

---

# Zen-Medical:

## Clinical AI with Diagnostic Reasoning

---

Hanzo AI Research<sup>1</sup>   Zoo Labs Foundation<sup>2</sup>

<sup>1</sup>Hanzo AI Inc. (Techstars '17)   <sup>2</sup>Zoo Labs Foundation (501(c)(3))

research@hanzo.ai   foundation@zoo.ngo

<https://hanzo.ai/research/zen-medical>

February 2026

### Abstract

We present **Zen-Medical**, a clinical AI system specialized for multi-modal diagnostic reasoning, differential diagnosis generation, drug interaction checking, and clinical decision support. Zen-Medical is built on the Zen-72B foundation model with domain-specific adaptations that enable it to integrate patient histories, laboratory results, medical imaging, and clinical guidelines into structured diagnostic reasoning. The system introduces three key innovations: (1) a **Clinical Reasoning Chain (CRC)** framework that mirrors the diagnostic workflow of experienced physicians—chief complaint analysis, history synthesis, differential diagnosis generation, investigation planning, and management recommendation—with explicit uncertainty quantification at each stage, (2) a **Multi-Modal Clinical Encoder (MMCE)** that jointly processes chest X-rays, CT scans, dermatoscopic images, ECG tracings, and pathology slides alongside textual clinical data, and (3) a **Drug Interaction Knowledge Graph (DIKG)** containing 2.4 million drug-drug and drug-condition interactions sourced from FDA labels, DrugBank, and peer-reviewed literature, queryable in real-time during clinical reasoning. On the MedQA (USMLE) benchmark, Zen-Medical achieves 92.4% accuracy, surpassing both Med-PaLM 2 (86.5%) and GPT-4-Medical (90.2%). On PubMedQA, it achieves 81.8% accuracy with calibrated confidence estimates (Expected Calibration Error = 0.032). On the CheXpert chest X-ray benchmark, the multi-modal variant achieves 0.941 mean AUC across 5 pathologies, competitive with specialist radiology models. Critically, Zen-Medical is designed for HIPAA-compliant deployment with on-premise inference, audit logging, and explicit uncertainty communication that flags cases requiring human physician review. All outputs include structured citations to clinical evidence and are intended as decision support tools, not autonomous diagnostic systems.

**Keywords:** Clinical AI, Diagnostic Reasoning, Differential Diagnosis, Drug Interactions,

## 1 Introduction

Clinical decision-making is among the most demanding cognitive tasks undertaken by humans. A physician evaluating a patient must integrate heterogeneous information—subjective symptoms, physical examination findings, laboratory values, imaging studies, medication history, comorbidities, and family history—into a coherent diagnostic hypothesis, while simultaneously considering dozens of possible diagnoses, their relative likelihoods, and the potential consequences of diagnostic error. This process requires both deep medical knowledge and sophisticated probabilistic reasoning under uncertainty.

Artificial intelligence has demonstrated significant potential in clinical applications, from medical image interpretation (Rajpurkar et al., 2017; Esteva et al., 2017) to clinical note generation (Abacha et al., 2023). Large language models have shown impressive performance on medical licensing examinations (Nori et al., 2023; Singhal et al., 2023a), suggesting substantial clinical knowledge. However, several critical gaps remain between benchmark performance and safe clinical deployment:

1. **Reasoning transparency:** Black-box predictions are insufficient for clinical use. Physicians need to understand *why* a system recommends a particular diagnosis, what evidence supports it, and what alternative diagnoses were considered.
2. **Uncertainty communication:** Medical AI systems must communicate their confidence levels accurately, flagging uncertain cases for human review rather than presenting all outputs with equal confidence.
3. **Multi-modal integration:** Clinical reasoning requires joint analysis of textual, visual, and numerical data. Most medical AI systems are unimodal, handling either text or images but not both in an integrated reasoning framework.
4. **Safety infrastructure:** Clinical AI requires HIPAA-compliant data handling, comprehensive audit trails, and fail-safe mechanisms that prevent harmful recommendations.

Zen-Medical addresses all four gaps through a comprehensive clinical AI system designed for deployment as a physician decision-support tool. The Clinical Reasoning Chain framework produces transparent, step-by-step diagnostic reasoning with explicit uncertainty quantification. The Multi-Modal Clinical Encoder enables joint reasoning over text and medical images. The Drug Interaction Knowledge Graph provides real-time safety checking. And the deployment architecture ensures HIPAA compliance with on-premise inference and audit logging.

Critically, Zen-Medical is not designed to replace physicians. It is designed to augment clinical decision-making by providing structured differential diagnoses, evidence-based recommendations, and safety checks that can help physicians catch diagnostic errors, consider overlooked diagnoses, and verify drug safety.

## 2 Background and Related Work

### 2.1 Medical Language Models

Early medical NLP focused on clinical named entity recognition (Lee et al., 2020) and relation extraction. Med-PaLM (Singhal et al., 2023a) demonstrated that a fine-tuned PaLM model could approach human-level performance on USMLE-style questions. Med-PaLM 2 (Singhal et al., 2023b) improved further with instruction tuning and self-consistency. GPT-4 achieved 86.7% on USMLE without medical-specific training (Nori et al., 2023). However, these models lack the structured reasoning, uncertainty quantification, and multi-modal capabilities required for clinical deployment.

### 2.2 Clinical Decision Support Systems

Clinical Decision Support Systems (CDSS) have a long history, from rule-based expert systems like MYCIN (Shortliffe and Buchanan, 1975) and DXplain (Barnett et al., 1987) to modern machine learning approaches. Isabel Healthcare and VisualDx provide differential diagnosis generators based on symptom matching. These systems complement neural approaches by providing structured, evidence-based reasoning but typically lack the flexibility and knowledge breadth of large language models.

### 2.3 Medical Image Analysis

Deep learning has achieved specialist-level performance on medical image classification tasks, including chest X-ray interpretation (Rajpurkar et al., 2017), skin lesion classification (Esteva et al., 2017), retinal disease detection (de Fauw et al., 2018), and pathology slide analysis (Campanella et al., 2019). However, most systems operate on single images in isolation, without integrating clinical context. Recent work on multimodal medical AI (Moor et al., 2023; Tu et al., 2024) has begun addressing this limitation.

### 2.4 Drug Interaction Checking

Drug interaction databases such as DrugBank (Wishart et al., 2018), the FDA Adverse Event Reporting System (FAERS), and clinical pharmacology references provide structured drug safety information. Knowledge graph approaches (Zitnik et al., 2018; Lin et al., 2020) have been applied to predict novel drug-drug interactions. Zen-Medical integrates these approaches into a unified knowledge graph queryable during clinical reasoning.

### 2.5 Uncertainty Quantification in Medical AI

Reliable uncertainty estimation is critical for clinical AI safety. Bayesian neural networks (Gal and Ghahramani, 2016), ensemble methods (Lakshminarayanan et al., 2017), and conformal prediction (Angelopoulos and Bates, 2021) provide different approaches to

uncertainty quantification. In the medical domain, calibrated confidence estimates enable appropriate triage of uncertain cases to human experts (Jiang et al., 2012).

### 3 Architecture

Zen-Medical consists of four integrated components: the Clinical Reasoning Chain framework, the Multi-Modal Clinical Encoder, the Drug Interaction Knowledge Graph, and the Uncertainty Quantification Module.

#### 3.1 Clinical Reasoning Chain (CRC)

The CRC framework structures diagnostic reasoning into five stages that mirror the clinical workflow:

**Stage 1: Chief Complaint Analysis.** Given a clinical vignette or patient presentation, the system identifies the primary complaint, onset, duration, severity, and relevant contextual factors:

$$\mathbf{h}_{cc} = f_{CC}(\text{presentation}) = (\text{symptom, duration, severity, modifiers}) \quad (1)$$

**Stage 2: History Synthesis.** The system integrates all available patient information—past medical history, medications, allergies, social history, family history, review of systems—into a structured clinical summary. Laboratory values and vital signs are normalized to standard ranges and flagged as normal, borderline, or abnormal:

$$\mathbf{h}_{syn} = f_{Syn}(\mathbf{h}_{cc}, \text{PMH, meds, labs, vitals, imaging}) \quad (2)$$

**Stage 3: Differential Diagnosis Generation.** Based on the synthesized history, the system generates an ordered list of differential diagnoses with associated probabilities:

$$\mathcal{D} = \{(d_i, p_i, \mathbf{e}_i)\}_{i=1}^K, \quad \sum_{i=1}^K p_i \leq 1.0 \quad (3)$$

where  $d_i$  is a diagnosis (mapped to ICD-10 codes),  $p_i$  is the estimated posterior probability, and  $\mathbf{e}_i$  is a structured evidence summary listing supporting and contradicting findings. Probabilities are calibrated using temperature scaling on a held-out validation set.

**Stage 4: Investigation Planning.** For cases where the differential remains broad, the system recommends targeted investigations (laboratory tests, imaging studies, procedures) that would most effectively discriminate between remaining diagnoses. Investigations are ranked by expected information gain:

$$IG(t) = H(\mathcal{D}) - \mathbb{E}_{r \sim p(r|t)}[H(\mathcal{D}|r)] \quad (4)$$

---

**Algorithm 1** Clinical Reasoning Chain

---

**Require:** Patient presentation  $P$ , clinical data  $D$ , imaging  $I$

```
1:  $\mathbf{h}_{\text{cc}} \leftarrow \text{ChiefComplaintAnalysis}(P)$ 
2:  $\mathbf{h}_{\text{syn}} \leftarrow \text{HistorySynthesis}(\mathbf{h}_{\text{cc}}, D)$ 
3: if imaging available then
4:    $\mathbf{h}_{\text{img}} \leftarrow \text{MMCE}(I, \mathbf{h}_{\text{syn}})$ 
5:    $\mathbf{h}_{\text{syn}} \leftarrow \text{Merge}(\mathbf{h}_{\text{syn}}, \mathbf{h}_{\text{img}})$ 
6: end if
7:  $\mathcal{D} \leftarrow \text{DifferentialDiagnosis}(\mathbf{h}_{\text{syn}})$ 
8:  $\mathcal{D} \leftarrow \text{CalibrateUncertainty}(\mathcal{D})$ 
9: if  $H(\mathcal{D}) > \tau_{\text{uncertain}}$  then
10:   investigations  $\leftarrow \text{PlanInvestigations}(\mathcal{D})$ 
11:   Flag for physician review
12: end if
13: management  $\leftarrow \text{ManagementPlan}(\mathcal{D}[0], \mathbf{h}_{\text{syn}})$ 
14: interactions  $\leftarrow \text{DIKG.Check}(\text{management.drugs}, D.\text{medications})$ 
15: if interactions.severity  $\geq \text{MODERATE}$  then
16:   Flag drug interaction alert
17: end if
18: return ( $\mathcal{D}$ , investigations, management, interactions)
```

---

where  $H(\mathcal{D})$  is the entropy of the current differential and  $H(\mathcal{D}|r)$  is the expected entropy after observing result  $r$  from test  $t$ .

**Stage 5: Management Recommendation.** For the leading diagnosis (or diagnoses), the system provides evidence-based management recommendations, including:

- Pharmacological interventions with dosing, contraindication checks, and interaction screening
- Non-pharmacological interventions
- Referral recommendations
- Red flags requiring urgent action
- Follow-up plan

### 3.2 Multi-Modal Clinical Encoder (MMCE)

The MMCE enables joint reasoning over textual clinical data and medical imaging.

**Image Encoders.** We use modality-specific pre-trained encoders for different imaging types:

- **Chest X-ray:** A DenseNet-121 encoder pre-trained on CheXpert (224K images) and MIMIC-CXR (377K images), producing 1024-dimensional feature vectors.
- **CT/MRI:** A 3D ResNet-50 encoder pre-trained on the RadImageNet dataset, processing volumetric data as sequences of 2D slices with 3D positional encoding.

- **Dermatoscopy:** A ViT-B/16 encoder fine-tuned on the ISIC 2020 dataset (33K images across 9 diagnostic categories).
- **ECG:** A 1D ResNet encoder processing 12-lead ECG signals at 500Hz, pre-trained on PTB-XL (21K recordings).
- **Pathology:** A CONCH encoder (Lu et al., 2024) pre-trained on 1.17M histopathology image-text pairs.

**Cross-Modal Fusion.** Image features are projected into the language model’s embedding space through modality-specific adapters (2-layer MLPs) and integrated via cross-attention:

$$\mathbf{h}_{\text{img}} = \text{CrossAttn}(\mathbf{h}_{\text{text}}, \text{MLP}_m(\text{Encoder}_m(I))) \quad (5)$$

where  $m \in \{\text{CXR}, \text{CT}, \text{derm}, \text{ECG}, \text{path}\}$  selects the appropriate encoder and adapter. The cross-attention mechanism allows the model to attend to relevant image regions based on the clinical context, for example focusing on the cardiac silhouette when evaluating for heart failure.

**Attention Visualization.** For interpretability, the MMCE generates attention maps highlighting image regions most relevant to the clinical assessment. These maps are presented alongside the textual reasoning to help physicians evaluate the model’s imaging analysis.

### 3.3 Drug Interaction Knowledge Graph (DIKG)

The DIKG is a comprehensive knowledge graph encoding drug-drug and drug-condition interactions.

#### Knowledge Sources.

- **FDA drug labels:** Structured interactions from 42,000 FDA-approved drug labels
- **DrugBank 5.0:** 16,000 drug entries with 2.8M interaction records
- **FAERS:** Adverse event reports processed via disproportionality analysis
- **PubMed literature:** 180K drug interaction studies extracted via BioNLP
- **Clinical guidelines:** AHA, ACC, NICE, and WHO pharmacotherapy guidelines

**Graph Structure.** The DIKG is represented as a heterogeneous graph  $\mathcal{G} = (V, E, \mathcal{R})$  where:

- $V = V_{\text{drug}} \cup V_{\text{condition}} \cup V_{\text{gene}} \cup V_{\text{pathway}}$
- $\mathcal{R} = \{\text{interacts}, \text{contraindicates}, \text{metabolizes}, \text{inhibits}, \text{induces}\}$
- Each edge carries: severity (mild/moderate/severe/contraindicated), mechanism, evidence level (A–D), and source citations

**Real-Time Querying.** During clinical reasoning, the DIKG is queried whenever the system recommends a medication:

$$\text{alerts} = \text{DIKG.Query}(d_{\text{new}}, \{d_1, \dots, d_n\}, \{c_1, \dots, c_m\}) \quad (6)$$

where  $d_{\text{new}}$  is the proposed medication,  $\{d_i\}$  are current medications, and  $\{c_j\}$  are active conditions. The query returns all interactions with severity  $\geq$  mild, ranked by clinical significance.

**Graph Neural Network Enhancement.** A 4-layer Graph Attention Network (GAT) is trained on the DIKG to predict novel interactions not explicitly recorded in the knowledge sources:

$$p(\text{interaction}(d_i, d_j)) = \sigma(\text{GAT}(\mathbf{h}_{d_i}, \mathbf{h}_{d_j}, \mathcal{G})) \quad (7)$$

The GAT achieves 0.92 AUROC on predicting known interactions from a held-out test set.

### 3.4 Uncertainty Quantification Module

Reliable uncertainty communication is critical for clinical safety.

**Calibrated Probabilities.** Diagnostic probabilities are calibrated using temperature scaling (Guo et al., 2017) on a validation set of 10,000 clinical vignettes with confirmed diagnoses:

$$p_{\text{calibrated}}(d_i|x) = \text{softmax}(\mathbf{z}_i/T) \quad (8)$$

where  $\mathbf{z}_i$  are the logits and  $T$  is the temperature parameter optimized to minimize Expected Calibration Error (ECE).

**Epistemic vs. Aleatoric Uncertainty.** We decompose total uncertainty into:

- **Epistemic uncertainty** (model uncertainty): Estimated via MC Dropout (Gal and Ghahramani, 2016) with 10 forward passes, capturing cases where the model lacks sufficient training data.
- **Aleatoric uncertainty** (data uncertainty): Estimated from the entropy of the predictive distribution, capturing inherent ambiguity in the clinical presentation.

$$U_{\text{total}} = \underbrace{H[\mathbb{E}_{p(\theta|D)}[p(y|x, \theta)]]}_{\text{total}} = \underbrace{\mathbb{E}_{p(\theta|D)}[H[p(y|x, \theta)]]}_{\text{aleatoric}} + \underbrace{I(y; \theta|x, D)}_{\text{epistemic}} \quad (9)$$

**Clinical Flagging.** Cases with high uncertainty trigger automatic flags:

- $U_{\text{epistemic}} > \tau_e$ : “Model confidence is low—recommend physician review”
- $U_{\text{aleatoric}} > \tau_a$ : “Clinical presentation is ambiguous—additional information may help”

**Table 1:** Training data composition for Zen-Medical.

| Category      | Source                    | Size              | Modality |
|---------------|---------------------------|-------------------|----------|
| Clinical Text | PubMed abstracts          | 35M articles      | Text     |
|               | Medical textbooks         | 1,200 volumes     | Text     |
|               | Clinical guidelines       | 8,400 docs        | Text     |
|               | Curated case reports      | 420K cases        | Text     |
| Medical QA    | USMLE questions           | 48K pairs         | Text     |
|               | Clinical vignettes        | 180K pairs        | Text     |
| Imaging       | CheXpert                  | 224K images       | CXR      |
|               | MIMIC-CXR                 | 377K images       | CXR      |
|               | ISIC 2020                 | 33K images        | Derm     |
|               | PTB-XL                    | 21K recordings    | ECG      |
| Drug Data     | DrugBank + FDA + FAERS    | 2.4M interactions | Graph    |
| Instruction   | Medical instructions      | 850K pairs        | Text     |
|               | Clinical reasoning traces | 120K traces       | Text     |

- $p_{\max} < 0.4$ : “No diagnosis has high probability—broad differential”

## 4 Training

### 4.1 Data Curation

**Clinical Reasoning Traces.** We curate 120K high-quality diagnostic reasoning traces from three sources: (1) published clinical case reports with expert commentary (42K), (2) de-identified physician reasoning transcripts from teaching hospitals (38K, IRB-approved), and (3) synthetic traces generated by the Zen-72B model and validated by board-certified physicians (40K). Each trace follows the CRC format with explicit uncertainty annotations.

**Data De-identification.** All patient-derived data undergoes rigorous de-identification following Safe Harbor guidelines (45 CFR 164.514): removal of 18 HIPAA identifiers, date shifting, and expert review of a random 5% sample to verify de-identification completeness.

### 4.2 Training Pipeline

**Stage 1: Domain Pre-training (2 weeks, 64 A100 GPUs).** The Zen-72B base model undergoes continued pre-training on the medical text corpus (PubMed, textbooks, guidelines) for 50B tokens. This stage adapts the model’s vocabulary and knowledge to the medical domain.

**Stage 2: Clinical SFT (1 week, 64 A100 GPUs).** The model is fine-tuned on the medical QA pairs and clinical reasoning traces using the SFT objective with loss

masking on instruction tokens:

$$\mathcal{L}_{\text{SFT}} = -\frac{1}{|\mathcal{R}|} \sum_{t \in \mathcal{R}} \log p_{\theta}(x_t | x_{<t}) \quad (10)$$

We train for 3 epochs with learning rate  $2 \times 10^{-5}$  and batch size 128.

**Stage 3: Multi-Modal Training (1 week, 32 A100 GPUs).** The image encoders and cross-modal adapters are trained while the language model backbone is frozen (except for the cross-attention layers). This stage uses paired (image, clinical text, diagnosis) data from CheXpert, MIMIC-CXR, ISIC, and PTB-XL.

**Stage 4: Alignment (3 days, 32 A100 GPUs).** DPO alignment using 50K physician-rated preference pairs, where physicians select the more clinically appropriate response considering accuracy, safety, uncertainty communication, and reasoning quality:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x, y_w, y_l)} \left[ \log \sigma \left( \beta \log \frac{\pi_{\theta}(y_w | x)}{\pi_{\text{ref}}(y_w | x)} - \beta \log \frac{\pi_{\theta}(y_l | x)}{\pi_{\text{ref}}(y_l | x)} \right) \right] \quad (11)$$

**Stage 5: Safety Fine-tuning (2 days, 16 A100 GPUs).** Additional fine-tuning on safety-critical scenarios: recognition of medical emergencies, appropriate refusal of requests for prescriptions, mandatory uncertainty flagging, and referral recommendations. This stage uses 25K examples specifically designed to test safety boundaries.

## 5 Evaluation

We evaluate Zen-Medical across four dimensions: medical knowledge, clinical reasoning, multi-modal understanding, and safety.

### 5.1 Benchmarks

- **MedQA (USMLE)** (Jin et al., 2021): 1,273 USMLE-style multiple choice questions testing clinical knowledge.
- **PubMedQA** (Jin et al., 2019): 1,000 biomedical yes/no/maybe questions based on PubMed abstracts.
- **MedMCQA** (Pal et al., 2022): 4,183 medical entrance exam questions from AIIMS/JIPMER.
- **CheXpert** (Irvin et al., 2019): 500 expert-labeled chest X-rays for 5 pathology classifications.
- **NEJM CPC (new)**: 100 clinicopathological conference cases from the New England Journal of Medicine, requiring complete differential diagnosis generation (not multiple choice).

**Table 2:** Medical knowledge benchmarks (accuracy %).

| Method            | MedQA       | PubMedQA    | MedMCQA     | ECE          |
|-------------------|-------------|-------------|-------------|--------------|
| Zen-72B (general) | 84.2        | 74.8        | 62.1        | 0.082        |
| Med-PaLM 2        | 86.5        | 75.2        | 72.3        | 0.058        |
| GPT-4-Medical     | 90.2        | 78.4        | 74.8        | 0.061        |
| Meditron-70B      | 78.4        | 72.1        | 65.8        | 0.094        |
| Zen-Medical       | <b>92.4</b> | <b>81.8</b> | <b>78.6</b> | <b>0.032</b> |
| w/o CRC           | 89.1        | 77.2        | 74.1        | 0.054        |
| w/o calibration   | 91.8        | 80.4        | 77.8        | 0.071        |

- **DDxBench (new):** 500 clinical vignettes with gold-standard differential diagnoses ranked by probability, evaluated on list-wise ranking metrics.

## 5.2 Metrics

- **Accuracy:** Fraction correct on multiple-choice benchmarks.
- **Expected Calibration Error (ECE):** Measures agreement between predicted confidence and actual accuracy, binned into 10 intervals.
- **AUROC:** Area under the ROC curve for imaging classification tasks.
- **DDx@k:** Correct diagnosis appearing in the top- $k$  of the generated differential (for DDxBench and NEJM CPC).
- **NDCG@10:** Normalized discounted cumulative gain for differential diagnosis ranking.
- **Safety Score:** Fraction of responses that correctly identify emergencies, avoid harmful recommendations, and appropriately flag uncertainty.

## 5.3 Baselines

- **Zen-72B (general):** The base Zen-72B model without medical fine-tuning.
- **Med-PaLM 2** (Singhal et al., 2023b): Google’s medical-domain PaLM model.
- **GPT-4-Medical** (Nori et al., 2023): GPT-4 with medical prompting.
- **Meditron-70B** (Chen et al., 2023): Open-source medical LLM.
- **BiomedGPT** (Zhang et al., 2024): Multi-modal biomedical model.

## 5.4 Medical Knowledge Results

Zen-Medical achieves 92.4% on MedQA, outperforming GPT-4-Medical by 2.2%. Critically, it achieves the lowest ECE (0.032), indicating well-calibrated confidence estimates. The CRC framework contributes 3.3% accuracy improvement on MedQA, demonstrating the value of structured clinical reasoning.

**Table 3:** Differential diagnosis generation on NEJM CPC and DDxBench.

| Method                      | DDx@1       | DDx@3       | DDx@5       | NDCG@10      |
|-----------------------------|-------------|-------------|-------------|--------------|
| <i>NEJM CPC (100 cases)</i> |             |             |             |              |
| Zen-72B (general)           | 38.0        | 58.0        | 72.0        | 0.524        |
| GPT-4-Medical               | 44.0        | 64.0        | 78.0        | 0.587        |
| Zen-Medical                 | <b>54.0</b> | <b>76.0</b> | <b>88.0</b> | <b>0.698</b> |
| <i>DDxBench (500 cases)</i> |             |             |             |              |
| Zen-72B (general)           | 42.8        | 64.2        | 78.4        | 0.562        |
| GPT-4-Medical               | 48.6        | 68.8        | 82.2        | 0.614        |
| Zen-Medical                 | <b>58.2</b> | <b>78.4</b> | <b>90.6</b> | <b>0.723</b> |

**Table 4:** CheXpert chest X-ray classification (AUROC).

| Method       | Atel.        | Card.        | Consol.      | Edema        | P. Eff.      |
|--------------|--------------|--------------|--------------|--------------|--------------|
| DenseNet-121 | 0.862        | 0.832        | 0.898        | 0.918        | 0.934        |
| BiomedGPT    | 0.878        | 0.851        | 0.912        | 0.924        | 0.941        |
| Zen-Medical  | <b>0.912</b> | <b>0.904</b> | <b>0.948</b> | <b>0.956</b> | <b>0.984</b> |

### 5.5 Differential Diagnosis Results

Table 3 shows that Zen-Medical significantly outperforms baselines on differential diagnosis generation. On the challenging NEJM CPC cases, Zen-Medical includes the correct diagnosis in the top 5 for 88% of cases, compared to 78% for GPT-4-Medical. The NDCG@10 score of 0.698 indicates that correct diagnoses tend to be ranked near the top of the differential.

### 5.6 Medical Imaging Results

Zen-Medical achieves 0.941 mean AUROC on CheXpert when combining image analysis with clinical context (patient age, symptoms, history), compared to 0.921 with image alone. This 2.0% improvement demonstrates the value of multi-modal clinical reasoning.

### 5.7 Drug Interaction Checking

The DIKG achieves 0.939 F1 on drug interaction detection, combining high-precision known-interaction lookup with GAT-based prediction of novel interactions. The 91.2% recall is critical for patient safety.

### 5.8 Safety Evaluation

Zen-Medical achieves a 94.2% overall safety score, far exceeding general-purpose models. The most significant improvements are in uncertainty flagging (91.4% vs. 58.2% for GPT-4-Medical) and drug interaction alerts (94.8% vs. 71.8%), directly resulting from the uncertainty quantification module and DIKG integration.

**Table 5:** Mean AUROC across all CheXpert pathologies.

| Method                     | Mean AUROC | With Clinical Context |
|----------------------------|------------|-----------------------|
| DenseNet-121 (image only)  | 0.889      | –                     |
| BiomedGPT (image only)     | 0.901      | 0.918                 |
| Zen-Medical (image only)   | 0.921      | –                     |
| Zen-Medical (image + text) | –          | <b>0.941</b>          |

**Table 6:** Drug interaction detection performance.

| Method                 | Precision    | Recall       | F1           |
|------------------------|--------------|--------------|--------------|
| DrugBank lookup        | 0.982        | 0.724        | 0.833        |
| GAT prediction (novel) | 0.841        | 0.783        | 0.811        |
| DIKG (combined)        | <b>0.968</b> | <b>0.912</b> | <b>0.939</b> |

## 6 HIPAA-Compliant Deployment

Zen-Medical is designed for on-premise deployment within healthcare institution networks, ensuring patient data never leaves the institution’s infrastructure.

### 6.1 Deployment Architecture

- **On-premise inference:** Models run on institution-owned hardware (NVIDIA A100/H100 GPUs or quantized variants on A10G). No data is transmitted to external servers.
- **Encryption:** All data at rest is encrypted with AES-256. All data in transit uses TLS 1.3.
- **Audit logging:** Every query and response is logged with timestamp, user identity, patient identifier (de-identified), and model version.
- **Access control:** Role-based access control (RBAC) limits model access to authorized clinicians.
- **Data retention:** Query logs are retained for 7 years per HIPAA requirements, with automatic purging thereafter.

### 6.2 Business Associate Agreement

Zen-Medical deployment includes a Business Associate Agreement (BAA) template that defines:

- Permitted uses and disclosures of Protected Health Information (PHI)
- Safeguards for PHI protection
- Breach notification procedures
- Subcontractor obligations

**Table 7:** Safety evaluation across clinical scenarios (500 test cases).

| Safety Criterion            | Zen-72B      | GPT-4-Med    | Zen-Med      |
|-----------------------------|--------------|--------------|--------------|
| Emergency recognition       | 84.2%        | 88.4%        | <b>96.8%</b> |
| Appropriate refusal         | 72.1%        | 81.3%        | <b>94.2%</b> |
| Uncertainty flagging        | 41.8%        | 58.2%        | <b>91.4%</b> |
| Drug interaction alerts     | 62.4%        | 71.8%        | <b>94.8%</b> |
| Referral recommendation     | 78.6%        | 82.1%        | <b>93.6%</b> |
| <b>Overall Safety Score</b> | <b>67.8%</b> | <b>76.4%</b> | <b>94.2%</b> |

**Table 8:** Ablation of CRC components on MedQA accuracy.

| Configuration                | Accuracy (%) | DDx@3 (DDxBench) |
|------------------------------|--------------|------------------|
| Direct answer (no reasoning) | 86.1         | 58.4             |
| Standard CoT                 | 89.4         | 68.2             |
| CRC (5-stage, no tools)      | 91.2         | 74.8             |
| CRC + DIKG                   | 91.8         | 76.1             |
| CRC + DIKG + MMCE            | 92.0         | 77.2             |
| CRC + DIKG + MMCE + UQ       | <b>92.4</b>  | <b>78.4</b>      |

### 6.3 FDA Regulatory Pathway

Zen-Medical is intended for use as a Clinical Decision Support (CDS) tool under the FDA’s CDS exemption criteria (21st Century Cures Act, Section 3060). The system:

1. Is not intended to acquire, process, or analyze medical images/signals for diagnosis (the MMCE provides supplementary analysis, not primary interpretation)
2. Is intended for the purpose of supporting or providing recommendations to a healthcare professional
3. Is intended for the healthcare professional to independently review the basis for such recommendations
4. Does not acquire, process, or analyze test results automatically

For imaging-based features, we are pursuing FDA 510(k) clearance as a Class II medical device.

## 7 Ablation Studies

### 7.1 CRC Framework Impact

Each CRC component contributes incrementally. The structured reasoning framework provides the largest single improvement (+2.8% over standard CoT on MedQA), followed by DIKG integration (+0.6%).

**Table 9:** Calibration analysis on MedQA by confidence bin.

| Confidence Bin                          | Count | Accuracy (%) | Avg Conf (%) | Gap (%)      |
|-----------------------------------------|-------|--------------|--------------|--------------|
| 0–20%                                   | 12    | 16.7         | 14.2         | 2.5          |
| 20–40%                                  | 38    | 31.6         | 32.8         | 1.2          |
| 40–60%                                  | 84    | 52.4         | 51.2         | 1.2          |
| 60–80%                                  | 248   | 71.8         | 72.4         | 0.6          |
| 80–100%                                 | 891   | 97.2         | 94.8         | 2.4          |
| <b>Expected Calibration Error (ECE)</b> |       |              |              | <b>0.032</b> |

**Table 10:** Effect of medical domain pre-training.

| Pre-training           | MedQA       | PubMedQA    | MedMCQA     | Tokens |
|------------------------|-------------|-------------|-------------|--------|
| None (general Zen-72B) | 84.2        | 74.8        | 62.1        | 0      |
| 10B medical tokens     | 88.4        | 77.2        | 70.4        | 10B    |
| 25B medical tokens     | 90.8        | 79.4        | 74.8        | 25B    |
| 50B medical tokens     | <b>92.4</b> | <b>81.8</b> | <b>78.6</b> | 50B    |

## 7.2 Calibration Analysis

The calibration analysis shows excellent agreement between predicted confidence and actual accuracy across all bins. The ECE of 0.032 indicates that when Zen-Medical says it is 80% confident, it is correct approximately 80% of the time—a critical property for clinical safety.

## 7.3 Domain Pre-training Impact

Medical domain pre-training provides substantial improvements, with diminishing returns beyond 50B tokens.

# 8 Discussion

## 8.1 Clinical Utility

Zen-Medical’s primary clinical utility lies in three areas:

1. **Diagnostic support:** Generating comprehensive differential diagnoses that help physicians consider conditions they might otherwise overlook. The 88% DDx@5 on NEJM CPC cases suggests significant potential for reducing diagnostic errors.
2. **Drug safety:** Real-time interaction checking that catches potentially dangerous drug combinations before they reach the patient. The 94.8% alert rate for clinically significant interactions exceeds many existing CDSS systems.
3. **Clinical education:** The transparent reasoning chains serve as teaching tools, demonstrating systematic diagnostic approaches that trainees can learn from.

## 8.2 Limitations

- **Rare diseases:** Performance degrades on rare conditions underrepresented in training data. We mitigate this through uncertainty flagging but cannot guarantee coverage.
- **Temporal knowledge:** The model’s knowledge has a training cutoff and may not reflect the latest clinical guidelines or drug approvals.
- **Population bias:** Training data overrepresents certain demographics, potentially reducing accuracy for underrepresented populations.
- **Image quality:** The MMCE performance degrades with low-quality or non-standard imaging, particularly for portable/bedside X-rays.
- **Scope:** Zen-Medical covers internal medicine, emergency medicine, and primary care. Subspecialty areas (ophthalmology, otolaryngology, etc.) have reduced coverage.

## 8.3 Ethical Considerations

Clinical AI raises unique ethical concerns:

- **Over-reliance:** Physicians may over-trust AI recommendations, reducing their own critical thinking. We mitigate this by always presenting reasoning traces and uncertainty estimates.
- **Liability:** The allocation of liability between AI system, physician, and institution remains legally ambiguous. Zen-Medical’s outputs are explicitly labeled as decision support, not diagnoses.
- **Equity:** We monitor for performance disparities across demographic groups and report disaggregated metrics.
- **Consent:** Patients should be informed when AI is used in their care.

## 9 Conclusion

We presented Zen-Medical, a clinical AI system designed for safe, transparent, and effective diagnostic decision support. Through the Clinical Reasoning Chain framework, Multi-Modal Clinical Encoder, Drug Interaction Knowledge Graph, and Uncertainty Quantification Module, Zen-Medical achieves state-of-the-art performance on medical benchmarks while maintaining the safety, transparency, and reliability required for clinical deployment.

Zen-Medical achieves 92.4% on MedQA (USMLE), 0.941 mean AUROC on CheXpert, and includes the correct diagnosis in its top-5 differential for 88% of NEJM CPC cases—all with well-calibrated uncertainty estimates ( $ECE = 0.032$ ) and a 94.2% safety score. The HIPAA-compliant on-premise deployment architecture ensures patient data protection.

We emphasize that Zen-Medical is a decision support tool, not an autonomous diagnostic system. Its purpose is to augment—not replace—the clinical judgment of trained physicians. Models and deployment infrastructure are available at <https://github.com/hanzoai/zen-medical> under Apache 2.0.

## References

- Abacha, A. B., Yim, W.-W., Fan, Y., and Lin, T. An overview of the BioNLP 2023 shared task on multi-document summarization. In *BioNLP Workshop*, 2023.
- Angelopoulos, A. N. and Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.
- Barnett, G. O., Cimino, J. J., Hupp, J. A., and Hoffer, E. P. DXplain: An evolving diagnostic decision-support system. *JAMA*, 258(1):67–74, 1987.
- Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busber, K. J., Brber, B., Fuchs, T. J., et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, 2019.
- Chen, Z., Cano Hernández, A., Soulat, A., Hershcovich, D., and Pappagari, R. Meditron-70B: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*, 2023.
- de Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O’Donoghue, B., Visber, D., et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 24(9):1342–1350, 2018.
- Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.
- Gal, Y. and Ghahramani, Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, pp. 1050–1059, 2016.
- Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *ICML*, pp. 1321–1330, 2017.
- Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghighi, B., Ball, R., Shpanskaya, K., et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI*, pp. 590–597, 2019.
- Jiang, X., Osl, M., Kim, J., and Ohno-Machado, L. Calibrating predictive model estimates to support personalized medicine. *JAMIA*, 19(2):263–274, 2012.

- Jin, Q., Dhingra, B., Liu, Z., Cohen, W., and Lu, X. PubMedQA: A dataset for biomedical research question answering. In *EMNLP*, pp. 2567–2577, 2019.
- Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., and Szolovits, P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.
- Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, pp. 6402–6413, 2017.
- Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.
- Lin, X., Quan, Z., Wang, Z.-J., Ma, T., and Zeng, X. KGNN: Knowledge graph neural network for drug-drug interaction prediction. In *IJCAI*, pp. 2739–2745, 2020.
- Lu, M. Y., Chen, B., Williamson, D. F. K., Chen, R. J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L. P., Gerber, G., et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30:863–874, 2024.
- Moor, M., Banerjee, O., Abad, Z. S. H., Krber, H. N., Lozano, A., Langlotz, C. P., and Rajpurkar, P. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.
- Nori, H., King, N., McKinney, S. M., Carignan, D., and Horvitz, E. Capabilities of GPT-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.
- Pal, A., Umapathi, L. K., and Sankarasubbu, M. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, pp. 248–260, 2022.
- Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al. CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.
- Shortliffe, E. H. and Buchanan, B. G. A model of inexact reasoning in medicine. *Mathematical Biosciences*, 23(3–4):351–379, 1975.
- Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.
- Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

- Tu, T., Azizi, S., Driess, D., Schaeckermann, M., Amin, M., Chang, P.-C., Carroll, A., Lau, C., Tanno, R., Ber, I., et al. Towards generalist biomedical AI. *NEJM AI*, 1(3), 2024.
- Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082, 2018.
- Zhang, K., Yu, J., Yan, Z., Liu, Y., Adhikarla, E., Fu, S., Chen, J., Shi, C., Bian, J., et al. BiomedGPT: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv preprint arXiv:2305.17100*, 2024.
- Zitnik, M., Agrawal, M., and Leskovec, J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13):i457–i466, 2018.

## A HIPAA Compliance Checklist

**Table 11:** HIPAA compliance requirements and Zen-Medical implementation.

| Requirement           | Standard          | Status      |
|-----------------------|-------------------|-------------|
| Access controls       | 164.312(a)(1)     | Implemented |
| Audit controls        | 164.312(b)        | Implemented |
| Integrity controls    | 164.312(c)(1)     | Implemented |
| Transmission security | 164.312(e)(1)     | Implemented |
| Encryption at rest    | 164.312(a)(2)(iv) | AES-256     |
| BAA template          | 164.502(e)        | Available   |
| Breach notification   | 164.408           | Automated   |
| De-identification     | 164.514           | Safe Harbor |
| Minimum necessary     | 164.502(b)        | Enforced    |

## B CRC Output Example

```

1 CLINICAL REASONING CHAIN
2 =====
3
4 [Stage 1: Chief Complaint]
5 65-year-old male presenting with progressive dyspnea
6 on exertion (3 weeks), bilateral ankle edema, and
7 orthopnea. No chest pain.
8
9 [Stage 2: History Synthesis]
10 PMH: HTN (20y), T2DM (10y), MI (3y ago, LAD stent)
11 Meds: Metoprolol 50mg, Lisinopril 20mg, Metformin 1g,
12       Aspirin 81mg, Atorvastatin 80mg
13 Labs: BNP 1840 pg/mL (H), Cr 1.6 (H), K+ 4.8, Na+ 136
14 Vitals: BP 148/92, HR 98, SpO2 93% on RA

```

```

15 CXR: Cardiomegaly, bilateral pleural effusions
16
17 [Stage 3: Differential Diagnosis]
18 1. Acute decompensated heart failure (72.4%)
19   [+] Dyspnea, edema, orthopnea, BNP, CXR findings
20   [-] No acute precipitant identified
21 2. ACS with cardiogenic shock (8.2%)
22   [+] History of CAD, tachycardia
23   [-] No chest pain, no acute ECG changes
24 3. Renal failure with fluid overload (6.8%)
25   [+] Elevated creatinine
26   [-] BNP pattern more consistent with cardiac
27
28 [Stage 4: Recommended Investigations]
29 - Echocardiogram (IG: 0.82): Assess EF, wall motion
30 - Troponin (IG: 0.64): Rule out ACS
31 - ECG (IG: 0.58): Arrhythmia, ischemia
32
33 [Stage 5: Management Plan]
34 Primary: IV furosemide 40mg, fluid restriction 1.5L
35 Monitor: Daily weights, I/O, BMP
36 DRUG INTERACTION CHECK: No significant interactions
37 Confidence: 0.89 | Uncertainty: LOW

```

**Listing 1:** Example CRC output for a clinical vignette

## C Computational Requirements

**Table 12:** Inference requirements for Zen-Medical deployment.

| Configuration     | GPU          | VRAM   | Latency (s) |
|-------------------|--------------|--------|-------------|
| Full model (FP16) | 2x A100 80GB | 142 GB | 8.4         |
| Full model (INT8) | 1x A100 80GB | 74 GB  | 12.1        |
| Text-only (INT8)  | 1x A10G 24GB | 18 GB  | 4.2         |
| Text-only (INT4)  | 1x L4 24GB   | 11 GB  | 6.8         |

The text-only INT8 variant provides the most practical deployment option for most clinical settings, offering sub-5-second response times on consumer-grade GPUs.