How Agent Role Structure Alters Operating Characteristics of Large Language Model Clinical Classifiers: A Comparative Study of Specialist and Deliberative Multi-Agent Protocols

Callum Anderson*
Faculty of Engineering
Master of Interdisciplinary Artificial Intelligence Program
University of Ottawa

Abstract

Large language models (LLMs) are increasingly evaluated for structured clinical decision support tasks, often using multi-agent architectures. Prior work has compared single-agent and multi-agent inference. However, the effect of internal role structure within multi-agent systems on classification behavior remains underexplored. We evaluate two multi-agent prompting protocols, implemented as deterministic Directed Acyclic Graph (DAG) systems, a *Generic Deliberative* (GD) protocol and a *Feature-Specialist* (FS) protocol, on tabular clinical heart disease data from the UCI Cleveland dataset. Structured variables are rendered into primarily text-based feature descriptions while preserving clinically relevant numeric values. The two protocols differ only in their prompt-level role decomposition and information routing, while base model, model weights, deterministic decoding with temperature set to 0, computational budget, and aggregation logic are held constant. The results show systematic differences in predictive behavior attributable solely to prompt-level role structure. The FS protocol improves overall accuracy by 0.07 and macro-F1 by 0.06. However, this improvement is accompanied by a marked operating-point shift in which specificity increases by 0.22 while sensitivity decreases by 0.13, with corresponding redistribution of class-wise precision. Notably, the increase in specificity corresponds to a reduction in false positive classifications, indicating decreased over-diagnosis under the FS configuration. These findings indicate that multi-agent role decomposition introduces a structured inductive bias in deterministic LLM-based classification. Prompt protocol and agent role design should therefore be regarded as core modeling decisions, as they show measurable influence on performance tradeoffs, particularly in safety-sensitive deployment contexts.

1 Introduction

Recent advances in large language models (LLMs) have increased interest in their use in clinical settings. Healthcare institutions now maintain extensive structured electronic health record (EHR) repositories, and artificial intelligence techniques are increasingly being applied to these data for clinical prediction, decision support, and workflow optimization [5]. Reviews of medical natural language processing in the LLM era show that modern models encode substantial biomedical knowledge and can support clinically relevant reasoning, particularly when adapted through instruction tuning and task-specific guidance [13]. Empirical studies further demonstrate that LLMs can process structured clinical artifacts, such as eligibility criteria and patient descriptors, and produce representations

*cande007@uottawa.ca

that support grouping and classification tasks [1]. LLM-based systems have also been explored for identifying social and contextual patient information from clinical records and incorporating these factors into predictive workflows [10]. Together, this body of work suggests that LLMs can operate over structured clinical information and support medical classification-oriented tasks. In medical contexts, locally deployable large language models are of particular interest, as they allow sensitive patient data to remain within institutional infrastructure while maintaining operational autonomy and regulatory compliance [9].

Efforts to improve reliability and reasoning depth in LLM-based systems have increasingly adopted multi-agent or chained architectures. Rather than relying on a single model invocation, these systems decompose complex tasks into coordinated sub-tasks executed by multiple LLM instances, often with explicit roles or sequential dependency structures [6]. These approaches draw conceptual parallels to ensemble-based decision systems, where multiple decision-makers contribute complementary perspectives to reduce error and improve robustness [12]. In computational settings, structured LLM workflows have been shown to make intermediate reasoning steps explicit and to introduce validation and adjudication mechanisms between agents, thereby increasing the auditability of system behavior and allowing more deliberate control over task execution [6]. These architectural patterns are particularly relevant for safety-sensitive domains such as clinical and diagnostic decision support, where structured multi-step reasoning and cross-checking may offer advantages over single-agent architectures. In these environments, shaping operating characteristics, including tradeoffs among sensitivity, specificity, and precision, is critical, as clinical workflows often prioritize minimizing missed diagnoses even at the expense of increased false positives [7].

Although a growing number of multi-agent LLM frameworks introduce layered, cooperative, or dynamically orchestrated agent structures, most prior research has emphasized architectural design and application-level performance rather than tightly controlled analysis of internal role decomposition [2, 6, 11]. Relatively little attention has been given to isolating prompt-level role structure while holding model weights, decoding temperature, computational budget, and aggregation logic and code fixed. Therefore, it remains unclear how much internal agent specialization or overlapping full-view reasoning alone contributes to observed differences in classification behavior and operating characteristics.

To address this specific gap, we conduct a controlled comparison of two structured multi-agent prompting protocols that differ only in internal role decomposition and patient data access. Both are implemented as deterministic Directed Acyclic Graph (DAG) workflows operating on the UCI Cleveland heart disease dataset [3], with agents running in parallel prior to a downstream adjudication step. In the GD configuration, generalist agents independently evaluate the complete feature set, whereas in the FS configuration, agents are assigned distinct clinically relevant feature domains and produce complementary partial assessments. Intermediate evidence representations are standardized across protocols, and model weights, decoding temperature, computational budget, aggregation logic, and implementation code are held constant. This design isolates role structure and information routing, enabling direct assessment of how internal agent organization shifts classification performance and sensitivity–specificity tradeoffs.

2 Methods

2.1 Dataset

We evaluate both protocols on the UCI Cleveland heart disease dataset [3], a widely used benchmark for binary diagnostic classification. The processed Cleveland subset contains 303 patient records with 13 structured clinical features spanning demographic variables, exercise test results, and imaging-derived indicators. The final dataset comprises $N = 303$ samples, with a positive class prevalence of 0.459 (45.9%).

2.1.1 Preprocessing and Labeling

The original target variable (`num`), which encodes heart disease severity on an ordinal scale from 0 to 4, was converted to a binary outcome. Records with a value of 0 were labeled as `no_disease`, whereas records with values 1 through 4 were labeled as `disease_present`. This binarization follows common practice in prior Cleveland-based diagnostic prediction studies.

Missing values were handled via per-feature mode imputation computed across the processed Cleveland subset. For each variable containing missing observations, the most frequent observed value was used to replace null entries. No explicit missing-value indicators were introduced into the serialized prompts presented to the language model.

The evaluation employed the standard 13-feature Cleveland subset comprising `age`, `sex`, `cp`, `trestbps`, `chol`, `fbs`, `restecg`, `thalach`, `exang`, `oldpeak`, `slope`, `ca`, and `thal`. Each variable was transformed into a structured textual representation while preserving the original categorical encodings and clinically meaningful numeric values as defined in the dataset documentation. A complete mapping of raw variables, encoded values, and LLM-facing feature names is provided in Appendix B.

Zero-shot inference was performed without model fine-tuning. All 303 patient records were evaluated under each experimental protocol using deterministic decoding (temperature = 0). Identical model weights and aggregation logic were applied across both conditions to ensure that observed performance differences were attributable solely to the prompting protocol rather than to stochastic or architectural variation.

2.2 Prompting Protocols

We implement two structured multi-agent prompting protocols that share identical model configuration, decoding parameters, evidence schema, and adjudication logic. Both protocols operate as DAG workflows in which upstream agents generate structured intermediate assessments that are consumed by a downstream adjudicator.

2.2.1 Protocol A: Feature-Specialist Architecture (FS)

The Feature-Specialist (FS) architecture comprises two parallel specialist agents, each restricted to a single clinically defined feature from the patient record. Each specialist produces a structured opinion describing the evidentiary contribution of its assigned feature, including (1) a directional signal (`supports_disease`, `supports_no_disease`, or `unclear`), (2) a categorical strength rating (`low`, `medium`, `high`), and (3) a brief justification note explicitly referencing the feature value. Specialists operate independently and are prohibited from reasoning over features outside their assigned scope.

The two selected features, `chest_pain_type` and `num_major_vessels_fluoroscopy`, were chosen because they reflect different but clinically relevant aspects of cardiac evaluation: symptom presentation and imaging-based evidence of coronary disease. Using one feature from each domain allows the architecture to model separate sources of diagnostic information. This enables assessment of role decomposition effects without selecting features purely to maximize predictive performance.

Formally, given a serialized patient record x, each specialist produces an opinion

$$o_i = (f_i,\ \text{signal}_i,\ \text{strength}_i,\ \text{note}_i),$$

where f_i is the assigned feature name. The adjudicator then receives the full record and both opinions and outputs the final binary prediction:

$$\hat{y} = f_\theta(x, o_1, o_2), \qquad \hat{y} \in \{\texttt{disease_present}, \texttt{no_disease}\},$$

where f_θ denotes a third LLM call under fixed model weights and decoding settings. The adjudicator is instructed to treat strength as ordinal evidence (high > medium > low); when specialist evidence conflicts or is weak/unclear, it may consult the full record to decide. If genuinely uncertain, `no_disease` is used as a conservative tie-break. If any agent fails to produce valid JSON, a conservative fallback output is substituted.

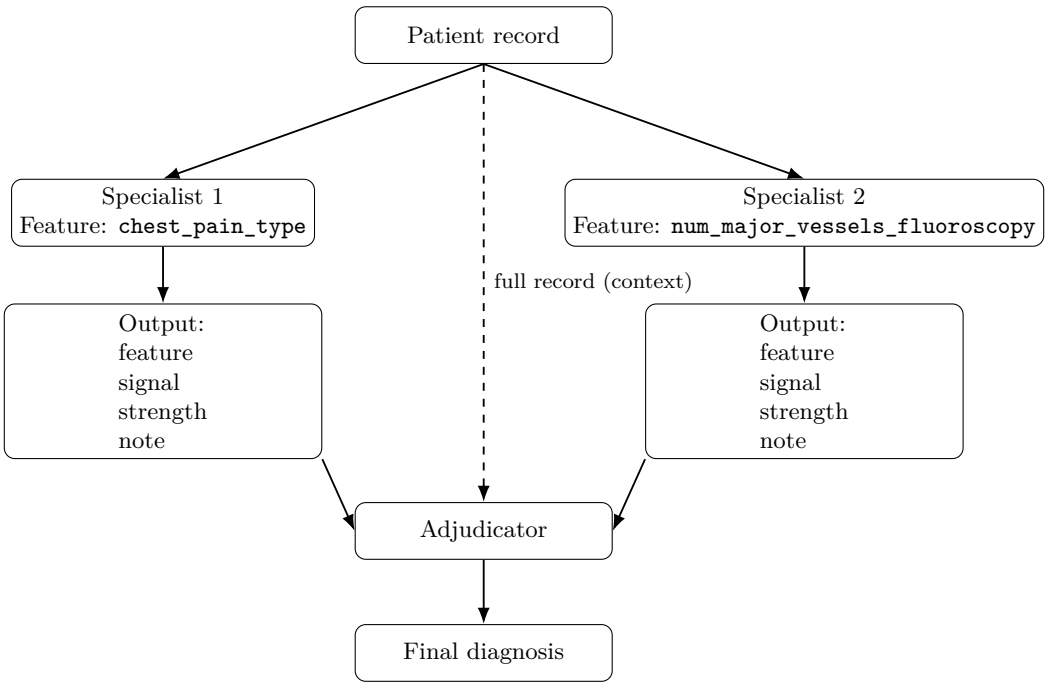

Figure 1: Feature-specialist (FS) architecture. Two agents independently evaluate `chest_pain_type` and `num_major_vessels_fluoroscopy`; their structured outputs are provided to an adjudicator agent, which also receives the full record and produces a binary diagnosis.

2.2.2 Protocol B: Generic Deliberative Architecture (GD)

The Generic Deliberative (GD) architecture comprises two parallel generalist clinician agents, each evaluating the full patient record. Each clinician produces a structured opinion consisting of (1)

a directional signal (`supports_disease`, `supports_no_disease`, or `unclear`), (2) a categorical strength rating (`low`, `medium`, `high`), and (3) a brief justification citing one to two key record-level factors. The clinicians operate independently without inter-agent communication.

Formally, given a serialized patient record x, each clinician produces an opinion

$$o_i = (\texttt{scope}_i,\ \texttt{signal}_i,\ \texttt{strength}_i,\ \texttt{note}_i), \qquad i \in \{1, 2\},$$

where `scope` is set to `overall_record`. The adjudicator then receives the full record and both opinions and outputs the final binary prediction:

$$\hat{y} = f_\theta(x, o_1, o_2), \qquad \hat{y} \in \{\texttt{disease_present}, \texttt{no_disease}\},$$

where f_θ denotes a third LLM call under fixed model weights and decoding settings. The adjudicator is instructed to treat strength as ordinal evidence (high > medium > low); when clinician evidence conflicts or is weak/unclear, it may consult the full record to decide. If genuinely uncertain, `no_disease` is used as a conservative tie-break, and the adjudicator is instructed not to systematically favor either clinician. If any agent fails to produce valid JSON, a conservative fallback output is substituted.

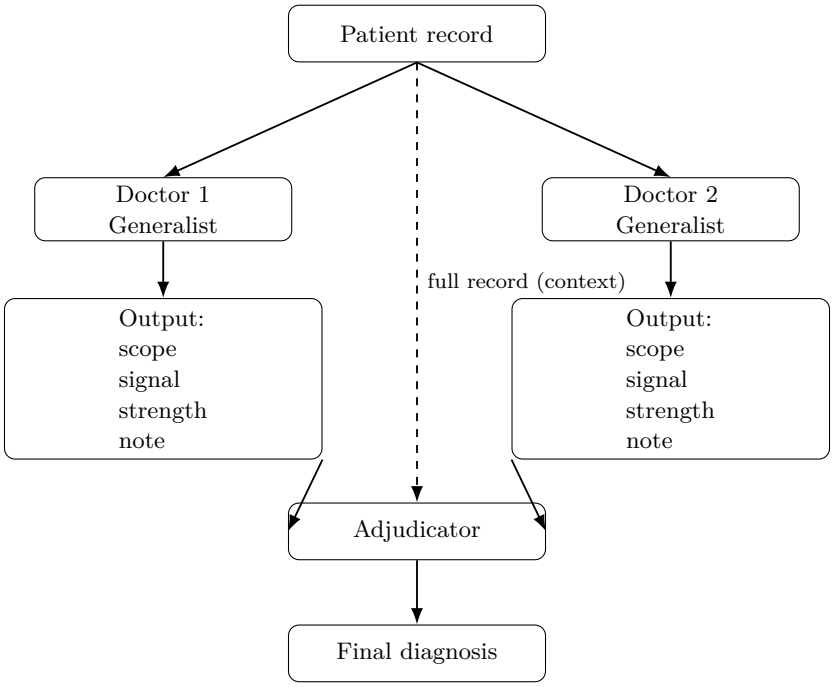

Figure 2: Generic deliberative (GD) architecture. Two generalist clinician agents independently evaluate the full patient record and produce structured opinions; an adjudicator agent, which also receives the full record, produces a binary diagnosis.

2.3 Implementation and Reproducibility

All experiments were implemented in Python 3.13 and executed using JetBrains DataSpell. LLM orchestration was handled via LangChain, with local inference performed through Ollama. All protocols used identical base model weights and decoding parameters to ensure that architectural

structure was the sole manipulated variable.

The underlying language model was `llama3.1:8b` (via Ollama), executed locally with deterministic decoding (temperature = 0). All structured outputs were enforced via strict JSON schema prompts and validated programmatically prior to adjudication.

Experiments were conducted on an Apple MacBook Air (M3) with 16 GB RAM. Core libraries included `ollama` for model inference, `langchain` for multi-agent orchestration and prompt templating, and `pandas` and `numpy` for data preprocessing and evaluation.

All source code, prompt templates, and evaluation scripts are publicly available at:

https://github.com/CallumGA/llm-role-decomposition-clinical-classification

The dataset used in this study is the UCI Heart Disease dataset (Cleveland subset), available at:

https://archive.ics.uci.edu/dataset/45/heart+disease

2.4 Evaluation Metrics

For all evaluated systems, we report standard classification performance metrics, including overall Accuracy as well as macro-averaged Precision, Recall, and F1-score.

Confusion matrices are presented for both the Generic Deliberative (GD) and Feature-Specialist (FS) systems to provide a transparent breakdown of class-wise predictions.

To characterize operating-point behavior, we explicitly report sensitivity–specificity trade-offs. Sensitivity is defined as recall for the `disease_present` class, and specificity as recall for the `no_disease` class.

3 Results

3.1 Overall Performance

Table 1 summarizes overall macro performance across the three protocols.

Table 1: Overall performance on the evaluation set.

System	Accuracy	Precision (Macro)	Recall (Macro)	F1 (Macro)
Single-Agent Protocol	0.69	0.71	0.70	0.69
Generic Deliberative (GD) Protocol	0.65	0.66	0.66	0.65
Feature-Specialist (FS) Protocol	0.72	0.72	0.71	0.71

3.2 Class-wise Performance

Table 2 reports class-wise precision, recall, and F1-score.

Table 2: Class-wise performance.

System	Class	Precision	Recall	F1-score
Single-Agent Protocol	no_disease	0.80	0.57	0.66
	disease_present	0.62	0.83	0.71
Generic Deliberative (GD) Protocol	no_disease	0.72	0.60	0.65
	disease_present	0.60	0.72	0.66
Feature-Specialist (FS) Protocol	no_disease	0.70	0.82	0.76
	disease_present	0.74	0.59	0.66

3.3 Confusion Matrices

Table 3 reports confusion matrix counts (rows = true class, columns = predicted class) for the two multi-agent protocols.

Table 3: Confusion matrices (rows = true class, columns = predicted class).

System	TN	FP	FN	TP
Generic Deliberative (GD) Protocol	98	66	39	100
Feature-Specialist (FS) Protocol	135	29	57	82

4 Discussion

4.1 Interpretation of Principal Findings

The performance differences between the Feature-Specialist (FS) and Generic Deliberative (GD) protocols indicate that internal role structure influences how clinical evidence is combined during inference. Although FS achieves higher overall accuracy (0.72 vs. 0.65) and macro-F1 (0.71 vs. 0.65), these improvements are not uniform across classes and instead reflect a redistribution of classification errors.

As shown in Table 3, FS substantially increases true negatives (135 vs. 98) and reduces false positives (29 vs. 66) relative to GD, indicating improved identification of patients without disease. One possible explanation is that separating feature evaluation constrains how positive evidence is accumulated: because each specialist must independently characterize direction and strength for a single feature, weak or ambiguous signals may be less likely to combine into an overall positive classification. In contrast, the GD protocol evaluates features jointly within a single reasoning process, which may allow multiple modest indicators to interact and collectively shift the prediction toward `disease_present`. The FS structure may therefore function as a more conservative aggregation mechanism, requiring stronger or more consistent evidence before assigning a positive label. This pattern is accompanied by a reduction in true positives (82 vs. 100) and an increase in false negatives (57 vs. 39), reflecting lower sensitivity for disease detection and a corresponding shift toward greater specificity.

In clinical contexts, such trade-offs are operational rather than purely statistical. A reduction in false positives may decrease unnecessary follow-up testing, specialist referral, or patient anxiety,

whereas an increase in false negatives may delay diagnosis or treatment initiation. The balance between these outcomes depends on intended use, such as screening (where sensitivity is typically prioritized) versus confirmatory decision support (where specificity may be emphasized).

Architecturally, the GD protocol evaluates features jointly within a single reasoning trajectory, whereas the FS protocol separates feature evaluation into specialist components prior to aggregation. Importantly, these differences arise despite identical base model weights, decoding parameters, and computational budget, indicating that role decomposition alone can alter clinically relevant operating behavior.

Overall, these findings demonstrate that multi-agent role structure does not simply improve or degrade performance globally. Rather, it changes the balance between sensitivity and specificity, producing systematic differences in classification outcomes. In clinical decision support systems, careful evaluation of this error profile is essential, as false positives and false negatives carry distinct clinical implications and may influence downstream patient management [14].

4.2 Relationship to Prior Work

The observed behavioral differences align with modular design principles in which inputs are routed through distinct components whose outputs are later combined. In sparsely activated mixture-of-experts models, routing determines which submodules contribute to a prediction and thus can change the effective input–output mapping without any weight updates [4]. This supports the broader view that internal organization and information routing can shape decision behavior even when model parameters are fixed.

Evidence from multi-agent prompting similarly indicates that explicit role decomposition can shift performance relative to single-agent baselines. Zunjare and Hsiao report that a three-agent simplification workflow achieved a 70% success rate versus 48% for a single-agent approach on the same dataset, attributing gains to staged role partitioning and validation rather than to changing the underlying model [15]. In multi-agent reinforcement learning, architectural asymmetry can also change outcomes: Kang et al. show that introducing a "special agent" with an independent actor–critic structure inside a value-decomposition framework improves performance in heterogeneous environments [8].

Workflow-level analyses of LLM chaining further emphasize that topology affects controllability, transparency, and output characteristics [6]. However, many prior comparisons confound architectural structure with changes in prompts, model configuration, or training. In contrast, the present study isolates internal role structure as the sole manipulated factor under fixed weights and deterministic decoding, enabling clearer attribution of operating-characteristic shifts to architectural organization.

4.3 Limitations

This study evaluates role-structured prompting on a single benchmark dataset (UCI Cleveland) and reduces the original severity label to a binary outcome, limiting generalizability to other populations and multiclass settings. Given the dataset's age and public availability, prior exposure during pretraining of the underlying model (Llama3:8b) cannot be excluded; however, identical base weights were used across all agents, so any such exposure would affect both protocols equally. All experiments used deterministic decoding (temperature = 0), and the analysis was restricted to a small set of agent configurations with fixed aggregation logic, without calibration or uncertainty

assessment. Results were not externally validated, and robustness should be evaluated on additional cohorts and noisier clinical data.

5 Conclusions and Future Work

This study presented a controlled comparison of two deterministic multi-agent prompting protocols for binary heart disease classification on the UCI Cleveland dataset, differing only in internal role decomposition and information routing while holding model weights, decoding settings, computational budget, and adjudication logic constant. Under these constraints, the Feature-Specialist protocol achieved higher overall accuracy and macro-F1 than the Generic Deliberative protocol, but produced a systematic operating-point shift characterized by increased specificity and reduced sensitivity, demonstrating that role structure can redistribute error types rather than uniformly improve performance. These findings indicate that prompt-level agent organization functions as a meaningful inductive bias in LLM-based clinical classifiers and should be treated as a core design variable. Future work should validate these effects across additional cohorts and structured clinical benchmarks with external testing, extend the comparison to multiclass and higher-dimensional EHR tasks, evaluate probabilistic decoding and calibration behavior, and conduct agent-count and role-ablation studies, ideally followed by limited clinical workflow simulations to assess deployment-relevant trade-offs.

Appendix

A Full Prompt Specifications

This appendix section provides the exact zero-shot prompt specifications, patient record rendering format, and fallback behavior used for all experimental conditions (single-agent, generic multi-agent, and specialist multi-agent). Prompts are reproduced verbatim to ensure full reproducibility.

A.1 Patient Record Serialization

All models received the patient record in the following structured text format:

```
Patient information:
- Age: {age_years}
- Sex: {sex}
- Chest pain type: {chest_pain_type}
- Resting blood pressure: {resting_blood_pressure_mmHg} mmHg
- Serum cholesterol: {serum_cholesterol_mg_dL} mg/dL
- Fasting blood sugar > 120 mg/dL: {fasting_blood_sugar_over_120_mg_dL}
- Resting ECG result: {resting_ecg_result}
- Maximum heart rate achieved: {max_heart_rate_bpm} bpm
- Exercise-induced angina: {exercise_induced_angina}
- ST depression induced by exercise: {st_depression_exercise_relative_to_rest}
- ST segment slope during peak exercise: {st_segment_slope_peak_exercise}
- Number of major vessels (fluoroscopy): {num_major_vessels_fluoroscopy}
- Thalassemia test result: {thalassemia_test_result}
```

A.2 Single-Agent Baseline Prompt

```
You are a clinician performing a binary diagnostic classification task.

You will receive the full patient record.

Your job:
Return the final diagnosis:
- "disease_present" or "no_disease"

Rules (do not mention these in output):
- Consider all relevant factors in the patient record.
- If evidence clearly supports disease, return "disease_present".
- Otherwise return "no_disease".
- If genuinely uncertain, choose "no_disease".

Output MUST be a single JSON object and nothing else.
{ "diagnosis": "disease_present" OR "no_disease" }
```

A.3 Generic Deliberative (GD) Protocol

Doctor Opinion Prompt

```
You are a clinician reviewing the FULL patient record.

Task:
Provide an overall opinion about whether the FULL RECORD supports:
- supports_disease
- supports_no_disease
- unclear

Guidelines (do not mention these in output):
- Use "supports_disease" only if overall evidence clearly supports disease.
- Use "supports_no_disease" only if overall evidence clearly supports no disease.
- Otherwise use "unclear".
- "strength" reflects how strong the overall evidence is (high > medium > low).
- "note" must be ONE short sentence citing 12 key factors from the record (no long reasoning).

Output MUST be a single JSON object and nothing else:
{
  "scope": "overall_record",
  "signal": "supports_disease" OR "supports_no_disease" OR "unclear",
  "strength": "low" OR "medium" OR "high",
  "note": "one short sentence"
}
```

Adjudicator Prompt

```
You are an adjudicator for a binary clinical classification task.

You will receive:
1) The full patient record
2) Two doctor opinions about the full record (signal + strength)

Your job:
Return the final diagnosis:
- "disease_present" or "no_disease"

Rules (do not mention these in output):
- Treat each doctor opinion as evidence about the overall record; it may be low quality or
    wrong.
- Weigh evidence by "strength" (high > medium > low).
- If both doctors strongly support the same direction, follow that direction.
- If evidence conflicts or is weak/unclear, use the full patient record to decide.
- If still genuinely uncertain after considering the full record, choose "no_disease" as a
    tie-break.
- Do not automatically prefer Doctor 1 or Doctor 2.

Output MUST be a single JSON object and nothing else.
{ "diagnosis": "disease_present" OR "no_disease" }
```

A.4 Feature-Specialist (FS) Protocol

Feature Specialist Prompt

```
You are a clinical feature specialist. You ONLY evaluate the feature: "{feature_name}".

Task:
Given the patient record, decide whether THIS FEATURE ALONE:
- supports_disease
- supports_no_disease
- unclear

Guidelines (do not mention these in output):
- Use "supports_disease" only if this feature value is a meaningful risk indicator by itself.
- Use "supports_no_disease" only if this feature value is meaningfully reassuring by itself.
- Otherwise use "unclear".
- "strength" reflects how strongly this single feature points in that direction (not overall
    diagnosis):
  - high: strong / clear signal
  - medium: moderate signal
  - low: weak, borderline, missing, or noisy
- "note" must be ONE short sentence that explicitly cites the patients "{feature_name}" value
    and why it points that way.
- Do NOT use any other features in your reasoning.

Output MUST be a single JSON object and nothing else, with EXACTLY these keys:
```

```
{
  "feature": "{feature_name}",
  "signal": "supports_disease" OR "supports_no_disease" OR "unclear",
  "strength": "low" OR "medium" OR "high",
  "note": "one short sentence referencing the patient's {feature_name} value"
}
```

Specialist Adjudicator Prompt

```
You are an adjudicator for a binary clinical classification task.

You will receive:
1) The full patient record
2) Two specialist opinions (each about ONE feature only)

Your job:
Return the final diagnosis:
- "disease_present" or "no_disease"

Rules (do not mention these in output):
- Treat each specialist opinion as evidence about a single feature; it may be low quality or
    wrong.
- Weigh evidence by "strength" (high > medium > low).
- If both specialists strongly support the same direction, follow that direction.
- If evidence conflicts or is weak/unclear, use the full patient record to decide.
- If still genuinely uncertain after considering the full patient record, choose "no_disease"
    as a tie-break.

Output MUST be a single JSON object and nothing else.
{ "diagnosis": "disease_present" OR "no_disease" }
```

B Full Feature Mapping

The dataset used in this study is the UCI Heart Disease dataset (Cleveland subset) [3]. Feature definitions and categorical encodings follow the official `heart-disease.names` documentation distributed with the dataset download.

Table 4: Full feature schema and categorical encodings for the UCI Heart Disease (Processed Cleveland subset).

#	Orig. Index	Raw Name	Description	Encoded Values / Units	LLM-Friendly Feature Name
1	#3	age	Age of the patient	Years	age_years
2	#4	sex	Biological sex	1 = male; 0 = female	sex
3	#9	cp	Chest pain type	1 = typical angina 2 = atypical angina 3 = non-anginal pain 4 = asymptomatic	chest_pain_type
4	#10	trestbps	Resting blood pressure	mm Hg	resting_blood_pressure_mmHg
5	#12	chol	Serum cholesterol	mg/dL	serum_cholesterol_mg_dL
6	#16	fbs	Fasting blood sugar > 120 mg/dL	1 = true; 0 = false	fasting_blood_sugar_over_120_mg_dL
7	#19	restecg	Resting electrocardiographic results	0 = normal 1 = ST-T wave abnormality 2 = left ventricular hypertrophy	resting_ecg_result
8	#32	thalach	Maximum heart rate achieved	Beats per minute	max_heart_rate_bpm
9	#38	exang	Exercise-induced angina	1 = yes; 0 = no	exercise_induced_angina
10	#40	oldpeak	ST depression induced by exercise relative to rest	Numeric (float)	st_depression_exercise_relative_to_rest
11	#41	slope	Slope of peak exercise ST segment	1 = upsloping 2 = flat 3 = downsloping	st_segment_slope_peak_exercise
12	#44	ca	Number of major vessels colored by fluoroscopy	Integer (0–3)	num_major_vessels_fluoroscopy
13	#51	thal	Thalassemia test result	3 = normal 6 = fixed defect 7 = reversible defect	thalassemia_test_result
14	#58	num	Diagnosis of heart disease (target variable)	0 = no disease 1–4 = disease present (increasing severity)	heart_disease_present (binary derived)

Funding

No external funding was received for this work.

Conflicts of Interest

The author declares no competing interests.

Ethics Statement

This study uses the publicly available UCI Heart Disease dataset, which contains fully de-identified records. No new data were collected and no human participants were recruited. Institutional review board approval was not required.

Data and Code Availability

The dataset used in this study is publicly available from the UCI Machine Learning Repository: https://archive.ics.uci.edu/dataset/45/heart+disease. Source code and prompt templates are available at: https://github.com/CallumGA/llm-role-decomposition-clinical-classification.

References

[1] Alban Bornet, Philipp Khlebnikov, Florian Meer, Quentin Haas, Anthony Yazdani, Boya Zhang, Poorya Amini, and Douglas Teodoro. Analysis of eligibility criteria clusters based on large language models for clinical trial design. *Journal of the American Medical Informatics Association*, 32(3):447–458, 2025. doi: 10.1093/jamia/ocae311.

[2] Y. Chen et al. A multiagent summarization and auto-evaluation framework using large language models. *JMIR AI*, 4:e75932, 2025.

[3] Dheeru Dua and Casey Graff. Uci machine learning repository: Heart disease data set. https://archive.ics.uci.edu/dataset/45/heart+disease, 1988. Accessed 2026.

[4] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23 (120):1–39, 2022.

[5] Deepa Fernandes Prabhu, Varadraj Gurupur, Alexa Stone, and Elizabeth Trader. Integrating artificial intelligence, electronic health records, and wearables for predictive, patient-centered decision support in healthcare. *Healthcare*, 13(21):2753, 2025. doi: 10.3390/healthcare13212753.

[6] Madeleine Grunde-McLaughlin et al. Structured llm chains for reliable clinical reasoning. *Journal of Medical Internet Research*, 26:e55315, 2024. doi: 10.2196/55315.

[7] Vivek Harish, Daniel Cabrera, Daniel Ting, et al. Teaching old tools new tricks: preparing emergency clinicians to appraise machine learning models in clinical practice. *AEM Education and Training*, 7(2), 2023. doi: 10.1002/aet2.10823.

[8] Qitong Kang, Fuyong Wang, Zhongxin Liu, and Zengqiang Chen. Special agents policy gradient in value decomposition-based approach. In *2023 IEEE 12th Data Driven Control and Learning Systems Conference (DDCLS)*. IEEE, 2023. doi: 10.1109/DDCLS58216.2023.10165847.

[9] Andrey Neveditsin et al. A comprehensive review of language models in medicine. *PLOS Digital Health*, 4(5):e0000800, 2025. doi: 10.1371/journal.pdig.0000800.

[10] Jasmine Chiat Ling Ong, Benjamin Jun Jie Seng, Jeren Zheng Feng Law, Lian Leng Low, Andrea Lay Hoon Kwa, Kathleen M. Giacomini, and Daniel Shu Wei Ting. Artificial intelligence,

chatgpt, and other large language models for social determinants of health: current state and future directions. *Cell Reports Medicine*, 5:101356, 2024. doi: 10.1016/j.xcrm.2023.101356.

[11] R. Perera et al. Auto-scaling llm-based multi-agent systems: dynamic architectures for improved collaboration. *Frontiers in Artificial Intelligence*, 2025. doi: 10.3389/frai.2025.1638227.

[12] Robi Polikar. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6(3):21–45, 2006. doi: 10.1109/MCAS.2006.1688199.

[13] Abeed Sarker, Rui Zhang, Yanshan Wang, Yunyu Xiao, Sudeshna Das, Dalton Schutte, David Oniani, Qianqian Xie, and Hua Xu. Natural language processing for digital health in the era of large language models. *IMIA Yearbook of Medical Informatics*, 2024. doi: 10.1055/s-0044-1800750.

[14] Laure Wynants, Ben Van Calster, Gary S. Collins, Richard D. Riley, Georg Heinze, Ewoud Schuit, Marc M. J. Bonten, Johanna A. A. Damen, Thomas P. A. Debray, Maarten De Vos, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ*, 369:m1328, 2020. doi: 10.1136/bmj.m1328.

[15] Pratibha Zunjare and Michael S. Hsiao. A hybrid multi-agent prompting approach for simplifying complex sentences. In *2025 IEEE Smart World Congress (SWC)*. IEEE, 2025. doi: 10.1109/SWC65939.2025.00070.