedures. This aligns with recent literature which demonstrates that AI tools show promise for knowledge acquisition by effectively synthesizing large volumes of content in surgical education [25, 26].

However GPT-5.2 also demonstrated several important drawbacks that limit its use as an isolated educational resource. Accuracy ranged widely across independent runs of the same procedure, with standard deviations exceeding 10% for twelve out of the fifteen procedures. As such, learners consulting GPT-5.2 on multiple occasions could receive substantially varied procedural accounts with no indication of which version is more complete. Moreover, major procedural steps were absent in many (12 out of 15) procedures. While most omissions were sporadic, its unpredictability can be hazardous as it precludes learners from knowing whether the received output had randomly omitted any critical content.

An important caveat to the accuracy metrics reported above is that aggregate accuracy weighs every missed step equally, regardless of clinical consequence — a model can score well overall while omitting steps critical for safety. Of the 195 missed, incomplete, or inaccurate steps identified, the 41 involving failure to identify key neurovascular structures are arguably the most clinically consequential, as these represent safety maneuvers intended to prevent iatrogenic injury rather than technical refinements. By comparison, the 86 missed judgement-based maneuvers more often reflect optional, surgeon-specific adjuncts rather than steps required for safe completion of the procedure. The overall accuracy of 81.2% may therefore mask clinically meaningful risk for learners relying on GPT-5.2, as a single missed neurovascular structure carries disproportionate weight relative to its contribution to the aggregate score.

The largest category of omitted content involved judgement-based and nuanced maneuvers, accounting for 86/195 (44.1%) missed, incomplete, or inaccurate step instances and occurring in 80% of procedures. These are not frivolous details but rather represent technical adjuncts, preferential modifications, and conditional reasoning that distinguish technically adequate procedural performance from expert operative practices. Regression analyses further characterized the relationship between response structure and accuracy. In univariable analyses, procedural complexity and the number of reference-standard procedural steps were not significantly associated with GPT-5.2 accuracy. Word count demonstrated a small positive association with accuracy, whereas step density demonstrated a

significant negative association, indicating that responses containing more reference-standard steps per 100 generated words tended to be less accurate. This should not be interpreted to mean that greater verbosity itself improves procedural quality; rather, step density is mathematically determined by both procedural scope and response length, and a higher value reflects fewer generated words relative to the number of reference-standard steps expected for that procedure. Procedural complexity was also positively associated with word count, suggesting that more complex procedures tended to generate longer responses without improving accuracy. In the multivariable mixed-effects model including procedural complexity, word count, number of reference-standard procedural steps, and step density, no variable remained independently associated with accuracy, although word count retained a borderline positive association. Because step density is derived directly from reference-standard step count and word count, these predictors are inherently interrelated, limiting interpretation of their independent effects when modeled simultaneously. While one might expect for simpler procedures to be captured with greater precision, the absence of a complexity-accuracy relationship suggests that GPT-5.2's limitations may instead relate to the difficulty for LLMs to reproduce judgement-based surgical knowledge that Apramian et al. 's work highlights as central to expert practice [27–29].

Apramian et al. described this occurrence as “thresholds of principle and preference” in their grounded theory study on procedural variation in surgical education [28]. They found that residents spend considerable parts of their learning distinguishing required procedural steps (principles) from those reflecting trained but interchangeable judgement (preferences) through observation and experience with multiple attending surgeons [28]. The steps most commonly omitted by GPT-5.2 in this current study directly correspond to “preferences” where conditional decisions, alternative approaches, and surgeon-specific modifications represent individualized knowledge over years of operative experience [28]. This finding is further reinforced by Guerrero et al. where AI showed promise for simulation-based training with automated feedback on reproducible technical skills, but failed to recognize and adapt to surgeon-specific variations that require human mentorship and exposure to multiple expert approaches [26].

The educational implications of this principle-preference distinction become even more complex when considering that surgeons themselves hold variable thresholds, and that significant variation in technique is typical across surgical training environments [29]. Upon recruiting an entire surgical division to assess the same operative performance, surgeons agreed on five generic principles but conflicted on their application: some deemed the performance competent for independent practice while others disagreed [29]. This variability implies that while core surgical principles exist, there is no singular “gold standard” for how a surgical procedure should be conducted which lends to the rarity of comprehensive technique papers in surgical literature [29]. This has direct implications for the present study where our reference standard was derived from a single educational resource and subsequently reviewed by a board-certified plastic surgeon. Rather than a methodological shortcoming, this exhibits an inherent feature of surgical education where variable expert knowledge resists standardization. The use of a single reference source supplemented by expert review represents a practical approach to this challenge, acknowledging that no single gold standard exists while still providing a rigorous benchmark for evaluation.

These nuances in surgical training have direct implications for understanding GPT-5.2's efficacy across education levels. For medical students or junior residents seeking preliminary procedural understanding, GPT-5.2 provides accessible, organized content that likely meets basic educational needs: core steps are generally accurate and efficient generation makes it a convenient resource for orientation. However, as training advances and senior learners begin navigating procedural variation and developing personal operative styles, describing core steps but omitting nuances provide less value to this group.

These insights, however, exist on the assumption that GPT-5.2's observed strengths and limitations reflect a stable property of the LLM rather than an artifact of this specific version. GPT-5.2 already reflects one version of a model that continues to evolve, with OpenAI releasing GPT-5.5 and GPT-5.6 between data collection (December 2025) and study completion. ChatGPT's performance on a task can meaningfully change within a few months even without a formal update [30], so conclusions drawn from a single time-point risk becoming outdated. As such, reporting guidelines for chatbot-related health research now recommends the documentation of model version, access date, and configuration so that findings can be contextualized by time rather than be treated as durable claims [31].

This concern was addressed through the sensitivity analyses, where GPT-5.5 did not demonstrate improved accuracy relative to GPT-5.2 on the 5 procedures tested (81.4% vs. 81.7%), though variability across runs increased. Importantly, the qualitative pattern of omissions was sustained across GPT-5.2 and GPT-5.5 using the same structured prompt, with judgement-based, surgeon-dependent maneuvers remaining the dominant category of missed content. This consistency across model versions lends support to the interpretation above, revealing that the limitation reflects a structural feature of LLM text generation, as opposed to a deficiency specific to GPT-5.2, and is therefore unlikely to resolve upon the release of newer models.

Notably, prompting strategy, rather than model version, was the predominant influence on performance. Accuracy fell from 81.4% to 40.5% when using a simple, unstructured prompt, and the prevailing omissions shifted from surgeon-dependent nuance to core procedural steps. Simple-prompt responses also demonstrated substantially higher step density than structured-prompt responses (1.63 vs. 0.53 reference-standard steps per 100 words), reflecting their much shorter response length relative to the same procedural scope rather than greater procedural completeness. This distinction reinforces that step density should be interpreted as a measure of response length normalized to procedural scope and not as a surrogate for accuracy. These differences suggest that prompting strategy may need to be weighed as much as model choice when evaluating AI-generated educational content, and that evaluations should report both to avoid misinterpreting a model's practical utility.

Several meaningful limitations of this study warrant discussion. The reference standard was derived from a single educational resource supplemented by review from a board-certified plastic surgeon. As discussed, there is no single "gold standard" for surgical procedures, with other surgeons likely producing different, yet equally legitimate, procedural frameworks. Video transcript from the reference material used may not capture complete procedural aspects as operations like wound closure may be

abbreviated in high level instructional content. The decision to divide procedures into discrete scorable steps requires subjective judgement and introduces variability and possible step misclassification. Additionally, there is a discrepancy in word count between the reference standard and GPT-5.2 outputs, which may explain the higher likelihood of the reference material containing more technical considerations and judgement-based steps. Furthermore, the process of coding the detailed reference standard into discrete steps reduces granularity and nuance, potentially omitting higher-order insights that may be particularly relevant for more advanced learners. The procedural complexity scoring system used in this work, while based on established surgical difficulty classifications, remains arbitrary. Interpretation of results also poses a challenge, as what defines an “acceptable” performance for surgical education remains unresolved. It is unclear what level of accuracy is deemed “sufficient” and how this threshold should vary across learner levels and clinical contexts. Finally, this study evaluated 15 peripheral nerve procedures, which is a relatively small sample from a single surgical subspecialty; similarly, the subset of 5 procedures used for the model-version and prompt-style sensitivity analyses were randomly selected, and may not be of sufficient size to completely capture the range of accuracy performance observed in the primary analysis. Together, these factors may limit the generalizability of these findings to other types of surgery.

Future research should explore how learners use AI tools in their surgical preparation and whether this correlates with operative performance and patient outcomes. More robust methodological approaches, including assessment of different surgical subspecialties and procedures, would be informative to whether patterns identified in peripheral nerve surgery generalize to other procedural contexts. Comparative studies using multiple reference standards from different experts could help characterize the extent to which omissions reflect AI limitation as opposed to acceptable procedural variation.

To support practical application and prospective evaluation of AI-generated procedural descriptions, we propose a six-step checklist combining structured prompting with critical review of AI-generated procedural descriptions (Online Resource 7). The checklist emphasizes defining the procedure and intended learner, explicitly requesting relevant content categories, requiring anatomical and procedural specificity, and distinguishing essential actions from surgeon-dependent preferences. It also recommends verifying supporting references and comparing generated descriptions with a trusted expert source, prioritizing clinically consequential omissions over aggregate completeness. This approach provides a practical starting point for generating and assessing procedural content as models evolve, although the checklist itself requires prospective evaluation and does not establish content accuracy or educational effectiveness.

For surgical education, the more durable question is not the performance of any single model version, but how AI-generated procedural content should be evaluated and applied across different stages of training. The present framework provides one approach for prospectively assessing model- and prompt-specific accuracy, reproducibility, and clinically relevant omissions as these systems continue to evolve. Nevertheless, given the limitations of this study, learners must interpret AI-generated responses with caution. Critically, GPT-5.2 cannot replace traditional apprenticeship models where residents develop

nuanced operative judgement through exposure to multiple approaches by different mentors. As AI continues to evolve and integrate into surgical education, establishing clear frameworks for appropriate use and recognizing inherent limitations will be essential to ensuring that these tools enhance, rather than compromise training quality.

References

1. Sánchez-Margallo JA, Plaza de Miguel C, Fernández Anzules RA, Sánchez-Margallo FM (2021) Application of Mixed Reality in Medical Training and Surgical Planning Focused on Minimally Invasive Surgery. *Front Virtual Real* 2:. <https://doi.org/10.3389/frvir.2021.692641>
2. Braund H, Patel V, Dalgarno N, Mann S (2024) Exploring residents' perceptions of competency-based medical education across Canada: A national survey study. *MedEdPublish* 14:2. <https://doi.org/10.12688/mep.19247.1>
3. Nagi F, Salih R, Alzubaidi M, et al (2023) Applications of Artificial Intelligence (AI) in Medical Education: A Scoping Review. *Stud Health Technol Inform* 305:648–651. <https://doi.org/10.3233/SHTI230581>
4. Mohammad B, Supti T, Alzubaidi M, et al (2023) The Pros and Cons of Using ChatGPT in Medical Education: A Scoping Review. *Stud Health Technol Inform* 305:644–647. <https://doi.org/10.3233/SHTI230580>
5. Totlis T, Natsis K, Filos D, et al (2023) The potential role of ChatGPT and artificial intelligence in anatomy education: a conversation with ChatGPT. *Surg Radiol Anat SRA* 45:1321–1329. <https://doi.org/10.1007/s00276-023-03229-1>
6. Introducing ChatGPT. <https://openai.com/index/chatgpt/>. Accessed 6 June 2025
7. Wu Y, Zheng Y, Feng B, et al (2024) Embracing ChatGPT for Medical Education: Exploring Its Impact on Doctors and Medical Students. *JMIR Med Educ* 10:e52483. <https://doi.org/10.2196/52483>
8. Liu G, Ma X, Zhang Y, et al (2023) GPT4: The Indispensable Helper for Neurosurgeons in the New Era. *Ann Biomed Eng* 51:2113–2115. <https://doi.org/10.1007/s10439-023-03241-x>
9. Johnson D, Goodman R, Patrinely J, et al (2023) Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. *Res Sq* rs.3.rs-2566942. <https://doi.org/10.21203/rs.3.rs-2566942/v1>
10. Pugliese N, Wai-Sun Wong V, Schattenberg JM, et al (2024) Accuracy, Reliability, and Comprehensibility of ChatGPT-Generated Medical Responses for Patients With Nonalcoholic Fatty Liver Disease. *Clin Gastroenterol Hepatol Off Clin Pract J Am Gastroenterol Assoc* 22:886-889.e5. <https://doi.org/10.1016/j.cgh.2023.08.033>
11. Ali K, Barhom N, Tamimi F, Duggal M (2024) ChatGPT-A double-edged sword for healthcare education? Implications for assessments of dental students. *Eur J Dent Educ Off J Assoc Dent Educ Eur* 28:206–211. <https://doi.org/10.1111/eje.12937>

12. Wei Q, Yao Z, Cui Y, et al (2024) Evaluation of ChatGPT-generated medical responses: A systematic review and meta-analysis. *J Biomed Inform* 151:104620. <https://doi.org/10.1016/j.jbi.2024.104620>
13. Venerito V, Lalwani D, Del Vescovo S, et al (2024) Prompt engineering: The next big skill in rheumatology research. *Int J Rheum Dis* 27:e15157. <https://doi.org/10.1111/1756-185X.15157>
14. Meskó B (2023) Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial. *J Med Internet Res* 25:e50638. <https://doi.org/10.2196/50638>
15. Lebhar MS, Velazquez A, Goza S, Hoppe IC (2024) Dr. ChatGPT: Utilizing Artificial Intelligence in Surgical Education. *Cleft Palate Craniofacial J* 61:2067–2073. <https://doi.org/10.1177/10556656231193966>
16. WashU Surgical Video Education Website. In: WashU Med. <https://surgicaleducation.wustl.edu/>. Accessed 6 June 2025
17. Superficial Radial Sensory Nerve Release with Tenotomy of the Brachioradialis Tendon. In: WashU Med. <https://surgicaleducation.wustl.edu/superficial-radial-sensory-nerve-release-with-tenotomy-of-the-brachioradialis-tendon/>. Accessed 6 June 2025
18. Grose EM, Holmes CP, Aravinthan KA, et al (2021) Readability and quality assessment of internet-based patient education materials related to nasal septoplasty. *J Otolaryngol - Head Neck Surg J Oto-Rhino-Laryngol Chir Cervico-Faciale* 50:16. <https://doi.org/10.1186/s40463-021-00507-z>
19. Take control of your content with ReadablePro. In: Readable. <https://readable.com/pro/>. Accessed 1 Feb 2026
20. Golan R, Ripps SJ, Reddy R, et al (2023) ChatGPT's Ability to Assess Quality and Readability of Online Medical Information: Evidence From a Cross-Sectional Study. *Cureus* 15:e42214. <https://doi.org/10.7759/cureus.42214>
21. Cocci A, Pezzoli M, Lo Re M, et al (2024) Quality of information and appropriateness of ChatGPT outputs for urology patients. *Prostate Cancer Prostatic Dis* 27:103–108. <https://doi.org/10.1038/s41391-023-00705-y>
22. Lacour-Gayet F, Clarke D, Jacobs J, et al (2004) The Aristotle score: a complexity-adjusted method to evaluate surgical results. *Eur J Cardio-Thorac Surg Off J Eur Assoc Cardio-Thorac Surg* 25:911–924. <https://doi.org/10.1016/j.ejcts.2004.03.027>
23. Braun V, and Clarke V (2006) Using thematic analysis in psychology. *Qual Res Psychol* 3:77–101. <https://doi.org/10.1191/1478088706qp063oa>
24. Government of Canada IAP on RE (2023) Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans – TCPS 2 (2022). https://ethics.gc.ca/eng/policy-politique_tcps2-eptc2_2022.html. Accessed 6 June 2025
25. Pakkasjärvi N, Luthra T, Anand S (2023) Artificial Intelligence in Surgical Learning. *Surgeries* 4:86–97. <https://doi.org/10.3390/surgeries4010010>
26. Guerrero DT, Asaad M, Rajesh A, et al (2023) Advancing Surgical Education: The Use of Artificial Intelligence in Surgical Training. *Am Surg* 89:49–54. <https://doi.org/10.1177/00031348221101503>

27. Apramian T, Cristancho S, Watling C, et al (2016) "Staying in the Game": How Procedural Variation Shapes Competence Judgments in Surgical Education. *Acad Med* 91:S37–S43.
<https://doi.org/10.1097/ACM.0000000000001364>
28. Apramian T, Cristancho S, Watling C, et al (2015) Thresholds of Principle and Preference: Exploring Procedural Variation in Postgraduate Surgical Education. *Acad Med* 90:S70–S76.
<https://doi.org/10.1097/ACM.0000000000000909>
29. Apramian T, Cristancho S, Sener A, Lingard L (2018) How Do Thresholds of Principle and Preference Influence Surgeon Assessments of Learner Performance? *Ann Surg* 268:385–390.
<https://doi.org/10.1097/SLA.0000000000002284>
30. Chen L, Zaharia M, Zou J (2024) How Is ChatGPT's Behavior Changing Over Time? *Harv Data Sci Rev* 6:. <https://doi.org/10.1162/99608f92.5317da47>
31. Huo B, Collins G, Chartash D, et al (2025) Reporting guideline for chatbot health advice studies: The CHART statement. *Artif Intell Med* 168:103222. <https://doi.org/10.1016/j.artmed.2025.103222>

Figures

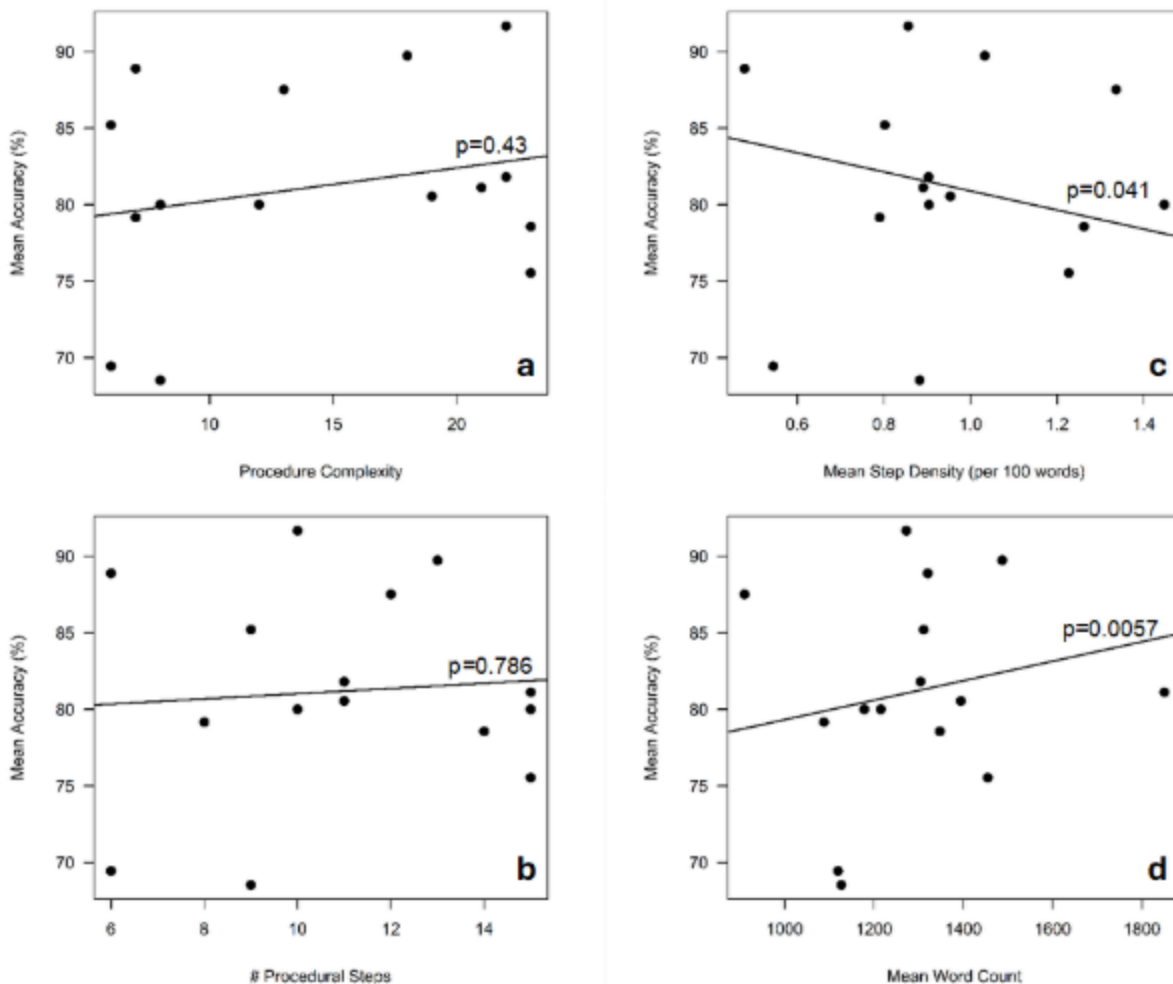


Figure 1

Association between GPT-5.2 response accuracy and (a) procedural complexity score, (b) number of reference-standard procedural steps, (c) step density per 100 words, and (d) word count. Each point represents the procedure-level mean across six independent GPT-5.2 responses. Lines represent the fixed-effect estimates from the corresponding univariable mixed-effects models fitted to all 90 responses, with procedure included as a random intercept.

Supplementary Files

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

- [ESM1.docx](#)
- [ESM2.docx](#)
- [ESM3.docx](#)
- [ESM4.docx](#)
- [ESM5.docx](#)
- [ESM6.docx](#)
- [ESM7.docx](#)