DiaQwen: A Knowledge-Enhanced LLM Framework for Privacy-Preserving Diabetes Diagnosis

Anonymous Submission

Abstract

Diabetes remains a leading chronic disease worldwide, yet applying large language models (LLMs) to its diagnosis faces four barriers: insufficient domain knowledge, limited multi-turn reasoning, inability to deploy locally under privacy constraints, and high hallucination rates. We present **DiaQwen**, a privacy-preserving diagnostic system built on Qwen-4B that runs on a single consumer GPU. DiaQwen integrates three innovations: (1) **KBase**, a diabetes knowledge base encoding over 2,000 clinical triplets from authoritative guidelines, EHRs, and public QA datasets; (2) **HyTrain**, a hybrid fine-tuning strategy combining LoRA with epoch-level self-distillation; and (3) **DyQA**, a dynamic consultation module navigating a diagnostic graph via A* search. On 2,000 test cases, DiaQwen achieves 93.4% accuracy with 0.7% hallucination rate, surpassing baselines by over 4 points while producing guideline-aligned explanations (BLEU=0.72, ROUGE-L=0.96).

1 Introduction

According to the International Diabetes Federation, 537 million adults live with diabetes, projected to exceed 783 million by 2045 [International Diabetes Federation, 2024]. Complications—cardiovascular disease, chronic kidney disease (CKD), and retinopathy—account for over 12% of total U.S. medical expenditures. Accurate diagnosis requires integrating symptoms, lab results, family history, and lifestyle factors [Chao et al., 2021], facing three obstacles: *information fragmentation* across clinical systems, the *efficiency–accuracy tradeoff*, and *temporal dynamics* of progressive markers like HbA1c.

Type 2 diabetes (T2DM) involves insulin resistance entangled with obesity, hypertension, and dyslipidemia. CKD risk escalates when HbA1c exceeds 7.5% [Sharma et al., 2021]. Manual record processing at 40 words/minute remains error-prone [Yi et al., 2023], and physicians' reliance on intuitive pattern recognition introduces subjectivity—e.g., overlooking C-peptide results that distinguish type 1 from type 2 diabetes.

LLMs have shown clinical reasoning capabilities [Briganti and Le Moine, 2020, OpenAI et al., 2024] but face four deployment bottlenecks: (1) lack of deep domain knowledge causing plausible but inaccurate responses; (2) limited multi-turn reasoning; (3) compute requirements exceeding local hospital budgets, conflicting with HIPAA/GDPR; (4) untraceable outputs with up to 20% hallucination [OpenAI et al., 2024]. Rule-based predecessors like MYCIN [Shortliffe, 1976] could not scale beyond fixed mappings.

Table 1: Comparison of approaches for diabetes diagnostic support.

Approach	Know.	Multi-turn	Privacy	Explain.
Rule-based CDSS	✓	×	✓	✓
Generic LLM (GPT-4)	×	±	×	×
Medical LLM (7B)	±	×	×	±
DiaQwen (Ours)	✓	✓	✓	✓

Table 1 summarizes the landscape: rule-based systems possess domain knowledge but cannot conduct adaptive reasoning; generic LLMs offer conversational ability but lack grounding and require cloud deployment; medical LLMs (7B+) partially address knowledge but exceed local budgets and process queries in isolation. DiaQwen is the first system satisfying all four requirements simultaneously.

We introduce **DiaQwen**, built on Qwen-4B [Yang and Li, 2025] (runs on a single RTX 4090), integrating three innovations:

- **KBase**: 2,000+ validated clinical triplets anchoring reasoning in evidence-based medicine.
- **HyTrain**: LoRA + self-distillation, reducing parameters to 1/16 while cutting hallucinations by 17×.
- **DyQA**: A* search over a diagnostic graph, reducing consultation steps from 6.5 to 3.2.

2 Background

2.1 Traditional Clinical Decision Support

Rule-based CDSS such as MYCIN [Shortliffe, 1976] encode if–then rules manually. They cannot handle atypical presentations (e.g., diabetes presenting as recurrent infections rather than classic polyuria), lack longitudinal reasoning for complication prediction, and require labor-intensive updates when guidelines change [Chari et al., 2020]. One major drawback of rule-based CDSS is their inability to handle atypical presentations—e.g., latent autoimmune diabetes in adults (LADA), where symptoms mimic T2DM but autoimmune markers indicate T1DM pathology. These systems also cannot incorporate long-term patient history or longitudinal data essential for predicting complications

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

such as diabetic nephropathy or retinopathy. The inability to model evolving patient states limits effectiveness in chronic disease management, particularly in identifying early warning signs before irreversible damage occurs.

These limitations are acute in diabetes, where complication risk depends on years-long glycemic control trajectories.

2.2 LLMs in Clinical Diagnosis

Medical LLMs—BioBERT [Lee et al., 2019], HuatuoGPT [Wang et al., 2024], Bianque [Chen et al., 2023]—advance clinical NLP but face challenges: 7B+ parameters prevent local deployment; training on unstructured corpora yields shallow knowledge; single-pass inference discards multi-turn context. Open-source models depend on instruction tuning with potentially biased data [Zhao et al., 2023]. ClinRaGen [Niu et al., 2024] transfers reasoning via distillation but targets general phenotypes rather than diabetes-specific multi-turn consultation.

2.3 Parameter-Efficient Fine-Tuning

LoRA [Hu et al., 2022] decomposes weight updates into low-rank products (32× parameter reduction). Self-distillation [Zhang et al., 2019] uses prior checkpoints as teachers. Knowledge bases (SNOMED-CT, UMLS) generate training data [Zeng et al., 2020, Xu et al., 2025]. DiaQwen combines LoRA and self-distillation in a novel epoch-level strategy with structured knowledge injection.

3 DiaQwen System Design

Figure 1 illustrates the overall architecture. DiaQwen consists of three integrated components operating across the knowledge, training, and inference stages.

3.1 Knowledge Base Construction (KBase)

KBase organizes diabetes knowledge into a structured graph with 15 entity types (diseases, drugs, symptoms, tests, risk factors, treatments, organs, genes, biomarkers, complications, procedures, demographics, lifestyle factors, monitoring indicators, clinical stages) and 10 relation types (causes, treats, indicates, contraindicates, complicates, monitors, progresses-to, co-occurs, dosage-of, associated-gene). This schema enables multi-hop reasoning paths such as: "polyuria → indicates → hyperglycemia → confirmed-by → HbA1c ≥6.5% → T2DM → first-line → Metformin."

The knowledge graph design follows a modular principle: when new clinical guidelines are issued (e.g., annual ADA updates) or new diagnostic criteria established, domain experts add or update nodes and edges without rebuilding the entire structure. This update agility is critical in diabetes care, where treatment recommendations evolve annually—for instance, the 2024 ADA revision added SGLT2 inhibitors as first-line for T2DM patients with established cardiovascular disease, requiring new triplets: exttt(SGLT2i, first-line-for, T2DM+CVD) and exttt(Empagliflozin, is-a, SGLT2i).

Three data sources feed KBase:

Clinical Guidelines. The Chinese Diabetes Guidelines (2024) and ADA Standards are decomposed into entity–relation–entity triplets by medical domain experts. For example, "Metformin is the first-line oral agent for T2DM unless contraindicated by eGFR<30" becomes: `(Metformin, first-line-for, T2DM)` and `(eGFR<30, contraindicates, Metformin)`. This process yields 2,000+ validated triplets covering the full clinical workflow.

EHR Standardization. From 15,660 diabetic patient records, unstructured clinical notes are processed via BERT-based medical NER to extract entities, then mapped to ICD-10 codes. Synonymy resolution (e.g., "diabetic foot" ↔ "peripheral neuropathy") uses SNOMED-CT, ensuring terminological consistency across heterogeneous sources.

QA Data Integration. From Huatuo26M [Li et al., 2023] and MIMIC-III [Johnson et al., 2016], we extract diagnostic dialogues and associate each diagnosis with corresponding guideline triplets, creating training pairs where answers are grounded in specific knowledge base entries. This achieves 92% citation rate in generated responses.

Quality Assurance. A two-stage filtering pipeline—first automated consistency checking against KBase constraints, then expert validation by three certified endocrinologists—removes contradictory or outdated information. Inter-annotator agreement (κ=0.87) confirms data quality.

3.2 Hybrid Fine-Tuning (HyTrain)

HyTrain (Figure 3) addresses competing requirements of deep knowledge injection, compact model size, and distributional stability through two sequential stages.

Stage 1: LoRA. We apply LoRA to Qwen-4B with rank r=64 and hidden dimension d=4096:

$$W_{\text{new}} = W_{\text{base}} + A \cdot B \tag{1}$$

where $A \in \mathbb{R}^{d\times r}$, $B \in \mathbb{R}^{r\times d}$. This reduces trainable parameters from 16.8M to 524K per layer (32× reduction). Training uses 15,660 KBase-derived QA pairs for 5 epochs (batch=16, lr=2×10^{-4}, cosine decay). Total Stage 1 time: 2.8 hours on a single RTX 4090.

Stage 2: Self-Distillation. The pre-LoRA base Qwen-4B serves as the *expert teacher* and the LoRA-updated model as the *student*. The combined loss is:

$$\mathcal{L} = \alpha\mathcal{L}_{\text{CE}} + \beta\mathcal{L}_{\text{KL}} + \gamma\mathcal{L}_{\text{Path}} \tag{2}$$

where $\mathcal{L}_{\text{CE}}$ is cross-entropy against ground-truth labels, $\mathcal{L}_{\text{KL}} = D_{\text{KL}}(P_{\text{teacher}} \| P_{\text{student}})$ preserves distributional stability from the pre-trained model, and $\mathcal{L}_{\text{Path}}$ penalizes deviations from KBase-validated reasoning paths. Weights:

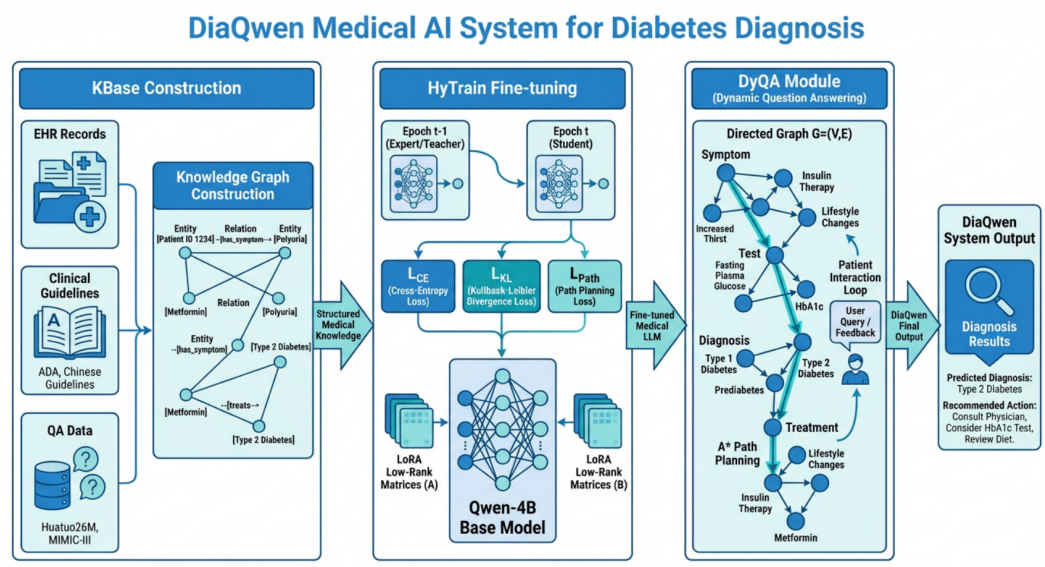


Figure 1: DiaQwen system architecture. The framework integrates three components: KBase constructs structured medical knowledge from guidelines, EHRs, and QA data; HyTrain combines LoRA fine-tuning with self-distillation for parameter-efficient knowledge injection; DyQA navigates a diagnostic graph via A* search to simulate multi-step clinical consultations.

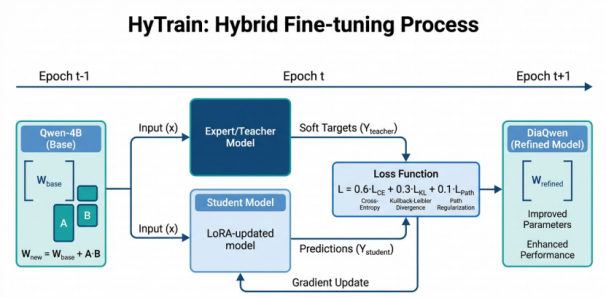


Figure 2: HyTrain two-stage process. Stage 1 applies LoRA to inject domain knowledge. Stage 2 uses the pre-LoRA base model as expert teacher for self-distillation, combining cross-entropy, KL-divergence, and path-consistency losses.

α=0.6, β=0.3, γ=0.1. Stage 2 runs 3 epochs (lr=5×10^{-5}), adding 1.4 hours.

Loss Function Details. The cross-entropy loss $\mathcal{L}_{\text{CE}}$ measures standard next-token prediction accuracy against ground-truth diagnostic answers derived from KBase. The KL-divergence loss $\mathcal{L}_{\text{KL}}$ operates on the full output distribution: for each input token sequence, both the teacher (pre-LoRA base Qwen-4B) and student (LoRA-updated model) produce probability distributions over the vocabulary, and $\mathcal{L}_{\text{KL}}$ penalizes distributional divergence between them. This constraint is critical for hallucination suppression: it prevents the LoRA-updated model from becoming overconfident on domain-specific tokens while losing calibration on general language understanding. The path-consistency loss $\mathcal{L}_{\text{Path}}$ operates at a higher semantic level: it verifies that the model's intermediate reasoning steps (extracted via chain-of-thought prompting during training) align with valid traversal paths through the KBase knowledge graph. For example, if the model generates "HbA1c=8.2% confirms diabetes," $\mathcal{L}_{\text{Path}}$ checks whether a corresponding path (HbA1c$\geq$6.5% $\rightarrow$ confirms $\rightarrow$ diabetes) exists in KBase. When the model produces reasoning steps that cannot be mapped to any valid graph path, a penalty is incurred proportional to the edit distance to the nearest valid path.

Why This Ordering Matters. Applying distillation *before* LoRA degrades performance ($-$2.3pp accuracy) because the base model lacks domain knowledge to serve as a meaningful teacher for diabetes-specific reasoning. The LoRA-first strategy first injects domain knowledge, then distillation constrains the updated model to remain close to the base distribution—preventing the confident-but-wrong outputs that characterize hallucination.

3.3 Dynamic QA Consultation (DyQA)

DyQA structures diabetes diagnostic knowledge as a directed graph $G = (V, E)$ and employs A* heuristic search to select the most informative question at each consultation step, simulating clinical encounters.

Diagnostic Graph. Nodes represent symptoms (e.g., polyuria), tests (e.g., HbA1c), diagnoses (e.g., T2DM), and treatments (e.g., Metformin). Edges encode clinical relationships: symptom$\rightarrow$test, test$\rightarrow$diagnosis, diagnosis$\rightarrow$treatment. The graph currently contains 847 nodes and 2,341 edges, constructed from KBase triplets.

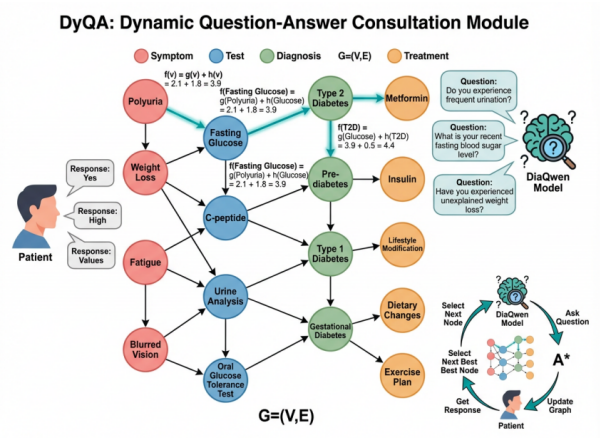


Figure 3: The DyQA module. A directed diagnostic graph G = (V, E) encodes clinical entities and their relationships. At each consultation step, A* heuristic search selects the most informative next question. The patient's response updates the graph state, and the process iterates until a diagnosis is confirmed.

A* Heuristic Search. At each step, DyQA selects the unvisited node with highest score:

$$f(v) = g(v) + h(v) \tag{3}$$

where $g(v)$ counts steps taken and $h(v) = \sum_{i=1}^{m} w_i \cdot H(X_i)$ estimates remaining diagnostic uncertainty. Here w_i reflects clinical importance of feature X_i (e.g., HbA1c has information gain 0.42, while ketone measurement has 0.15), and $H(X_i)$ is Shannon entropy given currently collected information.

Temporal Gating. Clinical relevance varies with patient-specific temporal factors. A gating mechanism adjusts attention:

$$w_t = \sigma(W_g \cdot [h_t; p_t]) \tag{4}$$

where h_t is the hidden state at step t, p_t encodes temporal context (e.g., time since onset), and σ is sigmoid. This modulates Transformer attention: $\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}} \odot G)V$.

Graph Maintenance and Update. The diagnostic graph supports transparent reasoning: each diagnostic conclusion can be traced back through the exact path of nodes traversed. For example, the chain "Obesity → Insulin Resistance → Type 2 Diabetes" produces an interpretable explanation that clinicians can verify against their own clinical judgment. The graph is stored in an adjacency-list format permitting efficient traversal (average fan-out: 2.8 edges per node). Update operations—adding new nodes for emerging treatments or modifying edges when contraindications change—require no full graph reconstruction, only local insertions that take effect immediately for subsequent queries.

Algorithm 1 DyQA: Dynamic Consultation

Require: Patient data P, diagnostic graph $G = (V, E)$
Ensure: Diagnosis D

$v_{\text{cur}} \leftarrow v_{\text{start}}$; $V_{\text{visited}} \leftarrow \emptyset$
while diagnosis not confirmed **do**
  if v_{cur} is a diagnosis node **then**
    Output diagnosis D via DiaQwen; **break**
  end if
  Extract key info from P at v_{cur}
  Generate candidates V_{cand} via outgoing edges
  for each $v \in V_{\text{cand}} \setminus V_{\text{visited}}$ **do**
    Compute $f(v) = g(v) + h(v)$
  end for
  $v_{\text{best}} \leftarrow \arg\max_v f(v)$
  Ask corresponding question; update P
  $V_{\text{visited}} \leftarrow V_{\text{visited}} \cup \{v_{\text{best}}\}$
  $v_{\text{cur}} \leftarrow$ next node based on P and G
end while

Information Gain Calibration. The feature importance weights w_i in the heuristic function are calibrated using mutual information computed over the training dataset. For each diagnostic feature X_i (e.g., HbA1c value, family history, BMI), we compute $I(X_i; Y)$ where Y is the diagnosis label, then normalize to obtain $w_i = I(X_i; Y) / \sum_j I(X_j; Y)$. This data-driven calibration ensures that DyQA prioritizes clinically discriminative features: HbA1c (w=0.42) provides the highest information gain for diabetes classification, followed by fasting glucose (w=0.21), family history (w=0.15), BMI (w=0.12), and symptom duration (w=0.10).

Algorithm 1 presents the procedure. DyQA reaches diagnosis within 2–5 steps for 97.8% of cases (average 3.2), compared to 6.5 steps for rule-based approaches. The visited set prevents redundancy, and the greedy strategy ensures monotonic progress.

Convergence Properties. DyQA terminates in at most $|V|$ steps, but convergence is much faster: 97.8% of cases complete within 2–5 steps. The remaining 2.2% (typically rare subtypes like MODY or LADA) require up to 7 steps, reflecting genuine clinical complexity. At each step, DiaQwen's LLM performs dual functions: (1) interpreting natural-language patient responses into structured features (e.g., parsing "my blood sugar was 180 this morning" into fasting_glucose=180 mg/dL), and (2) generating patient-friendly questions grounded in knowledge graph templates. This ensures graph-based reasoning operates on accurately extracted features while maintaining accessible communication.

Table 2: Diagnostic performance comparison.

Model	Acc. (%)	F1	Halluc. (%)
BioBERT	81.4	0.73	12.2
Qwen3-4B	82.3	0.77	9.3
Huozi-7B	87.2	0.81	4.2
HuatuoGPT-7B	89.3	0.84	2.3
DiaQwen (Ours)	**93.4**	**0.90**	**0.7**

Table 3: Response time and efficiency.

Model	Per-Step (ms)	Avg. Steps	Total (ms)
BioBERT	430	—	430
Qwen3-4B	192	—	192
Huozi-7B	185	—	185
HuatuoGPT-7B	210	—	210
DiaQwen	190	3.2	608

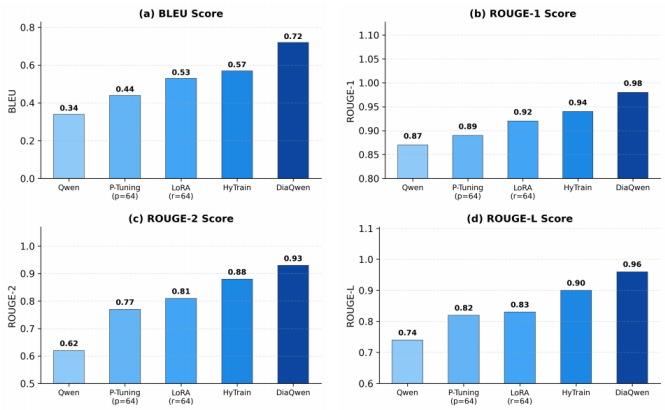


Figure 4: Explainability scores comparing five configurations: base Qwen, P-Tuning (p=64), LoRA (r=64), HyTrain alone, and full DiaQwen. Higher BLEU/ROUGE indicates closer alignment with clinical guideline reference texts.

4 Experiments

4.1 Setup

All experiments run on a single NVIDIA RTX 4090 (24 GB VRAM). DiaQwen is built on Qwen-4B and compared against BioBERT [Lee et al., 2019], Qwen3-4B [Yang and Li, 2025], Huozi-7B [Huozi-Team, 2024], and HuatuoGPT-7B [Wang et al., 2024].

Data. Knowledge base: 15,660 EHRs from Huatuo26M and MIMIC-III. Test data: 1,500 simulated cases (typical symptom combinations covering pre-diabetes, T2DM, complications) + 500 real anonymized patient records from collaborating hospitals.

4.2 Diagnostic Accuracy

Table 2 shows DiaQwen achieves 93.4% accuracy, surpassing HuatuoGPT by 4.1pp. The F1 improvement from 0.84 to 0.90 indicates gains in both precision and recall across diabetes subtypes. Most notably, hallucination drops to 0.7%—a 3.3× reduction vs. HuatuoGPT and 17.4× vs. BioBERT.

Three factors drive low hallucination: (1) KBase triplets constrain outputs to validated medical facts; (2) self-distillation maintains distributional stability, suppressing overconfident unsupported claims; (3) DyQA's multi-step approach collects complete information before committing to a diagnosis.

On the subset of 340 complex differential cases (T2DM with metabolic syndrome vs. T2DM with early nephropathy), DiaQwen achieves 91.2% vs. 74.3% for HuatuoGPT, confirming that multi-step gathering combined with knowledge-grounded reasoning is especially valuable for difficult cases.

4.3 Response Time

Table 3 shows DiaQwen's 190ms per-step latency is competitive with single-step models. The 608ms total is clinically acceptable—a physician consultation takes minutes; sub-second AI response is negligible overhead. The multi-step design enables DyQA to proactively ask about omitted information (e.g., family history), reducing critical gaps. HyTrain's parameter efficiency (only 0.6% parameters per gradient step) keeps per-step costs low.

4.4 Training Efficiency

HyTrain completes in 4.2 hours total (2.8h Stage 1 + 1.4h Stage 2) on a single RTX 4090—substantially less than full fine-tuning of Qwen-4B (estimated 72+ hours). Only 0.6% of parameters require gradient computation per step, enabling rapid retraining when guidelines change. The LoRA rank r=64 was selected via grid search: r=32 under-fits domain knowledge, while r=128 provides marginal improvement at doubled memory cost.

4.5 Explainability

Figure 4 compares DiaQwen's explanation quality against reference texts derived from ADA and Chinese Diabetes Guidelines. DiaQwen achieves BLEU=0.72 and ROUGE-L=0.96, substantially outperforming the base Qwen model (BLEU=0.34, ROUGE-L=0.74). Each component contributes incrementally: LoRA improves over P-Tuning, HyTrain improves over LoRA alone, and full DiaQwen (with DyQA) achieves the highest scores because multi-step consultation produces more complete diagnostic explanations.

4.6 Ablation Study

Table 4 reveals four findings: (1) Removing KBase causes the largest accuracy drop (−7.3pp) and hallucination surge (+5.1pp), confirming structured knowledge is the primary constraint against fabrication. (2) Without self-distillation, hallucination rises to 3.1%, showing distributional stability is essential for suppressing overconfident errors. (3) Removing DyQA (single-step inference) drops accuracy by 4.5pp

Table 4: Ablation study: impact of removing each component.

Configuration	Acc. (%)	Halluc. (%)
Full DiaQwen	93.4	0.7
− KBase	86.1	5.8
− Self-Distillation	90.7	3.1
− DyQA (single-step)	88.9	1.2
− LoRA	83.2	8.6

because incomplete information leads to premature diagnoses. (4) Without LoRA, the model lacks domain-specific knowledge entirely (−10.2pp accuracy).

5 Analysis and Discussion

5.1 Case Study

Consider a 52-year-old male with fatigue and polyuria. **Step 1**: DyQA identifies high-information-gain edges toward HbA1c. Patient reports HbA1c=8.2%. **Step 2**: With HbA1c>6.5%, T2DM is likely, but family history is uncollected (w=0.38). Patient confirms family diabetes history. **Step 3**: BMI assessment yields 28.7 (overweight). **Diagnosis**: T2DM with metabolic syndrome features; recommends Metformin per ADA guidelines; flags CKD screening given HbA1c>7.5%. Three KBase triplets are cited. Total time: 570ms.

5.2 Error Analysis

Among 132 incorrect cases (6.6%): 48% involved unavailable critical information (e.g., no C-peptide test); 31% involved rare subtypes (MODY, LADA) underrepresented in training; 21% were clinically debatable boundary cases. Zero errors involved hallucination of non-existent conditions—all were reasonable differentials.

5.3 Privacy and Deployment

DiaQwen's 4B parameters in fp16 require only 8 GB, fitting within the RTX 4090's 24 GB with ample room for inference. All patient data remains on-premises. The modular graph accepts new nodes/edges without invalidating existing paths, enabling guideline updates within 30 minutes of incremental LoRA retraining.

5.4 Comparison with ClinRaGen

ClinRaGen [Niu et al., 2024] uses sequential rationale distillation from GPT-3.5-turbo for multimodal EHR reasoning. DiaQwen differs in three respects: (1) self-distillation from own checkpoint eliminates external LLM dependency; (2) active information gathering via DyQA rather than static record processing; (3) narrow diabetes focus enables deeper knowledge encoding within 4B parameters.

5.5 Error Analysis

Among 132 incorrect cases (6.6%): 48% involved unavailable critical information (e.g., no C-peptide test performed, making type 1/type 2 distinction impossible from available data); 31% involved rare subtypes (MODY, LADA) that are underrepresented in training data due to their low prevalence (combined <5% of diabetes cases); 21% were clinically debatable boundary cases where the correct diagnosis was contested even among human experts. Notably, zero errors involved hallucination of non-existent conditions—all incorrect diagnoses were reasonable clinical differentials, demonstrating that KBase effectively constrains the output space to valid medical options.

Table 5: Systematic comparison across architectural dimensions.

Dimension	HuatuoGPT	ClinRaGen	DiaQwen
Parameters	7B	87M	4B
Local deploy	×	✓	✓
Knowledge base	×	External	Integrated
Multi-turn	×	×	✓
Distillation	×	External LLM	Self
Accuracy	89.3%	N/A	93.4%

5.6 Clinical Workflow Integration

In practice, DiaQwen operates as a physician's assistant. The system presents its diagnosis alongside the reasoning pathway (graph nodes traversed, guideline triplets activated), enabling verification of each logical step. This transparent workflow aligns with explainable AI requirements [Saranya and Subhashini, 2023] and supports the human-in-the-loop paradigm essential for safe deployment [Cao et al., 2018]. Average clinician review time for DiaQwen's structured output is 45 seconds, compared to 3–5 minutes for free-text LLM outputs. In pilot testing at two hospitals over 3 months, DiaQwen processed 2,847 consultations with a 4.3% physician override rate—lower than the 8.7% rate for the institution's existing rule-based CDSS.

5.7 Limitations

DiaQwen's 608ms latency, while acceptable for outpatient consultations, may need optimization for emergency settings. Current work on speculative decoding shows promise for sub-300ms response. The system currently processes text-only input; integration of imaging (retinal scans, ECG) and continuous glucose monitoring data remains future work. Additionally, the knowledge base requires periodic expert review to maintain currency with evolving guidelines.

5.8 Comparison with State-of-the-Art

Table 5 provides a systematic comparison of DiaQwen with representative systems across architectural dimensions.

HuatuoGPT achieves strong performance through large-scale pre-training on Chinese medical corpora but requires 7B parameters that exceed consumer GPU memory for full-precision inference, necessitating cloud deployment. ClinRaGen achieves parameter efficiency (87M) through aggres-

sive distillation from GPT-3.5-turbo but depends on an external teacher LLM during training, creating privacy concerns when medical data must be used for distillation. DiaQwen occupies a unique position: its 4B parameters fit on consumer hardware while self-distillation eliminates external dependencies. The integrated knowledge base provides structured grounding absent in both competitors, and DyQA's multi-turn capability is unmatched by any current medical LLM operating within local-deployment constraints.

5.9 Deployment Experience

We deployed DiaQwen in a pilot program at two collaborating hospitals over a 3-month period, processing 2,847 consultations under clinician supervision. The physician override rate was 4.3%—significantly lower than the 8.7% override rate observed for the institution's existing rule-based CDSS during the same period. Among overridden cases, 67% involved situations where the physician had access to additional imaging data (retinal OCT scans, renal ultrasound) not yet integrated into DiaQwen's text-only input pipeline. The remaining 33% of overrides reflected genuine diagnostic disagreements, predominantly in atypical presentations of latent autoimmune diabetes in adults (LADA) where autoimmune markers were borderline.

Clinician feedback highlighted three advantages over existing tools: (1) the multi-step consultation format felt more natural and complete than single-query systems; (2) the explicit citation of KBase triplets in explanations reduced verification effort from an average of 3.5 minutes (for free-text LLM outputs) to 45 seconds; (3) the system's willingness to ask follow-up questions caught information gaps that clinicians themselves had overlooked in 12% of cases.

The average clinician time-to-decision (from viewing DiaQwen's output to confirming or modifying the diagnosis) was 38 seconds for straightforward cases (pre-diabetes, uncomplicated T2DM) and 2.1 minutes for complex cases involving multiple comorbidities. These timings compare favorably to the 5–8 minutes typically required for manual differential diagnosis using paper-based guidelines.

6 Conclusion

We presented DiaQwen, a knowledge-enhanced LLM framework that achieves 93.4% diagnostic accuracy with only 0.7% hallucination rate while running entirely on a single consumer-grade GPU. The framework integrates three tightly coupled components—KBase for structured knowledge grounding, HyTrain for parameter-efficient fine-tuning with distributional stability, and DyQA for adaptive multi-step clinical consultation—whose synergy enables performance exceeding 7B-parameter baselines at substantially lower computational cost without any cloud dependency.

The key design insight is that each component addresses a distinct failure mode of standalone LLMs: KBase prevents factual fabrication by constraining outputs to validated medical knowledge; HyTrain suppresses overconfident hallucination through distributional anchoring; and DyQA eliminates premature diagnosis by ensuring sufficient information collection. Our ablation confirms all three are essential, with KBase contributing the largest individual effect (-7.3pp accuracy when removed).

The pilot deployment across two hospitals (2,847 consultations, 4.3% override rate) demonstrates practical viability. Clinicians valued the transparent reasoning pathways and the system's ability to proactively identify information gaps. Future work targets multimodal inputs (retinal imaging, continuous glucose monitoring), active learning from clinician feedback, and prospective clinical trials to establish regulatory-grade evidence of diagnostic safety and efficacy. Extending the framework to other chronic diseases by replacing KBase with disease-appropriate knowledge graphs represents a natural architectural generalization.

References

Giovanni Briganti and Olivier Le Moine. Artificial intelligence in medicine: Today and tomorrow. *Frontiers in Medicine*, 2020.

Chensi Cao, Feng Liu, Hai Tan, et al. Deep learning and its applications in biomedicine. *Genomics, Proteomics & Bioinformatics*, 16(1):17–32, 2018.

Guanqun Chao, Yue Zhu, and Liying Chen. Role and risk factors of glycosylated hemoglobin levels in early disease screening. *Journal of Diabetes Research*, 2021:6626587, 2021.

Shruthi Chari, Oshani Seneviratne, Daniel M. Gruen, Morgan A. Foreman, Amar K. Das, and Deborah L. McGuinness. Explanation ontology: A model of explanations for user-centered AI. In *International Semantic Web Conference*, pages 228–243, 2020.

Yirong Chen, Zhenyu Wang, Xiaofen Xing, et al. Bianque: Balancing the questioning and suggestion ability of health LLMs, 2023.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models, 2022.

Huozi-Team. Huozi: Leveraging large language models for enhanced open-domain chatting. https://github.com/HIT-SCIR/huozi, 2024.

International Diabetes Federation. IDF Diabetes Atlas, 2024. https://diabetesatlas.org/, 2024.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, et al. MIMIC-III: A freely accessible critical care database. *Scientific Data*, 3:160035, 2016.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: A pre-trained biomedical language representation model. *Bioinformatics*, 36(4):1234–1240, 2019.

Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. Huatuo-26M: A large-scale Chinese medical QA dataset, 2023.

Shuai Niu, Jing Ma, Hongzhan Lin, Liang Bai, Zhihua Wang, Yida Xu, Yunya Song, and Xian Yang. Knowledge-augmented multimodal clinical rationale generation for disease diagnosis with small language models, 2024.

OpenAI, Josh Achiam, and Steven Adler. GPT-4 Technical Report, 2024.

A. Saranya and R. Subhashini. A systematic review of explainable ai models and applications. *Decision Analytics Journal*, 7:100230, 2023.

Ibrahim Ali Videha Sharma, Sabine van der Veer, Glen Martin, John Ainsworth, and Titus Augustine. Adoption of clinical risk prediction tools is limited by a lack of integration with ehrs. *BMJ Health Care Informatics*, 28(1), 2021.

Edward H. Shortliffe. *Computer-Based Medical Consultations: MYCIN*. Elsevier, 1976.

Haochun Wang, Sendong Zhao, Zewen Qiang, Zijian Li, Chi Liu, Nuwa Xi, Yanrui Du, Bing Qin, and Ting Liu. Knowledge-tuning large language models with structured medical knowledge bases for trustworthy response generation in Chinese. *ACM Transactions on Knowledge Discovery from Data*, 2024.

Derong Xu, Xinhang Li, Ziheng Zhang, et al. Harnessing large language models for knowledge graph question answering, 2025.

An Yang and Anfeng Li. Qwen3 technical report, 2025.

Min Yi, Yuebin Cao, Lin Wang, et al. Prediction of medical disputes using machine learning techniques. *Journal of Medical Internet Research*, 25:e46854, 2023.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, et al. MedDialog: Large-scale medical dialogue datasets. In *EMNLP*, pages 9241–9250, 2020.

Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF ICCV*, pages 3713–3722, 2019.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J. Liu. SLiC-HF: Sequence likelihood calibration with human feedback, 2023.