Leveraging Evidence-Guided LLMs to Enhance Trustworthy Depression Diagnosis

Yining Yuan[1†], J. Ben Tamo[1†], Micky C. Nnamdi[1], Yifei Wang[1], May D. Wang[1]
[1]*Georgia Institute of Technology*, Atlanta, USA
{yyuan394, jtamo3, mnnamdi3, ywang4343, maywang}@gatech.edu

***Abstract*—Large language models (LLMs) show promise in automating clinical diagnosis, yet their non-transparent decision-making and limited alignment with diagnostic standards hinder trust and clinical adoption. We address this challenge by proposing a two-stage diagnostic framework that enhances transparency, trustworthiness, and reliability. First, we introduce evidence-guided diagnostic reasoning (EGDR), which guides LLMs in generating structured diagnostic hypotheses by interleaving evidence extraction and logical reasoning, grounded in DSM-5 criteria. Second, we propose a Diagnosis Confidence Scoring (DCS) module that evaluates the factual accuracy and logical consistency of generated diagnoses through two interpretable metrics: Knowledge Attribution Score (KAS) and Logic Consistency Score (LCS). Evaluated on the D4 dataset with pseudo-labels, EGDR outperforms Direct in-context prompting and Chain-of-Thought (CoT) across five LLMs. For instance, on OpenBioLLM, EGDR improves accuracy from 0.31 (Direct) to 0.76 and DCS from 0.50 to 0.67. On MedLlama, DCS rises from 0.58 (CoT) to 0.77. EGDR yields up to +45% accuracy and +36% DCS gains over baselines, offering a clinically grounded, interpretable foundation for trustworthy AI-assisted diagnosis.**

***Index Terms*—Large Language Models, Diagnostic Reasoning, Mental Health AI, DSM-5, Knowledge Graphs, Explainable AI, Evidence-Guided Reasoning.**

I. Introduction

Depression is a widespread and debilitating mental health disorder [1], affecting approximately 280 million people globally, according to the World Health Organization (WHO), and contributing significantly to the global burden of disease [2]. WHO further estimates that depression and anxiety together cost the global economy over $1 trillion each year in lost productivity, driven by factors such as absenteeism, reduced work performance, and unemployment [2]. These figures underscore the urgent need for improved strategies in early detection and diagnosis of depression, as timely intervention can significantly reduce morbidity and save lives [3].

Recent advances in large language models (LLMs) offer promising tools for analyzing textual data and identifying subtle linguistic markers of psychological distress [4], [5]. However, despite their fluency and general reasoning capabilities, most LLM-based diagnostic tools suffer from a lack of transparency [6]. They often function as black-box classifiers, directly mapping input text to diagnostic labels without providing interpretable justification [7], [8]. This limitation critically undermines clinical reliability, as medical decision-making requires interpretable, evidence-based justification. A recent scoping review in digital pathology emphasizes that one of the major hurdles to safely deploying large language models (LLMs) in medical settings is their limited ability to provide contextualized and interpretable outputs tailored to specific clinical scenarios [9]. In addition, broader literature points out that merely offering diagnostic predictions is often inadequate for clinical use, as the opaque reasoning behind LLM-generated results can weaken clinician trust, which underscore the need for explainable reasoning in diagnostic tools [7].

We propose a two-stage diagnostic pipeline designed to improve the trustworthiness, transparency, and clinical alignment of LLM-generated diagnoses. In the first stage, the Evidence-Guided Diagnostic Reasoning (EGDR) framework guides LLMs to interleave evidence extraction and logical inference, producing structured diagnostic hypotheses. Using multi-turn patient-clinician dialogues and the DSM-5 [10] as sources of medical knowledge, EGDR aligns reasoning with formal diagnostic criteria to enhance interpretability and clinical grounding. We evaluate EGDR using pseudo-labels from a strong baseline model and further validate its performance on MDDial [11], a secondary dataset with gold-standard annotations.

In the second stage, our DCS framework evaluates diagnosis reasoning from first stage via micro-level claim attribution and macro-level reasoning consistency. While prior claim verification methods usually treat claims in isolation [12], our approach jointly assesses the full reasoning chain. DCS combines two components: the Knowledge Attribution Score (KAS), inspired by ClaimVer [13], which checks alignment with DSM-5 triplets, and the Logic Consistency Score (LCS), which verifies rule-based criteria like symptom thresholds and exclusions. Together, they offer a comprehensive, interpretable measure of diagnostic validity, supported by alignment with pseudo-label accuracy and case-level analysis.

II. Related Works

The evolution of clinical diagnostic reasoning has increasingly embraced LLMs, shifting away from rule-based and traditional machine learning approaches towards flexible, generative systems. Prompt engineering, retrieval-augmented generation (RAG), fine-tuning, and domain-specific pre-training have emerged as prominent methods for interpretable diagnoses across a range of medical conditions.

Prompt-based methods use crafted inputs or few-shot ex-

† The first two authors contributed equally to this work.

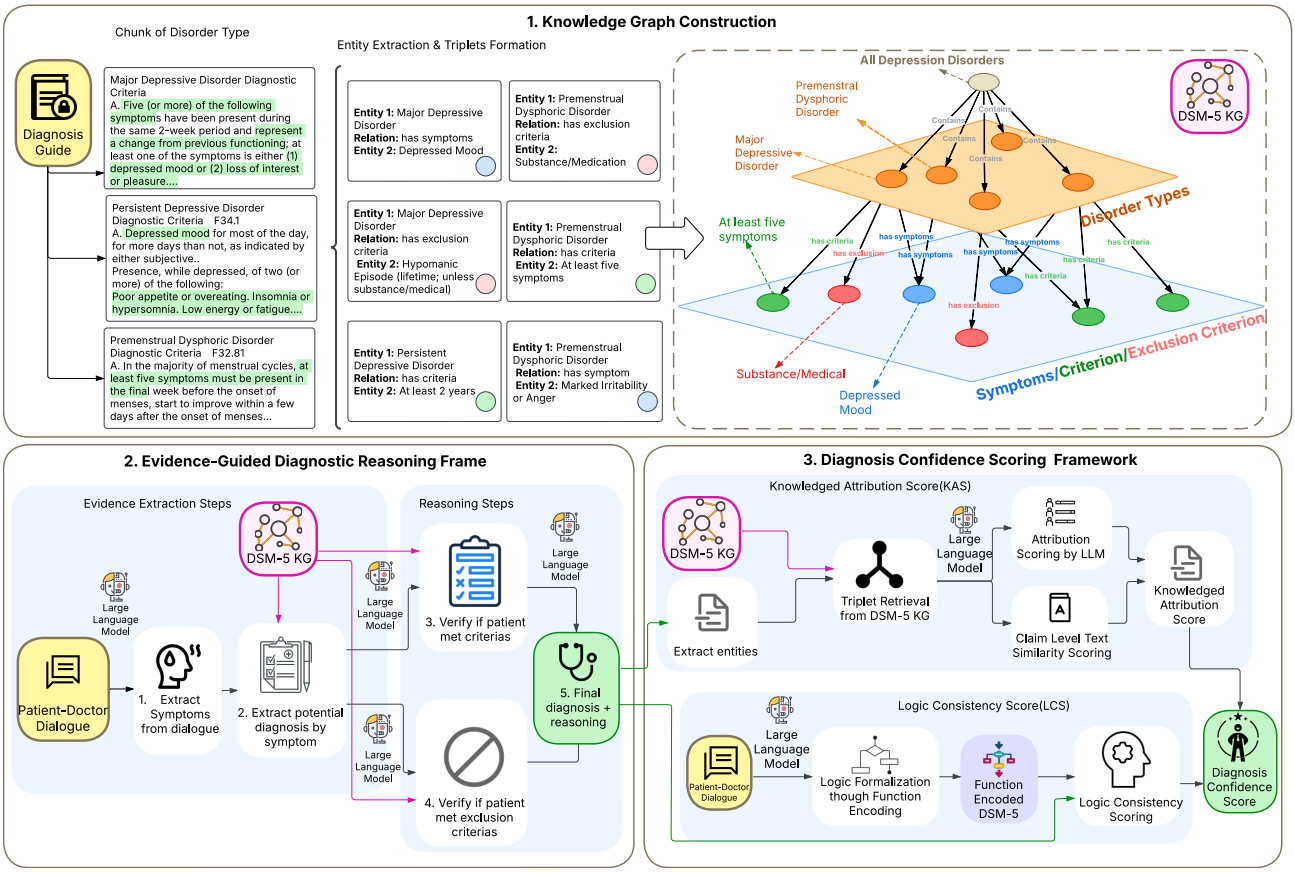

Fig. 1: Overview of the two-stage Evidence-Guided Diagnostic Reasoning (EGDR) and Diagnosis Confidence Scoring (DCS) framework. A DSM-5-based knowledge graph is constructed from diagnostic manuals using medical triplets. Stage 1 (EGDR) processes multi-turn patient dialogues to extract symptoms, retrieve relevant criteria, and generate diagnosis with reasoning. Stage 2 computes two evaluation scores: Knowledge Attribution Score (KAS) for semantic alignment with DSM-5 triplets, and Logic Consistency Score (LCS) for rule-based diagnostic validity. These are aggregated into a final Diagnosis Confidence Score (DCS) ranging from 0 to 1.

emplars to guide reasoning without additional training. Wu et al. [14] introduced Diagnostic-Reasoning Chain-of-Thought (CoT), significantly improving diagnostic accuracy by emulating physician reasoning. Kwon et al. [15] further advanced reasoning-awareness by prompting explicit clinical rationales before diagnoses. Nachane et al. ClinicR [16] applied iterative CoT strategies for open-ended medical questions, achieving superior accuracy.

Fine-tuning LLMs for depression diagnosis specifically has shown notable success. Zhang et al. [17] develop a dialogue-based system combining ChatGLM-6B with a multi-step diagnostic pipeline and knowledge-aware prompting, trained on the D4 dataset to output symptom summaries and risk assessments. ChatGLM-based models focus on severity scoring rather than subtype classification. Li et al. [18] propose DepLLM, a modular mixture of specialized experts approach to handle complex depressive symptom clusters, albeit language-specific and lacking direct DSM-5 interpretability. In parallel, causal reasoning can enhance the trustworthiness of clinical AI. For instance, Tamo et al. [19] show that incorporating causal inference into predictive frameworks improves individualized treatment-effect estimation in high-stakes surgical settings [20].

To improve factual correctness, Wang & Shu [21] introduced FOLK, a First-Order Logic approach that grounds model outputs through symbolic claim decomposition. Liu et al. [22] proposed retrieval-augmented frameworks that integrate structured medical knowledge to reduce hallucinations, while Dmonte et al. [12] surveyed LLM claim-verification methods spanning multi-hop reasoning, semantic entailment, and hybrid symbolic-neural verification. Yet most methods still assess claims in isolation, overlooking global logical consistency. A recent review [23] found that fine-tuned or ensemble transformers, including BERT, RoBERTa, DistilBERT, and GPT-3/4 with methods like prefix-tuning and domain-specific pretraining, consistently outperformed traditional models in detecting, classifying, and managing depression from social media, EHR, and clinical data.

Despite these gains, current systems remain limited. Prompting strategies (CoT, DoT) lack mechanisms for global diagnostic coherence, fine-tuned models focus narrowly, and retrieval or symbolic methods seldom integrate formal diagnostic logic or offer interpretable reasoning. Reliance on surface metrics (accuracy, F1) further constrains evaluation. These gaps highlight the need for hybrid frameworks that couple LLMs with clinical knowledge and symbolic reasoning for more reliable, interpretable depression diagnosis.

III. Methodology

Our method employs a two-stage diagnostic framework as depicted in Figure 1:

1) **Evidence-Guided Diagnostic Reasoning (EGDR)** extracts and structures diagnostic reasoning from patient–doctor dialogues, grounding it in DSM-5 criteria.
2) **Diagnosis confidence scoring** verifies these claims against a DSM-5–based knowledge graph.

Our DSM-5 knowledge graph ($KG_{\text{DSM-5}}$) builds on the general principles of medical knowledge graphs described by Wu et al. [24], with extensions to capture DSM-5-specific entities and relations. We extract entity–relation–entity triplets from DSM-5 text (e.g., "Major Depressive Disorder (MDD) → has symptom → depressed mood"). A root node links all disorders, with directed edges such as "has_criterion", "has_exclusion", and "has_specifier" to capture relationships among disorders, symptoms, and criteria, enabling efficient clinical reasoning.

A. Evidence-Guided Diagnostic Reasoning (EGDR)

Given a multi-turn dialogue between a patient and psychologist/clinician, $D = \{(u_1, r_1), \ldots, (u_n, r_n)\}$, where u_i and r_i are the patient and doctor utterances at turn i.

We prompt an LLM f_θ with dialogue context D using EGDR, Figure 2, which guides the model to produce structured diagnostic output R_{LLM}.

Formally, the model output is structured as:

$$R_{LLM} = f_\theta(\text{prompt}(D, KG_{\text{DSM-5}})) \tag{1}$$

This process yields a semi-structured final diagnosis and diagnostic reasoning C,that consist of analysis on symptoms, criterion, exclusion criterion, and final diagnosis.

B. Diagnosis Confidence Score Calculation

To evaluate the generated diagnosis hypotheses, we propose Diagnosis Confidence Scoring (DCS), which decomposes into two components, knowledge attribution score (KAS) and logic consistency score (LCS).

a) Knowledge Attribution Score (KAS)

KAS measures the factual alignment of individual diagnostic claims with the DSM-5 knowledge graph $KG_{\text{DSM-5}}$, combining symbolic matching and neural semantic similarity to assess whether each claim is clinically valid. This design builds on prior work in natural language claim verification and medical fact attribution [13], adapting it to the diagnostic setting where claims must be grounded in medical guidelines.

In $KG_{\text{DSM-5}}$, each diagnosis represents a central node connected to nodes representing symptoms, exclusion rules, specifiers, and modifiers. Edges encode semantic relationships such as "has symptom," "has exclusion," or "has criteria," allowing graph-based retrieval of relevant knowledge for each claim.

The KAS pipeline consist of:

- **Entity Extraction and Triplet Retrieval**. We extract psychiatric entities from the reasoning passage R using medical named entity recognition (NER), then retrieve relevant triplets $T(R) \subset KG_{\text{DSM-5}}$ via WoolNet [25], a graph-walk strategy that incrementally traverses clinically relevant subgraphs using a hybrid of symbolic reasoning and neural relevance scoring, enhancing both interpretability and reasoning accuracy in LLM-based diagnosis.
- **Claim Decomposition and Verification**. Diagnosis reasoning R is decomposed into atomic, verifiable claims $\{c_1, c_2, \ldots, c_n\}$ using an LLM. Each claim c_i is evaluated for factual alignment by comparing it against the retrieved DSM-5 triplets $T(R) \subset KG_{\text{DSM-5}}$ from last step.
- **Triplet-Based Attribution Classification**. Each claim c_i is classified based on its symbolic alignment with retrieved triplets $T(R) \subset KG_{\text{DSM-5}}$, and assigned a symbolic score $cs(c_i)$ as follows:

$$cs(c_i) = \begin{cases} 2 & \text{Attributable} \\ 1 & \text{Extrapolatory} \\ -1 & \text{Contradictory} \\ 0 & \text{No Attribution} \end{cases}$$

- **Triplets Match Score (TMS)**. Each claim c_i is also assigned a Triplets Match Score :

$$TMS_i = \alpha \cdot \text{sim}(c_i, T(R)) + (1 - \alpha) \cdot EPR_i \tag{2}$$

where $\text{sim}(c_i, T(R))$ is the cosine similarity between the claim and retrieved triplets using a BERT-based encoder, EPR_i the entity-level precision/recall between c_i and matched DSM entities, and $\alpha \in [0, 1]$ is a weighting hyperparameter, we set it to 0.5, through analyzing TMS and EPR distribution and experimenting a set of configuration and finding a suitable trade-off balance between TMS and EPR.

- **Claim Score Composition**. A weighted trust score for each claim by combining symbolic classification and semantic similarity:

$$w_i = cs(y_i) \cdot TMS_i \tag{3}$$

- **KAS Aggregation.** We apply a sigmoid function that adjusts sensitivity based on the aggregate evidence weight. This allows the KAS to effectively distinguish flawed reasoning from valid justifications.

$$KAS = \sigma\left(\sum_{i=1}^{n} w_i\right), \quad \text{where } \sigma(x) = \frac{1}{1 + e^{-x}} \tag{4}$$

b) Logic Consistency Score (LCS)

While KAS evaluates the factual accuracy of individual claims, LCS assesses whether the overall reasoning aligns with the full diagnostic logic of the DSM-5. Each DSM-5 disorder criterion is encoded as an executable function that explicitly captures the required symptom count, presence of core symptoms, and any exclusionary conditions. The LLM is prompted to execute a step-wise application of these diagnostic rules over its reasoning trace. The LCS score is then assigned

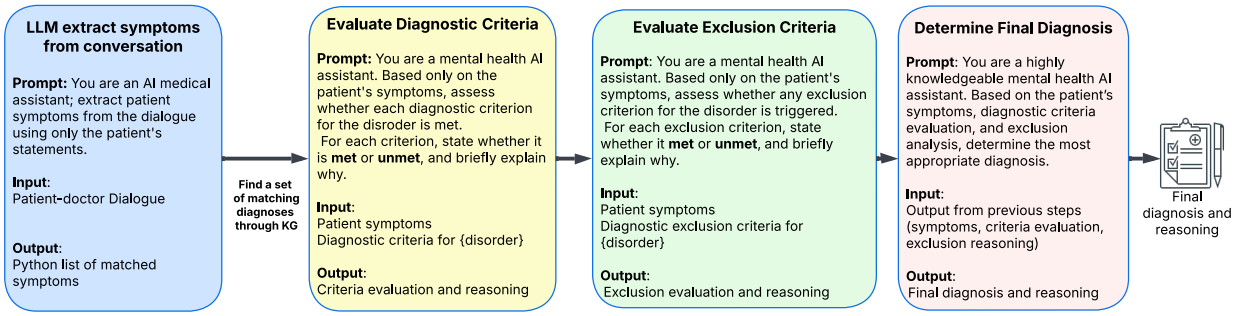

Fig. 2: EGDR Framework Overview: (1) An AI assistant extracts patient symptoms from doctor-patient dialogue; (2) Matches symptoms to top-3 candidate disorders via the DSM-5 knowledge graph; (3) Evaluates whether diagnostic criteria are met; (4) Checks exclusion criteria; (5) Determines the final diagnosis and provides reasoning.

based on how closely the reasoning conforms to the structured diagnostic criteria.

$$\text{LCS} = \begin{cases} 0, & \text{Contradicts DSM-5 diagnostic logic} \\ 1, & \text{Partially correct logic, incorrect conclusion} \\ 2, & \text{Partially correct logic, correct conclusion} \\ 3, & \text{Fully correct logic and conclusion} \end{cases}$$

The final DCS combines KAS and LCS, using a normalized LCS score scaled to $[0, 1]$:

$$DCS = \lambda \cdot KAS + (1 - \lambda) \cdot \left(\frac{LCS}{3}\right), \quad \text{where } \lambda \in [0, 1] \quad (5)$$

This unified score provides a normalized, interpretable measure of diagnostic reliability from 0 (invalid) to 1 (fully valid).

IV. Experiments

We evaluate our framework across a range of LLMs to assess diagnostic quality and reasoning fidelity in depression diagnosis and suicide risk assessment. To isolate the effect of our clinical evidence-guided reasoning pipeline, we include both domain-specific and general-purpose models. Experiments were conducted using a mix of local and cloud resources: local evaluations ran on dual NVIDIA A100 GPUs, while API-based models (e.g., Claude, GPT-4o-mini, Gemini) were accessed as available.

We begin by benchmarking all models on the Dialogue Dataset for Depression-Diagnosis (D4) [26], using a four-class depression and suicide risk scoring task to establish baseline performance and identify the top model for generating diagnostic pseudo-labels. This model is then used to create silver-standard labels for the D4 dataset by applying DSM-5 criteria to ground truth symptom annotations, which serve as evaluation targets for all models. In addition to the D4 dataset, to further demonstrate the effectiveness of our EGDR reasoning pipeline, we conducted experiments on the MDDial dataset [11], which includes a different set of diagnoses and gold-standard labels.

To assess the impact of our reasoning pipeline, we compare it against two prompting baselines: (1) direct instruction prompting, which elicits structured diagnostic outputs, and (2) chain-of-thought prompting, which encourages stepwise reasoning. All approaches include DSM-5 criteria, but our pipeline uniquely integrates them in a structured, diagnostic-oriented manner, systematically grounding reasoning in the formal framework. Baselines use the same information but with less structured guidance.

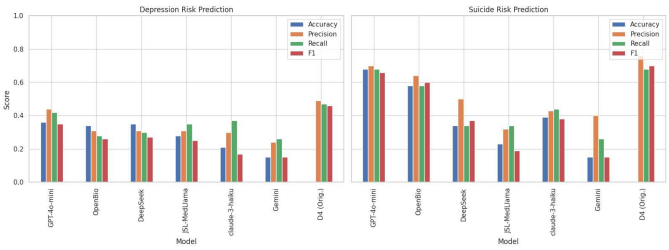

Fig. 3: Depression and suicidal risk prediction result. GPT-4o-mini performance approaches D4 baseline

Finally, we apply our DCS framework to evaluate factual accuracy and logical validity, analyzing alignment with pseudo-labels, per-case reasoning quality, and the contribution of individual scoring components via ablation. These results demonstrate the robustness and interpretability of the DCS method.

A. Dataset

We use the D4 dataset [26], which contains 1,339 annotated clinician–patient dialogues focused on mental health evaluation, including expert-labeled symptom mentions, depression severity, and suicide risk. The dataset is slightly imbalanced, with a risk score of 0 dominating both tasks. These annotations enable supervised evaluation of both diagnostic performance and reasoning quality.

Based on benchmarking results (Figure 3), GPT-4o-mini was selected as the pseudo-label generator for downstream tasks due to its strong alignment with expert risk scores and superior performance among evaluated models.

Alongside evaluating the D4 dataset, we use the MDDial dataset [11], which comprises 235 patient dialogue instances in its test set. This dataset offers authoritative labels across a variety of diagnoses, such as Esophagitis, Enteritis, Asthma, Coronary Heart Disease, Pneumonia, Rhinitis, Thyroiditis, Traumatic Brain Injury, Dermatitis, External Otitis, Conjunctivitis, and Mastitis. We assessed a selection of models on this dataset, illustrating the versatility and robustness of our proposed framework in handling both general and disease-specific clinical dialogue tasks.

TABLE I: Performance by Model and Prompting Method evaluated on GPT-4o-mini pseudo labels. DCS score represents the average Diagnosis Confidence Score.

Model	Prompting	Acc	Prec	Rec	F1	DCS
MedLlama [27]	EGDR	0.68	0.79	0.68	0.73	0.77
	Direct	0.65	0.71	0.83	0.77	0.72
	CoT	0.69	0.85	0.69	0.70	0.58
Claude [28]	EGDR	0.77	0.83	0.77	0.80	0.73
	Direct	0.75	0.82	0.75	0.78	0.69
	CoT	0.70	0.91	0.70	0.78	0.59
DeepSeek [29]	EGDR	0.29	0.86	0.29	0.30	0.56
	Direct	0.21	0.86	0.21	0.32	0.44
	CoT	0.11	0.76	0.12	0.20	0.31
OpenBioLLM [30]	EGDR	0.76	0.81	0.76	0.75	0.67
	Direct	0.31	0.77	0.31	0.44	0.50
	CoT	0.46	0.84	0.46	0.59	0.50
Gemini [31]	EGDR	0.88	0.85	0.86	0.86	0.77
	Direct	0.79	0.81	0.79	0.80	0.70
	CoT	0.76	0.85	0.76	0.80	0.74

TABLE II: Performance of different models and prompting methods on MDDial dataset.

Model	Prompting	Accuracy	Precision	Recall	F1 Score
Claude	EGDR	0.69	0.77	0.69	0.71
	Direct	0.57	0.45	0.46	0.46
	CoT	0.57	0.55	0.56	0.56
Gemini	EGDR	0.75	0.74	0.74	0.75
	Direct	0.71	0.71	0.72	0.72
	CoT	0.69	0.36	0.31	0.33
DeepSeek	EGDR	0.48	0.15	0.16	0.16
	Direct	0.21	0.15	0.17	0.17
	CoT	0.41	0.15	0.18	0.18

B. Diagnosis, Evaluation, and Confidence Score

As shown in Table I, the evidence-guided reasoning prompting consistently outperforms both Direct and Chain-of-Thought (CoT) prompting across most evaluation metrics, including accuracy, recall, and F1 score across all models. The most substantial improvement is observed in OpenBioLLM, where EGDR achieves a 0.45 gain in accuracy over Direct prompting (0.76 vs. 0.31). Notably, even strong general-purpose models like Claude and Gemini benefit from EGDR, with Gemini reaching the highest accuracy overall (0.88) under EGDR. These results demonstrate that integrating structured clinical evidence and diagnostic reasoning steps significantly enhances classification performance, regardless of model domain specialization.

On the MDDial dataset (Table II), EGDR similarly outperforms both Direct and CoT prompting across most models. Gemini shows the highest overall performance with EGDR (Accuracy: 0.75, F1: 0.75), while Claude also benefits significantly (F1: 0.71 with EGDR vs. 0.46 with Direct). Even with lower-performing models like DeepSeek, EGDR consistently improves results over baseline prompting strategies. These findings reinforce EGDR's robustness and adaptability in handling complex, multi-turn clinical dialogues.

Our analysis further demonstrates that DCS functions effectively as a diagnostic confidence metric. As illustrated in Figure 4, DCS scores for correct cases are consistently higher across prompting strategies, while incorrect cases receive markedly lower scores, indicating strong diagnostic confidence.

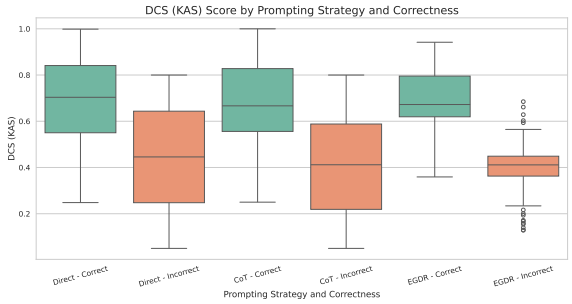

Fig. 4: Diagnosis Confidence Scores (DCS) by correctness under different prompting strategies.

Notably, EGDR yields a more concentrated distribution of DCS values for correct and partially correct outputs, with scores tightly clustered around higher values. This suggests that EGDR improves reasoning structure and calibrates confidence more reliably. In contrast, Direct and CoT prompting tend to produce more vague or unsupported reasoning, resulting in wider and more dispersed DCS distributions with greater overlap between correct and incorrect cases.

C. Fairness Analysis Across Age and Gender

TABLE III: Average accuracy across all models and prompting methods by demographic group.

	≤17	18–25	26–35	36–45	46–60	60+	Men	Women
Avg Acc	0.635	0.616	0.558	0.496	0.495	0.364	0.528	0.578

To assess the fairness of large language model (LLM) reasoning across demographic groups, we conducted a subgroup performance analysis by age and gender. The dataset itself is notably imbalanced. In terms of age, the 18–25 group dominates with 685 samples, followed by 26–35 (307), 36–45 (138), 46–60 (105), ≤17 (100), and 60+ (4). Gender distribution is similarly skewed, with 982 female and 357 male samples.

Because the LLM was not trained or fine-tuned on this dataset, dataset imbalances do not affect the model parameters directly. Instead, the subgroup disparities observed in performance reflect two main factors: (1) evaluation reliability—performance estimates are more stable in larger subgroups (e.g., 18–25), while very small groups such as 60+ yield volatile results where a few errors can drastically change accuracy; and (2) potential pretraining and reasoning biases within the LLM itself, which may influence responses differently across demographic attributes.

Overall, models show consistently higher accuracy for the majority subgroups (18–25, 26–35, and female participants), while smaller or minority groups exhibit lower or more variable performance, shown in Table III. This highlights the importance of cautious interpretation of fairness metrics in zero-shot LLM settings, where disparities may arise from a combination of intrinsic model biases and uneven group representation in the evaluation data.

D. Impact of α, and λ on DCS

We performed an ablation study, Table IV, to explore how two key parameters, α and λ, shape the behavior and interpretability of the Diagnostic Consistency Score (DCS). Rather than focus-

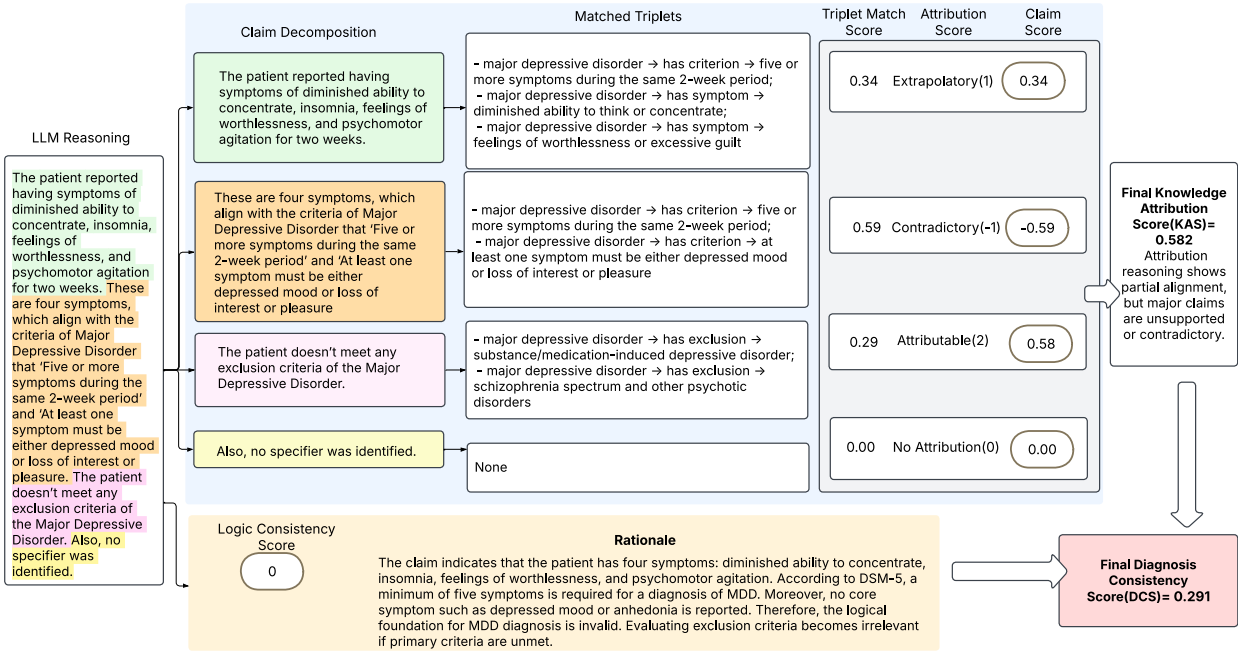

Fig. 5: DCS evaluation of diagnostic reasoning for an invalid MDD case. The top section shows a Knowledge Attribution Score (KAS) breakdown, revealing mostly weak alignment with DSM-5 diagnostic knowledge. The bottom section presents the logic consistency evaluation, which scores 0.0 due to missing core symptoms and insufficient symptom count. Together, these produce a low Diagnostic Consistency Score (**DCS = 0.291**).

TABLE IV: Ablation study showing how variations in α and λ affect the distribution of DCS. α controls the trade-off between semantic similarity and entity-level precision/recall in the Triplet Match Score (TMS), while λ balances KAS and LCS in the final DCS. The table reports summary statistics across reasoning outputs, highlighting the impact of each parameter on scoring behavior and stability.

Param	Value	Mean	Std Dev	Min	25%	Median	75%
α	0.00	0.3517	0.2790	0.0085	0.1471	0.2171	0.6422
	0.25	0.7454	0.2274	0.0083	0.7510	0.8153	0.8768
	0.50	0.8462	0.1160	0.0258	0.8219	0.8534	0.9197
	0.75	0.8653	0.1129	0.0094	0.8459	0.8614	0.9331
	1.00	0.8811	0.0781	0.2583	0.8559	0.8623	0.9362
λ	0.00	0.5700	0.2726	0.0000	0.4500	0.4500	0.7500
	0.25	0.6621	0.2069	0.0086	0.5743	0.5859	0.8074
	0.50	0.7541	0.1501	0.0172	0.6981	0.7215	0.8648
	0.75	0.8462	0.1160	0.0258	0.8219	0.8534	0.9197
	1.00	0.9382	0.1250	0.0152	0.9409	0.9600	0.9820

ing solely on numerical performance, our analysis highlights how these parameters modulate core trade-offs in semantic attribution and reasoning fidelity.

a) Alpha α – Balancing Semantic Similarity and Entity-Level Precision/Recall in TMS.

The α parameter controls the balance in the Triplet Match Score (TMS) between semantic similarity (SS) and entity-level precision/recall (EPR). Higher α emphasizes SS, which captures embedding-level closeness between claims and retrieved knowledge triplets. However, strong SS alone does not ensure that a claim is attributable (A) or extrapolating (E); it may still reflect contradiction (C) or no attribution (N). In contrast, EPR enforces stricter, symbolic alignment at the level of clinical entities, but may miss paraphrased or inferred matches.

Thus, α controls how sensitive TMS is to surface-level fluency versus grounded clinical precision. High α values lead to higher TMS and trickle into higher KAS and DCS scores, but this can result in overconfident scoring for plausible-sounding but incorrect claims. Lower α brings EPR to the forefront, reducing false positives but potentially penalizing semantically valid variation. We choose $\alpha = 0.5$ to maintain a balance: enough semantic flexibility to reward paraphrased reasoning, but enough entity grounding to detect extrapolation or misattribution.

b) Lambda λ – Balancing KAS and LCS.

The λ parameter controls the weighting between KAS (claim-level attribution to retrieved knowledge) and LCS (global diagnostic logic consistency) in the final DCS score. High λ values prioritize KAS, rewarding models that justify individual claims well, even when their overall diagnostic reasoning is flawed. For instance, at $\lambda = 1$, the mean DCS reaches 0.9382, but this can reflect locally plausible yet globally inconsistent outputs. Conversely, low λ values favor LCS, emphasizing logical coherence. At $\lambda = 0$, the mean DCS drops to 0.57, showing that LCS alone is brittle. It penalizes valid reasoning that deviates from expected rule structures and underweights attribution.

To balance attribution fidelity and logical soundness, we set $\lambda = 0.75$, which achieves a strong mean DCS of 0.8462 with low variance. This choice helps prevent models from being rewarded for individually plausible claims that, when combined, fail to justify the overall diagnosis.

E. Case Study: Low Diagnosis Confidence Scoring

The case study presented in Figure 5 demonstrates the effectiveness and interpretability of the DCS. In the low confidence case, the DCS score of 0.291 reflects both limited knowledge grounding (KAS = 0.582) and a failed symbolic logic check (Logic Score = 0.0). Although some symptoms are recognized, the claims do not meet the MDD threshold, with only four

symptoms and no core symptom identified. This breaks a core diagnostic requirement outlined in the DSM-5, making the reasoning incomplete. This inconsistency is accurately penalized in both the semantic and logic evaluations. By aligning both structured diagnostic knowledge and rule-based logic, DCS ensures that the diagnosis is grounded not only in claim correctness but also in rigorous diagnostic logic.

V. Conclusion

This work proposes a unified two-step framework for LLM-based diagnostic reasoning: (1) the Evidence-Guided Diagnostic Reasoning (EGDR) pipeline, which enhances output quality through clinically grounded prompting, and (2) the Diagnosis Confidence Score (DCS), a dual-layered metric evaluating both factual alignment and diagnostic logic. Experiments on the D4 and MDDial datasets show that EGDR consistently improves diagnostic performance across diverse models. The best-performing model, Gemini-2.5-flash, achieved an 8% accuracy improvement on D4 and a 4% accuracy improvement on MDDial with EGDR. Our DCS evaluation also demonstrates strong alignment with silver-standard labels on D4, and case analyses confirm its ability to generate clinically coherent and well-grounded reasoning. Together, EGDR and DCS advance the transparency, reliability, and clinical fidelity of AI-assisted diagnosis.

Acknowledgement

This research was supported in part through research cyberinfrastructure resources and services, including the AI Makerspace of the College of Engineering, provided by the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology, Atlanta, Georgia, USA. We also gratefully acknowledge funding and fellowships that contributed to this work, including a Wallace H. Coulter Distinguished Faculty Fellowship, a Petit Institute Faculty Fellowship, and research funding from Amazon and Microsoft Research awarded to Professor May D. Wang.

References

[1] J.-P. Lépine and M. Briley, "The increasing burden of depression," *Neuropsychiatric disease and treatment*, vol. 7, no. sup1, pp. 3–7, 2011.

[2] World Health Organization, "Depressive disorder (depression)," https://www.who.int/news-room/fact-sheets/detail/depression, 2023, accessed June 2025.

[3] K. A. Siniscalchi, M. E. Broome, J. Fish, J. Ventimiglia, J. Thompson, P. Roy, R. Pipes, and M. Trivedi, "Depression screening and measurement-based care in primary care," *Journal of primary care & community health*, vol. 11, p. 2150132720931261, 2020.

[4] M. Y. Lu, B. Chen, D. F. Williamson, R. J. Chen, M. Zhao, A. K. Chow, K. Ikemura, A. Kim, D. Pouli, A. Patel *et al.*, "A multimodal generative ai copilot for human pathology," *Nature*, vol. 634, no. 8033, pp. 466–473, 2024.

[5] J. Kim, K. G. Leonte, M. L. Chen, J. B. Torous, E. Linos, A. Pinto, and C. I. Rodriguez, "Large language models outperform mental and medical health care professionals in identifying obsessive-compulsive disorder," *NPJ Digital Medicine*, vol. 7, no. 1, p. 193, 2024.

[6] H. R. Lawrence, R. A. Schneider, S. B. Rubin, M. J. Matarić, D. J. McDuff, and M. J. Bell, "The opportunities and risks of large language models in mental health," *JMIR Mental Health*, vol. 11, no. 1, p. e59479, 2024.

[7] S. Zhou, Z. Xu, M. Zhang, C. Xu, Y. Guo, Z. Zhan, Y. Fang, S. Ding, J. Wang, K. Xu *et al.*, "Large language models for disease diagnosis: A scoping review," *npj Artificial Intelligence*, vol. 1, no. 1, pp. 1–17, 2025.

[8] T. Savage, A. Nayak, R. Gallo, E. Rangan, and J. H. Chen, "Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine," *NPJ Digital Medicine*, vol. 7, no. 1, p. 20, 2024.

[9] E. Ullah, A. Parwani, M. M. Baig, and R. Singh, "Challenges and barriers of using large language models (llm) such as chatgpt for diagnostic medicine with a focus on digital pathology–a recent scoping review," *Diagnostic pathology*, vol. 19, no. 1, p. 43, 2024.

[10] M. Guha, "Diagnostic and statistical manual of mental disorders: Dsm-5," *Reference Reviews*, vol. 28, no. 3, pp. 36–37, 2014.

[11] S. Macherla, M. Luo, M. Parmar, and C. Baral, "Mddial: A multi-turn differential diagnosis dialogue dataset with reliability evaluation," *arXiv preprint arXiv:2308.08147*, 2023.

[12] A. Dmonte, R. Oruche, M. Zampieri, P. Calyam, and I. Augenstein, "Claim verification in the age of large language models: A survey," *arXiv preprint arXiv:2408.14317*, 2024.

[13] P. P. S. Dammu, H. Naidu, M. Dewan, Y. Kim, T. Roosta, A. Chadha, and C. Shah, "Claimver: Explainable claim-level verification and evidence attribution of text through knowledge graphs," *arXiv preprint arXiv:2403.09724*, 2024.

[14] C.-K. Wu, W.-L. Chen, and H.-H. Chen, "Large language models perform diagnostic reasoning," *arXiv preprint arXiv:2307.08922*, 2023.

[15] T. Kwon, K. T.-i. Ong, D. Kang, S. Moon, J. R. Lee, D. Hwang, B. Sohn, Y. Sim, D. Lee, and J. Yeo, "Large language models are clinical reasoners: Reasoning-aware diagnosis framework with prompt-generated rationales," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 16, 2024, pp. 18 417–18 425.

[16] S. Sandeep Nachane, O. Gramopadhye, P. Chanda, G. Ramakrishnan, K. Sharad Jadhav, Y. Nandwani, D. Raghu, and S. Joshi, "Few shot chain-of-thought driven reasoning to prompt llms for open ended medical question answering," *arXiv e-prints*, pp. arXiv–2403, 2024.

[17] X. Zhang, W. Cui, J. Wang, and Y. Li, "Chat, summary and diagnosis: A llm-enhanced conversational agent for interactive depression detection," in *2024 4th International Conference on Industrial Automation, Robotics and Control Engineering (IARCE)*. IEEE, 2024, pp. 343–348.

[18] Y. Li, W. Zhang, Y. Sun *et al.*, "Depllm: Fine-tuning large language models with a chinese dialogue dataset for depression diagnosis via mixture of specialized experts," *arXiv preprint arXiv:2405.12345*, 2024.

[19] J. B. Tamo, N. S. Chouhan, M. C. Nnamdi, Y. Yuan, S. S. Chivilkar, W. Shi, S. W. Hwang, B. R. Brenn, and M. D. Wang, "Causal machine learning for surgical interventions," *arXiv preprint arXiv:2509.19705*, 2025.

[20] J. B. Tamo, M. C. Nnamdi, A. Hornback, M. Chen, W. Shi, Y. Zhu, H. J. Iwinski, and M. D. Wang, "Heterogeneous treatment effects of spinal fusion surgery for adolescent idiopathic scoliosis patients," in *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 2024, pp. 1–10.

[21] H. Wang and K. Shu, "Explainable claim verification via knowledge-grounded reasoning with large language models," *arXiv preprint arXiv:2310.05253*, 2023.

[22] F. Liu, B. Yang, C. You, X. Wu, S. Ge, Z. Liu, X. Sun, Y. Yang, and D. A. Clifton, "Retrieval-augmented and knowledge-grounded language models for faithful clinical medicine," *arXiv preprint arXiv:2210.12777*, 2022.

[23] M. Omar and I. Levkovich, "Exploring the efficacy and potential of large language models for depression: A systematic review," *Journal of Affective Disorders*, vol. 371, pp. 234–244, 2025.

[24] J. Wu, J. Zhu, Y. Qi, J. Chen, M. Xu, F. Menolascina, Y. Jin, and V. Grau, "Medical graph rag: Evidence-based medical large language model via graph retrieval-augmented generation," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 28 443–28 467.

[25] C. T. Gutiérrez and A. Hogan, "Woolnet: Finding and visualising paths in knowledge graphs," 2024.

[26] H. Li, W. Shi, Q. Zhong, L. Hou, D. Song, and Y. Li, "D4: A generative dialogue system for differential diagnosis from depressive symptoms," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

[27] J. S. Labs, "Jsl-medllama-3-8b-v2.0," https://huggingface.co/johnsnowlabs/JSL-MedLlama-3-8B-v2.0, 2024.

[28] Anthropic, "Claude 3 haiku," https://www.anthropic.com/news/claude-3-family, 2024.

[29] D. AI, "Deepseek llm 7b chat," https://huggingface.co/deepseek-ai/deepseek-llm-7b-chat, 2024.

[30] A. Singh and contributors, "Openbiollm-llama3-70b," https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B, 2024.

[31] G. DeepMind, "Gemini 2.5 flash," https://deepmind.google/technologies/gemini/, 2024.