Supplementary Materials

Supplement A. In-Context Learning with Large Language Models: A Simple and Effective Approach to Improve Radiology Report Labeling

I. Method

1. Radiology Report Labeling

Two board-certified radiologists, with experience over 4 and 5 years of radiologic practice, performed the manual labeling of the radiology reports. If a radiology report contained findings described in Table S1, the corresponding label was annotated as positive. Following the manual labeling, the level of agreement was evaluated, and in instances of discrepancy, a consensus was achieved. These labels served as the ground truth for model performance evaluation. Considering the extensive length of the radiology reports, a pre-annotation process that employed regular expressions within the Python environment was introduced to improve the efficiency and accuracy of annotations. This process involved identifying keywords and phrases outlined in the annotation instructions. Subsequently, human annotators reviewed the entire document to confirm the accuracy of the pre-annotations generated by the regular expressions.

1) Experiment 1

Experiment 1 introduced a total of 10 labels: mass, vascular, volume loss, infarct, white matter, hydrocephalus, foreign body, hemorrhage, fracture, and pneumocephalus. Inspired by Wood's study, which utilized seven specialized categories of abnormality and five general abnormal categories in MRI report labeling, labels such as mass, vascular, volume loss, infarct, and hydrocephalus were chosen. The annotation rules for each label were defined, referring to the annotation rules from the cited study [1]. Furthermore, incorporating insights from Lorga's study, labels frequently identified in noncontrast head CT scans, such as hemorrhage, fracture, and pneumocephalus, were included [2].

2) Experiment 2

The definition of "actionable finding" in this study follows the categories outlined by the ACR actionable reporting work group [3]. Category 1 is defined as “critical or urgent findings, such as closed-loop intestinal obstruction, that require communication within minutes.” Category 2 is defined as “clinically significant findings, such as intra-abdominal infections including appendicitis and cholecystitis, that require communication within hours.” Category 3 findings are incidental or unexpected but do not require immediate treatment or other action, such as

liver cirrhosis.

To assess the ability of GPT-4 to identify urgent findings, this study merged Categories 1 and 2, defining "actionable findings" as those necessitating communication within hours. The categorization for this research incorporated findings from the gastrointestinal (GI), genitourinary (GU), musculoskeletal (MSK), and vascular sections within the abdomen and pelvis.

Findings from the lung base, such as pulmonary thromboembolism detected on abdominal CT, were excluded from this study. We also excluded the “General” actionable section, which involved subjective decisions such as “determining that the interpreting radiologist requires immediate physician notification.” Additionally, findings that were unobservable or unevaluable via Abdomen CT were also excluded. These include fetal Doppler ultrasound findings and those outside the abdomen, such as coronary artery occlusion. Findings that were previously deemed actionable but showed no significant interval change were not classified as actionable in this study. By contrast, findings that indicate progression were considered actionable. When no comparison with previous findings was mentioned, observations were considered novel for this study and annotated accordingly.

2. GPT-4 Prompt Engineering

1) Background

PE emerges as a field of study that effectively enhances the performance of LLMs without the necessity for fine-tuning by carefully crafting the input prompt [4]. Numerous prompt engineering techniques have been introduced, notably few-shot prompting [5], which improves performance by providing a few examples of input data and desired output; chain-of-thought prompting [6], which increases the accuracy of reasoning by inducing stepwise reasoning; and other advanced methodologies that apply these concepts.

However, for specific tasks, appropriate prompts should be developed. Among existing prompt engineering techniques, we referred to the following methods. First, we structured our prompts. Compared with free-text written prompts, structured prompts have the advantages that they provide a reusable framework and can easily be modified or adapted to other prompts. Moreover, structuring prompts increases the probability of obtaining precise and relevant responses [7-9].

Second, we addressed the consistency and accuracy of a model as distinct concepts [10]. Thus, we created two prompts for our experiments. One prompt served as a baseline, designed to perform the task consistently. The other prompt built upon the baseline with the additional goal of enhancing the reasoning capability of the model.

Table S1. List of labels used in Experiments 1 and 2 and the descriptions of annotation instructions provided to annotators

Experiment 1	
Label	**Annotation instruction**
Mass	- Neoplasm (tumor): both intra-axial and extra-axial neoplasm included - Abscess - Cysts: retrocerebellar cyst, arachnoid cyst, pineal cysts, and choroid fissure cysts
Vascular	- Aneurysm - Atherosclerosis - Thrombosis or occlusion - Arteriovenous malformation - Arteriovenous fistula - Other cerebrovascular malformation (such as cavernoma, capillary telangiectasia, or developmental venous anomaly) - Congenital anatomical variations are not included
Volume loss	- Diffuse brain atrophy - Encephalomalacia - Post-operative tissue changes - Chronic or old infarction with volume loss
Infarct	- Any findings suggestive of infarct - Old or chronic infarct with encephalomalacia are both labeled as “infarct,” and “volume loss”
White matter	- Any findings describing white matter inflammation or small vessel disease on CT.
Hydrocephalus	- Acute hydrocephalus - Trapped ventricle - Chronic/stable/improving hydrocephalus (whether it's compensated or not) - Ventricular enlargement - Normal pressure hydrocephalus (NPH)
Foreign body	- Shunts - Clips - Coils - Other materials related to surgery or procedure
Hemorrhage	- Parenchymal hemorrhage - Subarachnoid hemorrhage - Subdural hemorrhage - Extradural hemorrhage
Fracture	- Any displaced/non-displaced bony fracture on skull and upper cervical vertebra
Pneumocephalus	- Any findings suggestive of pneumocephalus on CT
Experiment 2	
Urgent	**Annotation instruction**
GI	Category 1: Communicate within minutes - Unexplained pneumoperitoneum - Closed loop intestinal obstruction - Intestinal ischemia and/or portal/mesenteric venous gas - Pseudoaneurysm or active hemorrhage (post-trauma, GI bleed, other) - High-grade intra-abdominal organ injury (liver, spleen, pancreas, other) and/or bowel injury post-

	trauma, acute intervention likely Category 2: Communicate within hours - Abscess, any location - Intestinal obstruction, no evidence of acute ischemia - Intra-abdominal infection, likely surgical or interventional candidate (appendicitis, cholecystitis, diverticulitis, abscess, other) - Large volume ascites - Low to moderate grade intra-abdominal organ injury and/or bladder or bowel injury post-trauma, observation likely - Pneumatosis in bowel wall, no other signs of ischemia
GU	Category 1: Communicate within minutes - Testicular torsion - Ovarian torsion - Ectopic pregnancy (high likelihood) - Placental abruption - Uterine rupture - High-grade kidney injury and/or ureteral or bladder injury post-trauma, acute intervention likely - Absent perfusion post-op kidney Category 2: Communicate within hours - Placenta previa or suspected placenta accreta, increta, percreta in third trimester - Urinary tract obstruction (stone, tumor, other) - Pyonephrosis/renal abscess - Indeterminate findings for ectopic vs normal pregnancy
MSK	Category 1: Communicate within minutes - Nonspinal fracture and/or dislocation with risk of vascular compromise - Necrotizing fasciitis Category 2: Communicate within hours - Bone lesion at risk for pathologic fracture - Nonspinal fracture and/or dislocation without vascular compromise, likely requiring intervention - Large hematoma without or with fracture, especially with compression of adjacent structures - Fracture follow-up imaging, significant change in alignment or concern regarding infection (including septic arthritis and osteomyelitis) - Hardware complication
Vascular	Category 1: Communicate within minutes - Ruptured/leaking arterial aneurysm - Arterial dissection or intramural hematoma Category 2: Communicate within hours - Hemodynamically significant arterial stenosis or occlusion, associated with acute symptoms - arterial pseudoaneurysm post-vascular access - Abdominal aortic aneurysm >5 cm, no evidence of acute instability - Previously unknown chronic arterial dissection or intramural hematoma

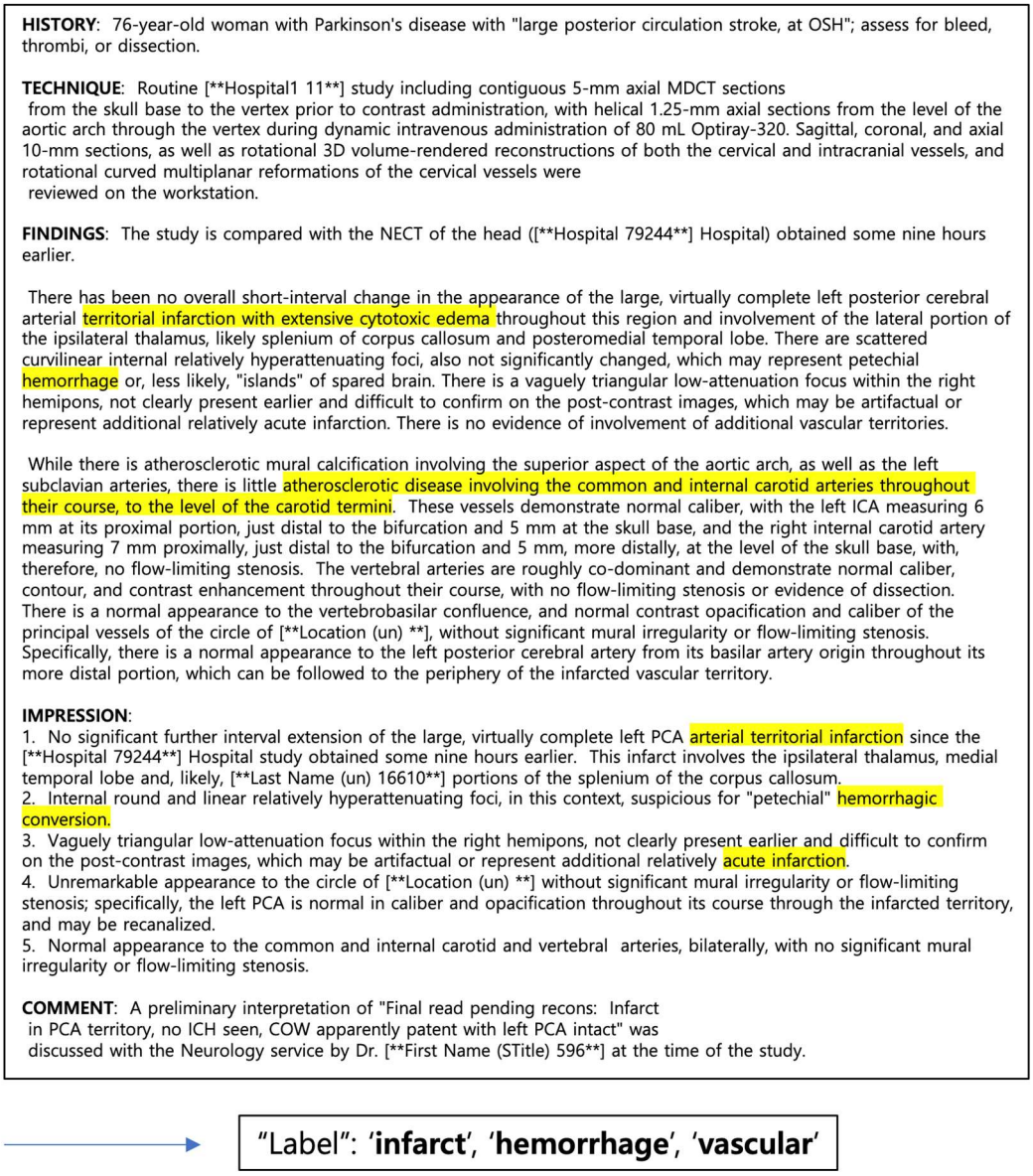

Figure S1. Example of Head CT report extracted from MIMIC-III database and its labeling. The report contains diverse clinical findings. Name, place, and time in the reports are sufficiently anonymized.

Third, we created a separate output section and specified the output to follow a particular format and content. This specification constraint the model to perform semantically similar tasks consistently, prevented the output from becoming verbose, improved performance, and reduced hallucination where the model provided plausible but incorrect answers [11].

Finally, we considered in-context learning, as the most important aspect of our research [12,13]. In-context learning, including few-shot learning, is an umbrella term referring to the improvement of the reasoning performance of a model via the context provided within the prompt. This context is typically considered to be a

few-shot set of input–output pairs that demonstrate the desired task. However, in addition to these real data, well-crafted human-written sentences can serve as effective context [12].

2) PE strategy used in this study

As described above, we designed two types of prompts. Initially, the "Basic prompt" was designed to perform the same task consistently and serve as a baseline for evaluating the performance of the "In-context prompt." Specifically, the "Task" section provided stepwise instructions tailored to each labeling scenario, while the "Output format" section was constrained to JavaScript Object Notation (JSON), a computer-friendly format that facilitates parsing multiple labels in the output and enables convenient post-processing. The "Basic prompt" did not include any domain-specific context, enabling the model to freely respond based on its knowledge and decisions.

Furthermore, for the "In-context prompt," the "Context" section aimed to provide relevant domain-specific information to leverage the in-context learning capabilities of LLMs. We included the annotation instructions used by human annotators, as they were the most appropriate and relevant context for the labeling task at hand. These carefully crafted summaries of the annotation instruction were provided in the input prompt, effectively guiding the model toward the desired labeling method while maintaining the prompt relatively concise compared with when providing entire examples of the report.

Our method maintained brevity and clarity, unlike other methods. For instance, few-shot prompts provided entire examples and correct labels as examples [5]. This approach was highly inefficient and impractical because providing examples for all labels and cases led to lengthy reports. Furthermore, this long prompt degraded performance [14], and consequently, all possible cases could not be covered. Moreover, it is expensive because the prompts are excessively long. Another approach, Chain-of-thought prompting, involves providing the model with examples or attempting stepwise reasoning without examples [6]. However, this also leads to lengthy outputs and does not provide the model with domain-specific context.

Additionally, in the GPT-4 application programming interface, prompts are categorized into User Prompts (entered by human users) and System Prompts. System Prompts, often not visible to users on the ChatGPT website, define the functions of the model (e.g., "You are a helpful assistant"). In this study, we assigned predefined prompts to the System Prompt, which assigned the model the identity of a report labeler and allocated only the radiology reports as User Prompts. This approach not only maintained concise prompts and Python code but also enabled the evaluation of our predefined prompts solely by clearing existing prompts in the System Prompt and assigning only the predefined prompts, thus excluding the possible effects from other prompts.

II. Results

1. Optimization Experiment

Table S2. Inconsistency and output format error in Head CT

	Temperature	Head CT				
		Inconsistency	Output format error	Syntactic error	Semantic error	Undefined label error
Basic prompt	0	0.02	0	0	0	0
	0.2	0.05	0	0	0	0
	0.4	0.05	0	0	0	0
	0.6	0.05	0	0	0	0
	0.8	0.07	0	0	0	0
	1	0.12	0.04	0.02	0.02	0
ICL prompt	0	0	0	0	0	0
	0.2	0.03	0.02	0	0.02	0
	0.4	0.02	0.01	0	0.01	0
	0.6	0.02	0.06	0.01	0.05	0
	0.8	0.03	0.05	0.03	0.02	0
	1	0.06	0.08	0.03	0.05	0

CT: computed tomography, ICL: In-context learning.

Table S3. Inconsistency and output format error in Abdomen CT

	Temperature	Abdomen CT				
		Inconsistency	Output format error	Syntactic error	Semantic error	Undefined label error
Basic prompt	0	0.03	0	0	0	0
	0.2	0.04	0	0	0	0
	0.4	0.06	0.01	0.01	0	0
	0.6	0.07	0.01	0	0	0.01
	0.8	0.06	0.02	0.02	0	0
	1	0.12	0.02	0.02	0	0
ICL prompt	0	0.01	0	0	0	0
	0.2	0.02	0.01	0.01	0	0
	0.4	0.01	0.01	0.01	0	0
	0.6	0.02	0.03	0.03	0	0
	0.8	0.03	0.03	0.03	0	0
	1	0.04	0.05	0.05	0	0

CT: computed tomography, ICL: In-context learning.

Table S4. Failed cases analysis

	Report example	Error analysis	In-context prompt	Ground truth
Experiment 1	"…There is minimal polypoid thickening in the frontal sinuses, and there is a small osteoma in the right frontal sinus. …"	The osteoma was incorrectly recognized as a tumorous lesion, resulting in a failure to apply the correct "mass" tag and leading to a false negative error. There may be low sensitivity for relatively uncommon tumor lesions.	Vascular	Mass, vascular
	"…Following recent mechanical thrombectomy for right MCA occlusion, there is patchy cortical hyperdensity along the right insular and frontal opercular cortex. Findings are favored to represent contrast staining rather than hemorrhage. No acute intracranial hemorrhage is seen."	The model possibly misinterpreted the term "hyperdensity" from the original report as hemorrhage, despite it representing contrast staining. Phrases explicitly negating hemorrhage, such as "favored to represent contrast staining rather than hemorrhage" and "no acute hemorrhage," were inadequately understood by the model, leading to a false positive error.	Hemorrhage, post-surgical, vascular	Post-surgical, vascular
Experiment 2	"INDICATION: 75-y M, sudden periumbilical pain & lactic acidosis … KEY FINDINGS: focal non-opacification of proximal SMA … ~25 cm jejunal segment: hypo-enhancement + mild pneumatosis … IMPRESSION: *Occlusive acute mesenteric ischemia* ⇒ emergent re-vascularization …"	The original report described vascular occlusion using ambiguous terms such as "non-opacification" instead of the explicit keyword "thrombosis," causing the model to fail to recognize these descriptions as indicative of occlusion, resulting in a false negative error.	GI	GI, Vascular

2. Main Experiment

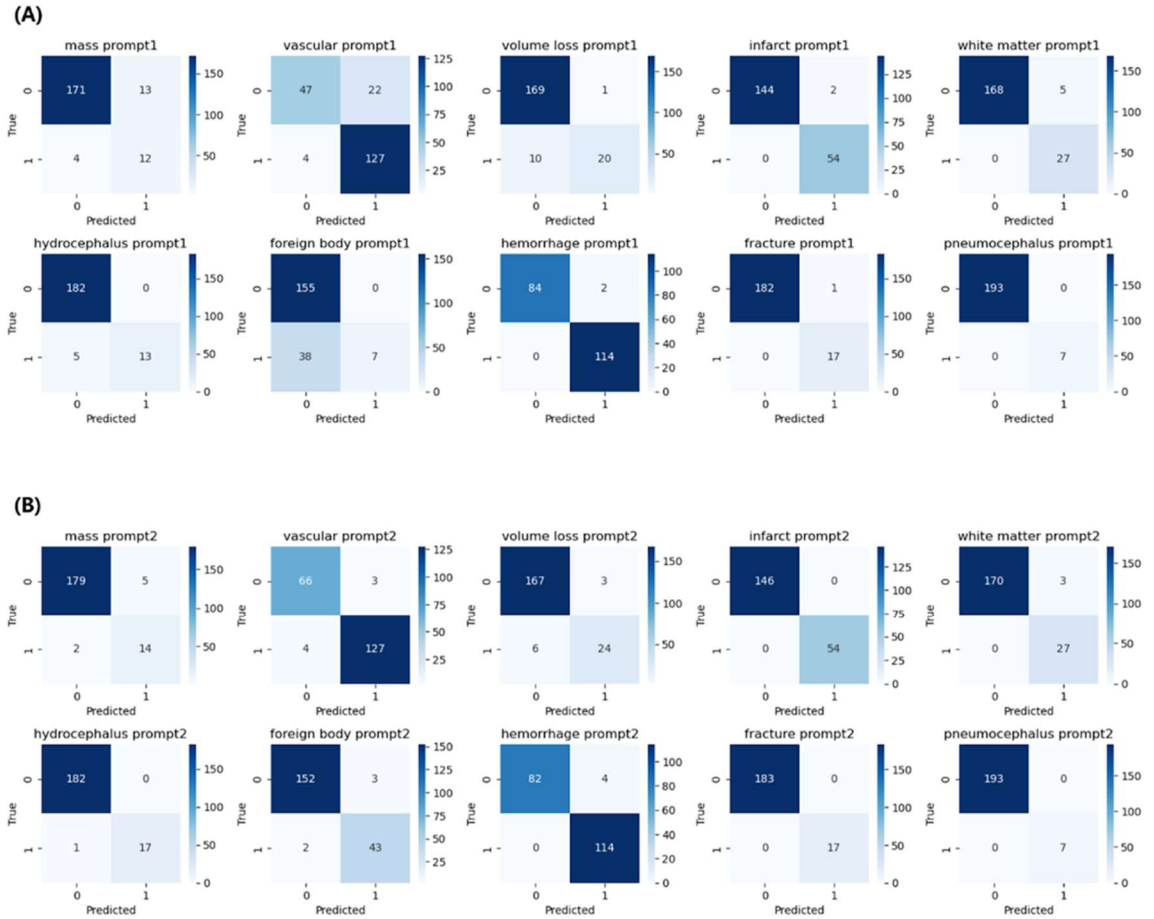

Figure 2. (A) Confusion matrix classified by GPT-4 for each label in Experiment 1 using "Basic prompt." (B) Confusion matrix classified by GPT-4 for each label in Experiment 1 using "In-context prompt."

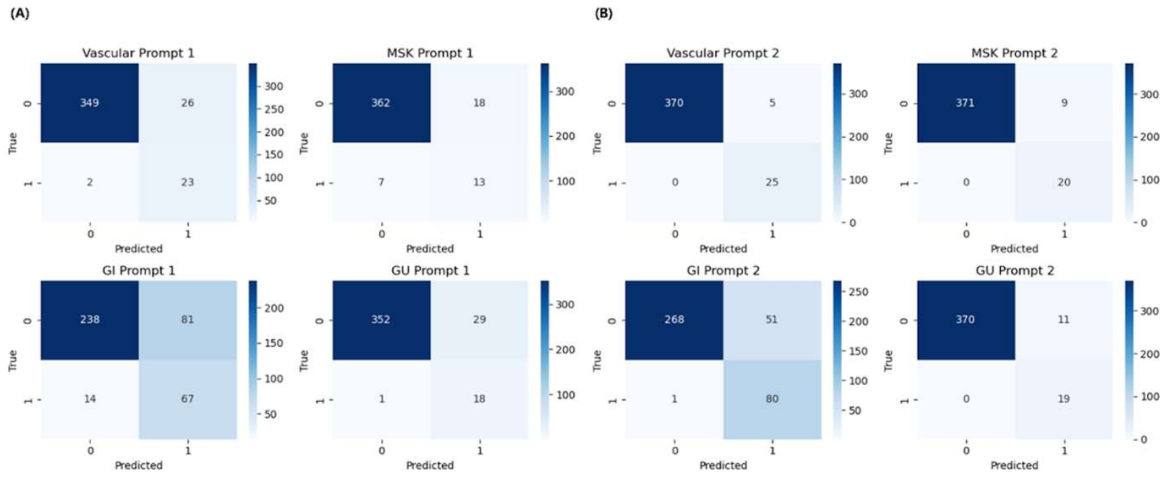

Figure S3. (A) Confusion matrix classified by GPT-4 for each label in Experiment 2 using "Basic prompt." (B) Confusion matrix classified by GPT-4 for each label in Experiment 2 using "In-context prompt."

References

1. Wood DA, Kafiabadi S, Al Busaidi A, Guilhem EL, Lynch J, Townend MK, et al. Deep learning to automate the labelling of head MRI datasets for computer vision applications. Eur Radiol 2022;32(1):725-36. https://doi.org/10.1007/s00330-021-08132-0
2. Iorga M, Drakopoulos M, Naidech AM, Katsaggelos AK, Parrish TB, Hill VB. Labeling noncontrast head CT reports for common findings using natural language processing. AJNR Am J Neuroradiol 2022;43(5):721-6. https://doi.org/10.3174/ajnr.A7500
3. Larson PA, Berland LL, Griffith B, Kahn CE, Liebscher LA. Actionable findings and the role of IT support: report of the ACR Actionable Reporting Work Group. J Am Coll Radiol 2014;11(6):552-8. https://doi.org/10.1016/j.jacr.2013.12.016
4. Ekin S. Prompt engineering for ChatGPT: a quick guide to techniques, tips, and best practices [Internet]. New York (NY): TechRxiv; 2023 [cited at 2025 Jul 1]. Available from: https://doi.org/10.36227/techrxiv.22683919.v2.
5. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners [Internet]. Ithaca (NY): arXiv.org; 2020 [cited at 2025 Jul 1]. Available from: https://arxiv.org/abs/2005.14165.
6. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, et al. Chain-of-thought prompting elicits reasoning in large language models. Adv Neural Inf Process Syst 2022;35:24824-37.
7. Lin Z. How to write effective prompts for large language models. Nat Hum Behav 2024;8(4):611-5. https://doi.org/10.1038/s41562-024-01847-2
8. Wang M, Liu Y, Liang X, Li S, Huang Y, Zhang X, et al. LangGPT: rethinking structured reusable prompt design framework for LLMs from the programming language [Internet]. Ithaca (NY): arXiv.org; 2024 [cited at 2025 Jul 1]. Available from: https://arxiv.org/abs/2402.16929.
9. Bansal M. A comprehensive guide to prompt engineering: Unveiling the power of the COSTAR template [Internet]. San Francisco (CA): medium.com; 2024 [cited at 2024 May 31]. Available from: https://levelup.gitconnected.com/a-comprehensive-guide-to-prompt-engineering-unveiling-the-power-of-the-costar-template-944897251101.
10. Raj H, Gupta V, Rosati D, Majumdar S. Semantic consistency for assuring reliability of large language models [Internet]. Ithaca (NY): arXiv.org; 2023 [cited at 2025 Jul 1]. Available from: https://arxiv.org/abs/2308.09138v1.
11. Liu MX, Liu F, Fiannaca AJ, Koo T, Dixon L, Terry M, et al. "We need structured output": towards user-

centered constraints on large language model output. Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA); 2024 May 11-16; Honolulu, HI, USA. p. 1-9. https://doi.org/10.1145/3613905.3650756

12. Dong Q, Li L, Dai D, Zheng C, Ma J, Li R, et al. A survey on in-context learning [Internet]. Ithaca (NY): arXiv.org; 2024 [cited at 2025 Jul 1]. Available from: https://arxiv.org/abs/2301.00234.
13. Wei J, Wei J, Tay Y, Tran D, Webson A, Lu Y, et al. Larger language models do in-context learning differently [Internet]. Ithaca (NY): arXiv.org; 2023 [cited at 2025 Jul 1]. Available from: https://arxiv.org/abs/2303.03846.
14. Liu NF, Lin K, Hewitt J, et al. Lost in the Middle: How Language Models Use Long Contexts. arXiv [csCL]. Published online July 6, 2023. http://arxiv.org/abs/2307.03172