Clinical Text Summarization: Adapting Large Language Models Can Outperform Human Experts

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, Akshay S. Chaudhari

Abstract

Motivation

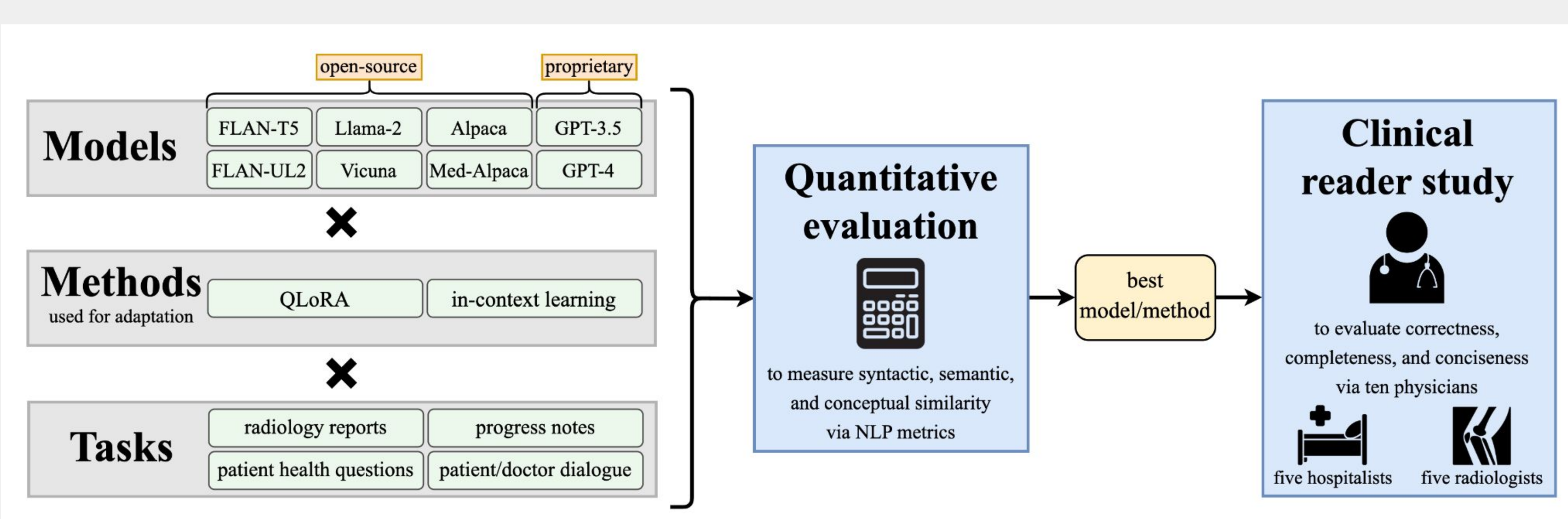
- Summarizing key information from electronic health records (EHR) imposes a substantial burden on how clinicians allocate their time.
- Large language models (LLMs) do well on general natural language processing (NLP) tasks, but their efficacy on summarizing clinical text has not been demonstrated.

Outcome

- Our research marks the first evidence of LLMs outperforming human experts for clinical text summarization.
- This implies that integrating LLMs into clinical workflows could alleviate documentation burden, enabling clinicians to focus more directly on patient care.

Overview

First, we quantitatively evaluate each valid combination (×) of LLM and adaptation method across four distinct summarization tasks comprising six datasets. We then conduct a clinical reader study in which ten physicians compare summaries of the best model/method against those of a human expert.



Approach

We compare two methods—quantized low-rank adaptation (QLoRA) and in-context learning (ICL)—for adapting eight models (**left**) to four summarization tasks (**right**).

Model	Context	Parameters	Proprietary?	Seq2seq	Autoreg.
FLAN-T5	512	2.7B	-	✔	
FLAN-UL2	2,048	20B	-	✔	
Alpaca	2,048	7B	-	-	✔
Med-Alpaca	2,048	7B	-	-	✔
Vicuna	2,048	7B	-	-	✔
Llama-2	4,096	7B, 13B	-	-	✔
GPT-3.5	16,384	175B	✔	-	✔
GPT-4	32,768	unknown	✔	-	✔

Task (Dataset)	Task description	Number of samples	Avg. number of tokens: Input	Avg. number of tokens: Target	Lexical variance
Radiol. reports (Open-i)	findings → impression	3.4K	52 ± 22	14 ± 12	0.11
Radiol. reports (MIMIC-CXR)	findings → impression	128K	75 ± 31	22 ± 17	0.08
Radiol. reports (MIMIC-III)	findings → impression	67K	160 ± 83	61 ± 45	0.09
Patient questions (MeQSum)	verbose → short question	1.2K	83 ± 67	14 ± 6	0.21
Progress notes (ProbSum)	notes → problem list	755	$1{,}013 \pm 299$	23 ± 16	0.15
Dialogue (ACI-Bench)	dialogue → assessment	126	$1{,}512 \pm 467$	211 ± 98	0.04

Task examples

1. **Radiology reports** (see right)

2. **Patient health questions** → **Summary**

Hello, I have been dealing with trimethylaminuria since I was a child. I have done some of my own research and it looks like not much can be done for this condition. I do not have it all over my body it's only in my armpits. In the past I've gone to doctors and dermatologist they gave me no answers until I looked online today and finally found out what I have. I don't know maybe I'm wrong. But this disease isn't even consider common because no one has done anything about it. I'm sure they're thousands of women withit... Can I be tested for it and help in some kind of way to finding a cure or something? What testing is done for this? And where? Thank you

How can I get tested and treated for trimethylaminuria?

3. **Progress notes** → **Problem list**

<ASSESSMENT>Ms. [**Known last name 12031**] is a [**Age over 90 **]yo female with HTN, CAD s/p CABG, osteoporosis, COPD, here with painless lower GI bleeding and active extravasation from branch of middle colic artery on CTA now s/p angiographic coiling of middle colic artery branch. <SUBJECTIVE>UOP low, gave 500cc NS bolus doing very well clinically track serialhctsstill having bloody bowel movements as expected ifhctstablelikely plan for scope 2amhct dropped to 29 from 35 [**Doctor FirstName 91**]-give 2 units and recheck 1hr after 2nd unit, 3-4 hours Lactose Intolerance (Oral) (Lactase) Unknown; Codeine Nausea/Vomiting Bactrim Ds (Oral)(Sulfamethoxazole/Trimethoprim) Unknown; Changes to and f Review of systems is unchanged from admission except as notedbelow Review of systems:<OBJECTIVE>Last dose of Antibiotics: Ciprofloxacin-[**2196-3-31**] 12:29 AM Infusions: Other ICU medications: Pantoprazole (Protonix)-[**2196-3-30**] 08:20 PM Other medications: Flowsheet Data as of[**2196-3-31**] 06:40 AM Vital signs Hemodynamic monitoring Fluid balance 24 hours Since [**98**] AMTmax: 36.3 C (97.3Tcurrent: 36.3 C (97.3 HR: 79 (79-92) bpm BP: 115/45(62) {93/32(48)-126/85(96)} mmHg RR: 19 (18-29)insp/min SpO2: 95% Heart rhythm: SR (Sinus Rhythm) Height: 62 Inch Total In: 3,554 mL 2,328 mL PO: TF: IVF: 179 mL 1,698 mL Blood products: 375 mL 630 mL Total out: 230 mL 191 mL Urine: 230 mL 191 mLNG: Stool: Drains: Balance: 3,324 mL 2,137 mL Respiratory support O2 Delivery Device: None SpO2: 95%ABG: ///27/General: Alert, oriented, no acute distress HEENT: Sclera anicteric, dry MM, oropharynx clear, dentures on upper teeth Neck: supple, JVP not elevated, no LAD Lungs: Clear to auscultation bilaterally, no wheezes, rales, rhonchi CV: Regular rate and rhythm, normal S1 + S2, II/VI SEM LUSB, well-healed thoracotomy scar Abdomen: soft, non-tender, very mildly distended, hyperactive bowel sounds, no rebound tenderness or guarding, no organomegaly appreciated Ext: upper extremities WWP, 2+ pulses; LE cool with weak butpalpable distal pulses107 K/uL12.6 g/dL 139 mg/dL 0.5 mg/dL 27mEq/L 4.4mEq/L 13mg/dL 107mEq/L 139mEq/L 29.7 % 10.7 K/uL image002.jpg][**2196-3-30**] 03:10 PM [**2196-3-30**] 09:25 PM [**2196-3-31**] 01:54 AMWBC 10.7 Hct 30 35.9 29.7 Plt107 Cr 0.5 Glucose 139 Other labs: PT / PTT / INR:13.5/28.2/1.2, ALT / AST:14/23, AlkPhos/ T Bili:43/2.0, Lactic Acid:1.1 mmol/L, Albumin:3.0 g/dL,LDH:223 IU/L, Ca++:7.8 mg/dL, Mg++:1.7 mg/dL, PO4:3.9 mg/dL

- GI bleed
- CAD
- UTI
- HTN
- Osteoporosis

4. **Doctor-patient dialogue** → **Assessment**

[DOCTOR] hi , [PATIENT] , how are you ? [PATIENT] hi . good to see you . [DOCTOR] it's good to see you as well . so i know that the nurse told you about dax . i'd like to tell dax a little bit about you . [PATIENT] sure . [DOCTOR] okay ? so , [PATIENT] is a 62-year-old male with a past medical history significant for a kidney transplant , hypothyroidism , and arthritis , who presents today with complaints of joint pain . [PATIENT] , what's going on with your joint ? what happened ? [PATIENT] uh , so , over the the weekend , we've been moving boxes up and down our basements stairs , and by the end of the day my knees were just killing me . [DOCTOR] okay . is , is one knee worse than the other ? [PATIENT] equally painful . [DOCTOR] okay . [PATIENT] both of them . [DOCTOR] and did you , did you injure one of them ? [PATIENT] um , uh , i've had some knee problems in the past but i think it was just the repetition and the weight of the boxes . [DOCTOR] okay . all right . and , and what have you taken for the pain ? [PATIENT] a little tylenol . i iced them for a bit . nothing really seemed to help , though . [DOCTOR] okay . all right . um , and does it prevent you from doing , like , your activities of daily living , like walking and exercising and things like that ? [PATIENT] uh , saturday night it actually kept me up for a bit . they were pretty sore . [DOCTOR] mm-hmm . okay . and any other symptoms like fever or chills ? [PATIENT] no . [DOCTOR] joint pain ... i mean , like muscle aches ? [PATIENT] no . [DOCTOR] nausea , vomiting , diarrhea ? [PATIENT] no . [DOCTOR] anything like that ? [PATIENT] no . [DOCTOR] okay . all right . now , i know that you've had the kidney transplant a few years ago for some polycystic kidneys . [PATIENT] mm-hmm . [DOCTOR] um , how are you doing with that ? i know that you told dr. gutierrez- [PATIENT] mm . [DOCTOR] . a couple of weeks ago . [PATIENT] yes . [DOCTOR] everything's okay ? [PATIENT] so far , so good . [DOCTOR] all right . and you're taking your immunosuppressive medications ? [PATIENT] yes , i am . [DOCTOR] okay . all right . um , and did they have anything to say ? i have n't gotten any reports from them , so ... [PATIENT] no , n- nothing out of the ordinary , from what they reported . [DOCTOR] okay . all right . um , and in terms of your hyperthyroidism , how are you doing with the synthroid ? are you doing okay ? [PATIENT] uh , yes , i am . [DOCTOR] you're taking it regularly ? [PATIENT] on the clock , yes . [DOCTOR] yes . okay . and any fatigue ? weight gain ? anything like that that you've noticed ? [PATIENT] no , nothing out of the ordinary . [DOCTOR] okay . and just in general , you know , i know that we've kind of battled with your arthritis . [PATIENT] mm-hmm . [DOCTOR] you know , it's hard because you ca n't take certain medications 'cause of your kidney transplant . [PATIENT] sure . [DOCTOR] so other than your knees , any other joint pain or anything like that ? [PATIENT] every once in a while , my elbow , but nothing , nothing out of the ordinary . [DOCTOR] okay . all right . now i know the nurse did a review of systems sheet when you checked in . any other symptoms i might have missed ? [PATIENT] no . [DOCTOR] no headaches ? [PATIENT] no headaches . [DOCTOR] anything like that w- ... okay . all right . well , i wan na go ahead and do a quick physical exam , all right ? hey , dragon , show me the vital signs . so here in the office , your vital signs look good . you do n't have a fever , which is good . [PATIENT] mm-hmm . [DOCTOR] your heart rate and your , uh , blood pressure look fine . i'm just gon na check some things out , and i'll let you know what i find , okay ? [PATIENT] perfect . [DOCTOR] all right . does that hurt ? [PATIENT] a little bit . that's tender . [DOCTOR] okay , so on physical examination , on your heart exam , i do appreciate a little two out of six systolic ejection murmur- [PATIENT] mm-hmm . [DOCTOR] . which we've heard in the past . okay , so that seems stable . on your knee exam , there is some edema and some erythema of your right knee , but your left knee looks fine , okay ? um , you do have some pain to palpation of the right knee and some decreased range of motion , um , on exam , okay ? so what does that mean ? so we'll go ahead and we'll see if we can ...

ASSESSMENT AND PLAN

[PATIENT] is a 62-year-old male with a past medical history significant for a kidney transplant, hypothyroidism, and arthritis. He presents today with complaints of joint pain.

Arthritis.
- Medical Reasoning: The patient reports increased joint pain in his bilateral knees over the past weekend. Given that his right knee x-ray was unremarkable, I believe this is an acute exacerbation of his arthritis.
- Additional Testing: We will order an autoimmune panel for further evaluation.
- Medical Treatment: Initiate Ultram 50 mg every 6 hours as needed.
- Patient Education and Counseling: I advised the patient to rest his knees. If his symptoms persist, we can consider further imaging and possibly a referral to physical therapy.

Hypothyroidism.
- Medical Reasoning: The patient is doing well on Synthroid and is asymptomatic at this time.
- Additional Testing: We will order a thyroid panel.
- Medical Treatment: Continue Synthroid.

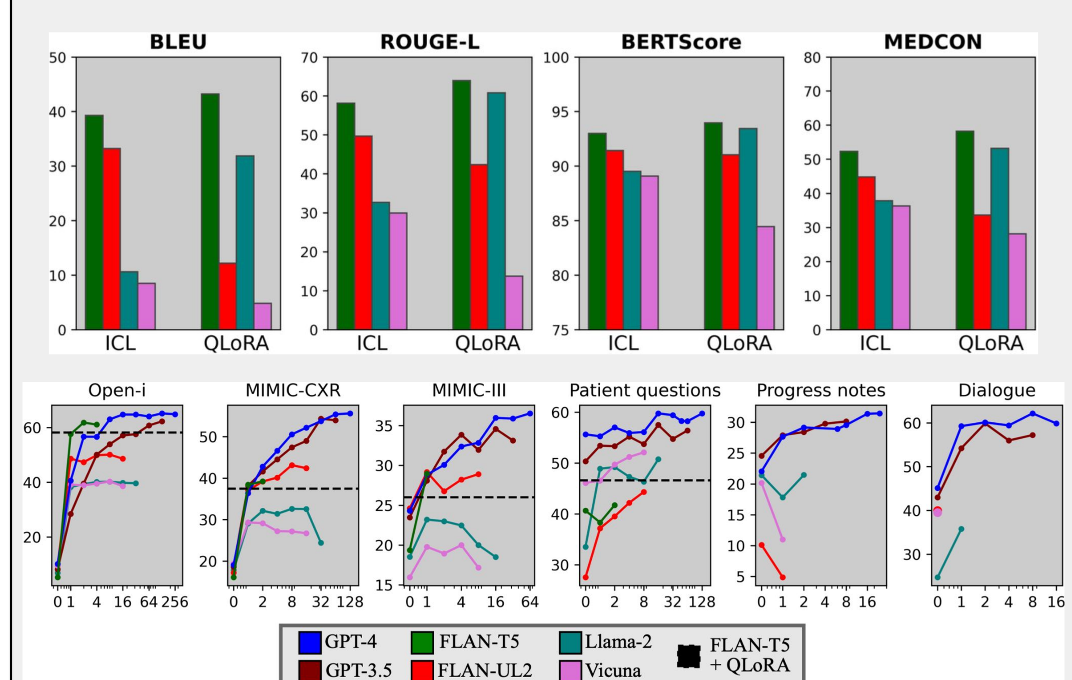
Status post renal transplant.
- Medical Reasoning: He is doing well and has been compliant with his immunosuppressive medications. On recent labs, his white blood cell count was within a normal limits and his kidney function is stable.
- Medical Treatment: Continue current regimen.

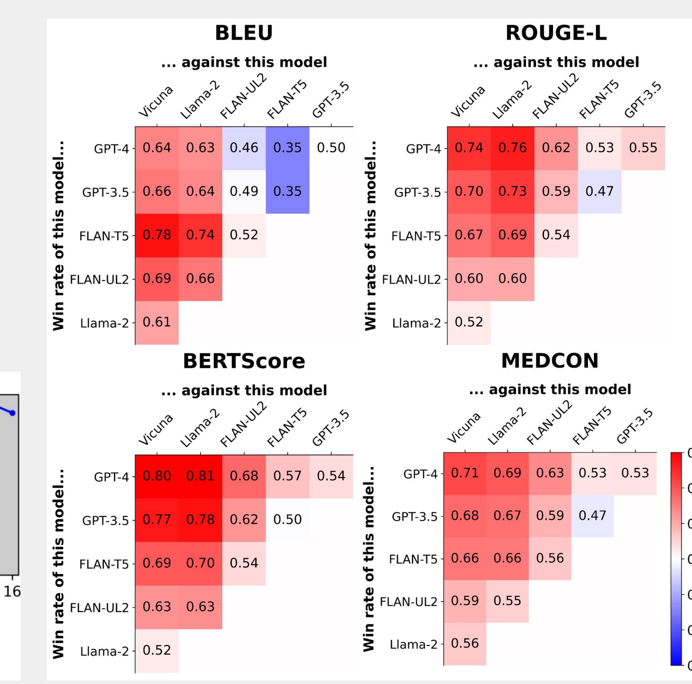
Prompt Anatomy

Expertise	`You are an expert medical professional.`
Instruction (task-specific)	`Summarize the [radiology report findings] into an [impression with minimal text].`
Examples $i = 1, \ldots, m$ #: delimiters *Note*: examples for ICL only, else $m = 0$	`Use the examples to guide word choice.` ⋮ input i: {example input} summary i: {example summary} ## ⋮
Input	input $m+1$: {input text} summary $m+1$:

Quantitative evaluation

Metrics ____ measure ____ similarity between generated and reference texts.
BLEU, ROUGE-L: syntactic **BERTScore**: semantic **MEDCON**: medical conceptual





Top left: FLAN-T5 emerges as the best open-source model fine-tuned with QLoRA.
Bottom left: Given enough examples, ICL surpasses QLoRA (dashed line). Including one example drastically improves performance compared to zero-shot prompting.
Right: Head-to-head win rates of each model combination. GPT-4 generally performs best. Seq2seq models outperform open-source autoregressive models.

→ *we conclude the best configuration is GPT-4 with maximal* in-context examples*

*determined by example size and model context length

Clinical Reader Study

Which summary...

[Completeness] ... more completely captures important information?
[Correctness] ... includes less false information?
[Conciseness] ... contains less non-important information?

Human significantly (-10) | Human slightly (-5) | neither (0) | GPT-4 slightly (5) | GPT-4 significantly (10)

Task	Completeness	Correctness	Conciseness
Radiology reports	2.8 ± 5.1 *	1.7 ± 3.7 *	0.0 ± 4.3
Patient questions	1.6 ± 6.5 *	0.6 ± 3.7 *	0.6 ± 3.9 *
Progress notes	2.6 ± 6.9 *	0.4 ± 4.8	0.6 ± 4.5 *
Overall	2.3 ± 5.8 *	0.8 ± 3.7 *	0.4 ± 4.0 *

Input: there is focal high attenuation overlying a superior left frontal gyrus, probably a dural calcification. subsequent mri shows no evidence of hemorrhage in this region. the brain parenchyma is normal. the ventricles and sulci are slightly prominent.

Summary A: there is no evidence of a hemorrhage and no area of low attenuation or mass effect is seen to suggest an acute infarct.

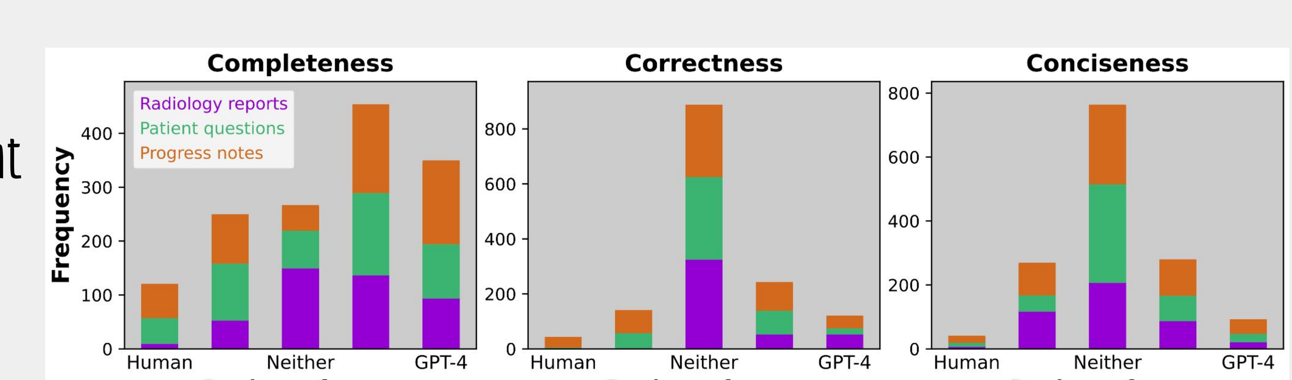
Summary B: no acute intracranial abnormality. probable dural calcification overlying a superior left frontal gyrus.

Which summary...

	A: significantly	A: slightly	neither	B: slightly	B: significantly
... more completely captures important information?	○	○	○	○	○
... includes less false information?	○	○	○	○	○
... contains less non-important information?	○	○	○	○	○

Left: Summaries generated via GPT-4 with in-context examples are rated higher with statistical significance (*) on all three attributes. **Right**: Example of study user interface.

Distribution of reader scores across five-point Likert scale.



Annotations of radiology report examples to illustrate the strengths and weaknesses of GPT-4 and humans.

Radiology reports

Example 1: GPT-4 performed *better* because the human made a laterality mistake (**red**).

Input:
The patient is s/p left frontal craniotomy. A small amount of intracranial gas is seen posterior to the surgical intervention, which could represent postoperative changes. Extensive edema is seen in the left frontal lobe at the site of presumed surgery. Additionally multiple foci of hemorrhage are seen in the region of the left frontal lobe. Midline shift to the right is seen in the frontal region. The ventricles, cisterns, and sulci are unremarkable, without effacement. Comparison with prior studies from outside institution would be helpful in further evaluation of these findings.

Summary (human expert):
1. Left frontal craniotomy. 2. Frontal midline shift to the right. 3. Extensive left frontal lobe edema. 4. Multiple foci of hemorrhage in the right frontal lobe.

Summary (GPT-4):
Postoperative changes following left frontal craniotomy with intracranial gas, extensive edema, and multiple foci of hemorrhage in the left frontal lobe. Midline shift to the right is noted.

Example 2: GPT-4 performed *worse* because it was less concise, i.e. included non-important information (**green**).

Input:
Right frontoparietal craniotomy again identified, with decreasing pneumocephalus and small amount of subdural fluid. There is persistent mild leftward shift of approximately 2-3 mm, not significantly changed from prior study. Size and configuration of the ventricles is unchanged from prior study. There is no new hemorrhage. Multiple facial fractures, better evaluated on prior dedicated maxillofacial CT. Increasing opacification is seen in the ethmoid and sphenoid sinuses. Large right subgaleal hematoma again identified.

Summary (human expert):
Postoperative changes in the right frontoparietal region identified, with decreasing pneumocephalus. No new hemorrhage. Persistent mild leftward shift, not significantly changed from prior.

Summary (GPT-4):
Stable post-craniotomy changes with decreasing pneumocephalus and persistent mild leftward shift. No new hemorrhage. Increasing opacification of ethmoid and sphenoid sinuses. Large right subgaleal hematoma unchanged.

Reader scores:

Attribute	Average	Example 1	Example 2
Completeness	3.9	5	-1.7
Correctness	2.5	10	0
Conciseness	-1.7	1.7	-5

Color key:

- **Blue**: correct; exists in input + human + GPT-4
- **Purple**: correct; exists in input + human only
- **Green**: correct; exists in input + GPT-4 only
- **Orange**: incoherent or filler
- **Red**: incorrect

Contact: Dave Van Veen | vanveen@stanford.edu | davevanveen.com