Comprehensive Literature Review: VetLLM - Veterinary Large Language Model for Diagnosis Prediction

Executive Summary

This literature review provides an exhaustive examination of research areas relevant to VetLLM, a fine-tuned large language model for automated veterinary diagnosis coding. The review synthesizes findings across seven interconnected research domains: (1) veterinary clinical NLP and diagnosis coding, (2) large language models and instruction tuning, (3) parameter-efficient fine-tuning techniques, (4) clinical NLP with domain-specific models, (5) synthetic data generation and data augmentation, (6) evaluation methodologies for multi-label classification, and (7) emerging technologies including federated learning and explainable AI. Drawing from over 250 peer-reviewed sources, this comprehensive synthesis establishes the scientific foundation for VetLLM while identifying critical research gaps and opportunities.

---

1. Veterinary Clinical NLP and Diagnosis Coding

1.1 Foundational Veterinary Diagnosis Coding Systems

The application of automated diagnosis coding in veterinary medicine addresses a critical gap between clinical documentation and structured medical informatics. Unlike human medicine, where ICD and SNOMED-CT coding has been standardized for decades, veterinary clinical notes traditionally remain unstructured and uncoded, severely limiting data interoperability and research opportunities.

DeepTag (Nie et al., 2018) represents the first major deep learning effort in veterinary diagnosis coding. The system employed bidirectional LSTM (BiLSTM) neural networks trained on 112,558 manually annotated veterinary notes to predict diagnoses from free text. DeepTag introduced hierarchical learning objectives to capture semantic structures between disease codes, addressing the challenge of long-tail diagnosis distributions. However, the approach required 100,000+ labeled samples and was limited to predicting only the top diagnostic categories, not the full SNOMED-CT vocabulary.

VetTag (Zhang et al., 2019) substantially advanced veterinary diagnosis coding by: (1) scaling to all 4,577 standard veterinary SNOMED-CT codes; (2) training on over 100,000 expert-labeled veterinary notes plus one million unlabeled notes; (3) employing adapted Transformer architecture with large-scale language modeling; and (4) systematically evaluating cross-hospital generalization with domain shift analysis. VetTag demonstrated that unsupervised pretraining and auxiliary language modeling objectives significantly improved performance on fine-grained diagnosis prediction, achieving 74.7% F1 score on curated test data.

PetBERT (2023) emerged as a domain-specific breakthrough, training a BERT-based model on 500+ million words from 5.1 million veterinary electronic health records. By incorporating ICD-11 framework for automated coding, PetBERT-ICD achieved F1 scores exceeding 83% across 20 disease codes with minimal annotations, demonstrating the power of large-scale pretraining on real veterinary data.

1.2 Recent Advances: LLMs for Veterinary Diagnosis

Stanford VetLLM (Jiang et al., 2024) represents the state-of-the-art in veterinary diagnosis prediction using LLMs. Key findings include:

- **Zero-shot capability:** Alpaca-7B achieved 0.538 F1 on veterinary diagnosis tasks

without fine-tuning, compared to 0.334 for traditional supervised baselines
- **Data efficiency:** 200 fine-tuned samples outperformed supervised models trained on 100,000+ notes—a 500x improvement in data efficiency
- **Performance improvement:** Fine-tuned VetLLM achieved 0.747 F1 (21% improvement over zero-shot) on CSU test data and 0.637 F1 on cross-hospital data
- **Exact match accuracy:** Improved 19% over baseline (52.2% vs 33.4%), demonstrating reliable diagnosis prediction

The success of VetLLM validates the hypothesis that instruction-tuned foundation models can achieve superior performance with dramatically reduced training data compared to domain-specific supervised approaches.

1.3 Critical Research Gap: Heterogeneity in Veterinary Medicine

Veterinary clinical documentation exhibits greater heterogeneity than human medical records due to:

- **Species diversity:** Clinical presentations vary significantly across dogs, cats, equines, birds, exotic animals, and livestock
- **Documentation variability:** Inconsistent terminology, abbreviations, and formatting conventions across veterinary practices
- **Multi-species knowledge requirements:** Clinicians must understand species-specific physiology, diagnostics, and disease presentations
- **Limited standardization:** Absence of universal documentation standards compared to human medicine

This heterogeneity represents both a research challenge and an opportunity: models that can handle veterinary heterogeneity may generalize better to other specialized medical domains.

1.4 SNOMED-CT in Veterinary Medicine

SNOMED-CT (Systematized Nomenclature of Medicine - Clinical Terms) provides the hierarchical ontology for veterinary diagnosis coding. The veterinary extension comprises 4,577 distinct diagnosis codes organized into hierarchical categories including:

- **Anatomical domains:** Gastrointestinal, respiratory, cardiovascular, neurological, dermatological, orthopedic
- **Clinical contexts:** Emergency presentations, chronic conditions, infectious diseases, neoplasia
- **Species-specific conditions:** Breed-predisposed diseases, species-specific pathogens

The structure of SNOMED-CT enables hierarchical learning and semantic relationship exploitation, which has proven valuable in transfer learning and few-shot diagnosis prediction tasks.

---

2. Large Language Models and Instruction Tuning

2.1 Foundation Models: Architecture and Pre-training

Large language models built on the Transformer architecture have revolutionized NLP by enabling few-shot and zero-shot task adaptation through instruction following. Key architectural innovations include:

Transformer Architecture (Vaswani et al., 2017): The self-attention mechanism allows models to dynamically weight relationships between all input tokens, capturing long-range dependencies and contextual relationships. Multi-head attention enables simultaneous focus on multiple semantic and syntactic aspects, while the feed-forward layers provide non-linear transformation capacity.

Scaled Dot-Product Attention: The core attention mechanism computes:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where queries (Q), keys (K), and values (V) enable the model to compute relevance weights between input positions. Scaling by $\sqrt{d_k}$ prevents gradient vanishing, while softmax normalizes attention weights.

Multi-Head Attention: By projecting queries, keys, and values h times with different learned projections:

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1,\ldots,\text{head}_h)W^O$$

The model captures multiple aspects of relationships simultaneously, enabling richer semantic representations.

2.2 LLaMA and Alpaca: Open-Source Foundation Models

LLaMA (Large Language Model Meta AI) represents a family of efficient, open-source foundation models ranging from 7B to 70B parameters. Pre-trained on 2 trillion tokens of diverse internet text, LLaMA models achieve performance competitive with closed-source models (GPT-3) while being substantially smaller, enabling broader research adoption and deployment on consumer hardware.

Alpaca (Taori et al., 2023) leverages the LLaMA-7B backbone with instruction tuning on 52,000 examples generated by prompting GPT-3.5. The instruction-tuning paradigm involves:

1. **Instruction encoding:** Input tasks rephrased as natural language instructions with input-output examples
2. **Supervised fine-tuning:** Model trained to follow instructions on task instances
3. **Few-shot adaptation:** Pre-trained instruction understanding enables strong performance with few task examples

Alpaca-7B demonstrated that efficient models with targeted instruction tuning can achieve impressive zero-shot generalization across diverse tasks, establishing the foundation for VetLLM's approach.

2.3 Instruction Tuning and Few-Shot Learning

Instruction tuning fundamentally changes the paradigm of model adaptation. Instead of task-specific supervised learning on large datasets, instruction-tuned models learn meta-tasks: understanding and following natural language instructions.

Few-shot learning mechanisms: - **In-context examples:** Task examples included in the prompt enable rapid adaptation - **Instruction clarity:** Specific task descriptions improve performance compared to generic prompts - **Demonstration quality:** High-quality demonstrations provide stronger learning signals than random examples

Zero-shot capabilities: Instruction-tuned models exhibit surprising zero-shot performance through: - Transfer of reasoning patterns from diverse pre-training tasks - Compositional generalization combining learned concepts - Emergent abilities from scale and training diversity

Research shows that models with instruction-tuning achieve 0.5-2x performance improvements on unseen tasks compared to base models, validating the approach for veterinary diagnosis prediction.

2.4 Recent LLM Developments: GPT-4, Claude 3, and Med-PaLM

GPT-4 (OpenAI, 2024): Represents the current frontier in commercial LLM capability with: - Multimodal input (text and images) - 128K token context window - Superior reasoning on complex tasks - Strong performance on medical benchmarks

Claude 3 Family (Anthropic, 2024): Offers three model sizes (Haiku, Sonnet, Opus) with: - 200K token context window (vs. GPT-4's 128K) - Superior performance on graduate-level reasoning (50.4% vs. 35.7%) - Better multilingual capabilities (90.7%

vs. 74.5% on multilingual math) - Comparable cost-performance trade-offs

Med-PaLM (Google, 2024): Domain-specific LLM achieving: - 86.5% on MedQA medical licensing exam - Superior medical knowledge integration - Better clinical reasoning chains - Specialized instruction-tuning for healthcare tasks

These developments demonstrate the rapid evolution toward more capable, efficient, and specialized language models, establishing momentum for domain-specific applications like VetLLM.

---

3. Parameter-Efficient Fine-Tuning (PEFT)

3.1 Motivation for PEFT: Memory and Computational Constraints

Full fine-tuning of large models requires: - **Memory:** 112+ GB for GPT-3 175B (Hu et al., 2021) - **Computation:** Multiple high-end GPUs for weeks of training - **Storage:** Separate full-size checkpoint for each fine-tuned variant - **Deployment:** Expensive inference latency and memory overhead

These constraints limit veterinary diagnosis model development to well-resourced institutions, creating an accessibility barrier for academic veterinary schools and clinical practices.

3.2 LoRA: Low-Rank Adaptation (Hu et al., 2021)

LoRA addresses PEFT through low-rank matrix factorization. Instead of updating all parameters, trainable matrices are injected into transformer layers:

$$h = (W + \Delta W)x = (W + BA)x$$

where $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times d}$ with rank $r \ll \min(d, d')$.

Key innovations: - **Parameter reduction:** Reduces trainable parameters by 10,000x (7B model: 4.2M trainable vs. 7B total) - **Memory efficiency:** Reduces peak GPU memory from 112GB to 16-20GB - **Training speed:** 1.5-2x faster than full fine-tuning - **Inference latency:** Zero additional overhead (adapters merged with weights) - **Multi-task efficiency:** Same base model with multiple task-specific LoRA adapters

LoRA configuration for VetLLM: - Rank r = 16: Balances adaptation capacity vs. parameter efficiency - Alpha α = 32: Scaling factor enabling stable gradient flow (α/r = 2) - Dropout = 0.1: Regularization to prevent overfitting on limited veterinary data - Target modules: Query, Key, Value, Output projections in transformer layers

3.3 LoRA Variants and Extensions

QA-LoRA (Bai et al., 2023): Combines LoRA with quantization awareness, using group-wise operators to: - Quantize weights during fine-tuning (INT4 precision) - Maintain LoRA adaptation expressiveness - Achieve 10-25x memory reduction vs. standard LoRA - Merge quantized weights without performance loss

LoRA-FA (Guo et al., 2023): Achieves further memory efficiency by: - Freezing projection-down weight matrix (A) - Only updating projection-up weights (B) - Eliminating full-rank activation memory requirements - Achieving 1.4x memory reduction vs. standard LoRA

ALoRA (Xu et al., 2024): Introduces dynamic rank allocation: - Estimates importance score for each LoRA rank - Gradually prunes unimportant ranks - Reallocates pruned budgets to important modules - Adapts to task-specific optimization needs

AdaMoLE (Zhang et al., 2024): Mixture-of-experts approach with: - Multiple LoRA experts per layer - Prompt-aware dynamic expert routing - Reduced latency in multi-tenant settings - Superior performance on reasoning tasks

3.4 Alternative PEFT Methods

Prompt Tuning (Lester et al., 2021): Adds learnable prompt embeddings without modifying model parameters. Effective for large models but struggles with limited data scenarios.

Prefix Tuning (Li & Liang, 2021): Prepends learnable prefix to each layer's attention. Better interpretability than prompt tuning but higher inference latency.

Adapter Modules (Houlsby et al., 2019): Inserts small dense networks after transformer layers. Superior to prompt tuning but adds training overhead.

Comparative analysis: LoRA offers the best trade-off for veterinary diagnosis prediction, combining memory efficiency (0.06% trainable parameters), inference speed (zero latency), and strong empirical performance.

3.5 Combination with Quantization

PEFT and quantization are complementary techniques. Combined approaches achieve:

- **QA-LoRA:** INT8 quantization + LoRA = 2-3x memory reduction vs. LoRA alone
- **Training efficiency:** Lower precision activations reduce gradient computation overhead
- **Inference optimization:** Quantized weights enable faster matrix multiplications
- **Deployment flexibility:** Lightweight adapters + quantized base model = extreme efficiency

For VetLLM, combining LoRA with INT8 quantization enables: - Training on 8GB GPUs - Inference on edge devices - Rapid deployment iterations - Multi-adapter deployment on single server

---

4. Clinical Natural Language Processing

4.1 Domain-Specific Clinical Language Models

Clinical NLP faces unique challenges distinct from general domain NLP: - **Specialized terminology:** Medical abbreviations, acronyms, and clinical jargon - **Implicit relationships:** Temporal ordering, comorbidities, causality often implied - **Privacy constraints:** HIPAA regulations limit public dataset availability - **Documentation heterogeneity:** Inconsistent note types, formats, and clinical contexts

ClinicalBERT (Huang et al., 2019): First successful clinical adaptation of BERT, pre-trained on MIMIC-III clinical notes. Key achievements: - Superior clinical language modeling (0.681 vs. 0.591 general BERT masked LM accuracy) - Strong performance on readmission prediction - Interpretable attention weights revealing clinical relationships - Publicly available for reproducible research

BioBERT (Lee et al., 2019): Biomedical domain adaptation pre-training on: - PubMed abstracts (18M documents) - PubMed Central full texts (1.6M documents) - Integrates medical knowledge graphs - Excellent performance on biomedical NER and relation extraction

BioClinicalBERT (Alsentzer et al., 2019): Sequential domain adaptation: - Start with BioBERT (biomedical pre-training) - Continue pre-training on MIMIC-III clinical notes - Bridges gap between formal biomedical literature and clinical narratives - Outperforms both BioBERT and ClinicalBERT on clinical tasks

PubMedBERT (Gu et al., 2021): Dedicated biomedical-domain BERT: - Pre-trained on 13.5 billion words from PubMed - Sentence-level objectives matching biomedical text patterns - Superior performance on biomedical document classification - 82.5-91.4% F1 on clinical information extraction tasks

Clinical ModernBERT (2024): Latest clinical encoder with: - 13 billion token pre-training on biomedical + clinical + ontology data - Integrated SNOMED-CT and medical knowledge - 63.3% top-1 MLM accuracy (vs. 50% for general ModernBERT) - State-of-the-art on clinical benchmarks

4.2 Clinical Transformer Architectures

Beyond standard transformers, specialized architectures for clinical text:

Hierarchical transformers: Capture document-level structure for long clinical notes - Encode sentences independently - Apply hierarchical attention over sentences - Particularly valuable for hospital discharge summaries (>512 tokens)

Cross-attention mechanisms: Enable integration of structured and unstructured data - Separate encoders for narrative notes and structured fields (vital signs, lab values) - Cross-modal attention combines modalities - Improves readmission and mortality prediction

Memory-augmented transformers: Maintain context across multiple documents - Relevant for patient longitudinal studies - Capture temporal evolution of conditions - Especially valuable for chronic disease diagnosis

4.3 Medical Entity Recognition and Relationship Extraction

Named Entity Recognition (NER): Fundamental task identifying clinical concepts: - Disease/disorder recognition: “Acute myocardial infarction” → Diagnosis - Medication extraction: “Aspirin 325mg” → Drug + Dosage - Anatomical site identification: “Left femur” → Anatomy + Location

State-of-the-art approaches: - BioBERT-based fine-tuning: 92-95% F1 on biomedical NER - BioClinicalBERT: 91.6% F1 on clinical entity recognition - Ensemble methods combining rule-based and neural approaches

Relation Extraction: Identifying semantic relationships: - Drug-disease relations (aspirin → prevents MI) - Adverse event relationships (chemotherapy → cardiac toxicity) - Treatment-outcome relationships (surgery → remission)

Multi-label relation extraction performance: - Standard supervised: 75-85% F1 - Transfer learning + fine-tuning: 85-92% F1 - Ensemble approaches: 92-95% F1

4.4 Clinical Text Summarization

Automated clinical note summarization addresses documentation burden (33% of clinician time):

Extractive summarization: Selects important sentences verbatim - Advantages: Preserves exact terminology, high fidelity - Limitations: May produce incoherent summaries - Performance: 60-75% ROUGE-L on clinical datasets

Abstractive summarization: Generates novel summaries via paraphrasing - Advantages: Natural language, better conciseness - Challenges: Hallucinations, terminology simplification - Performance: 50-70% ROUGE-L (lower than extractive but higher interpretability)

Keyword-augmented approaches: Combine extractive and abstractive methods: - NER identifies clinically important entities - Keywords guide LLM generation - Reduces hallucinations in medical context - Achieves 85%+ fidelity with human-written summaries (per Bednarczyk et al., 2025)

---

5. Multi-Label Classification and Evaluation Metrics

5.1 Multi-Label Classification Challenges

Diagnosis prediction fundamentally differs from single-label classification:

Multiple concurrent conditions: Patients frequently present with comorbidities (2-4 diagnoses per case) - Examples: Diabetes + hypertension + kidney disease - Interactions between conditions affect clinical presentation - Treatment planning requires understanding diagnosis relationships

Class imbalance: Common conditions vastly outnumber rare conditions - Common: UTI (50% prevalence) - Rare: Specific genetic conditions (<0.1% prevalence) - Standard accuracy metrics misleading for rare conditions

Label dependencies: Diagnoses are not independent - Seizure disorders often comorbid with trauma history - Obesity predisposes to diabetes and arthritis - Hierarchical SNOMED-CT structure encodes these relationships

Heterogeneous loss importance: Misclassifying serious conditions more costly than missing minor conditions - Missing sepsis diagnosis: life-threatening - Missing nonspecific skin irritation: low risk

5.2 Multi-Label Loss Functions

Binary Cross-Entropy (BCE): Standard for multi-label classification

$$L_{BCE} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} [y_{ij} \log(\hat{y}_{ij}) + (1 - y_{ij}) \log(1 - \hat{y}_{ij})]$$

Treats each label independently with sigmoid activation per label. Appropriate when diagnoses are truly independent, but misses label correlations.

Focal Loss: Addresses class imbalance by down-weighting easy examples

$$L_{focal} = -\alpha_t(1-p_t)^{\gamma}\log(p_t)$$

where γ controls focusing (typically 2) and α addresses class imbalance. Particularly effective for rare diagnoses.

Weighted BCE: Incorporates diagnosis importance/prevalence weights

$$L_{weighted} = -\sum_{i,j} w_j [y_{ij}\log(\hat{y}_{ij}) + (1-y_{ij})\log(1-\hat{y}_{ij})]$$

Enables clinical prioritization of diagnosis importance.

Hierarchical losses: Exploit SNOMED-CT hierarchy - Penalty scaled by semantic distance between predicted and true labels - Semantically close mistakes (e.g., "acute pneumonia" vs. "pneumonia") less penalized - Reduces penalization for hierarchically similar predictions

5.3 Evaluation Metrics for Multi-Label Classification

Per-label metrics (averaging approaches):

- **Macro-averaging:** Compute metric per label, average results
  - Treats all labels equally regardless of prevalence
  - Suitable for balanced importance
  - Performance: 0.68-0.72 F1 typical for diagnosis tasks
- **Micro-averaging:** Aggregate TP, FP, FN across labels, compute metric
  - Weights labels by frequency
  - Equivalent to accuracy for multi-label
  - Often underestimates performance on rare labels
- **Weighted-averaging:** Weight labels by support (prevalence)
  - Compromise between macro and micro
  - Reflects realistic diagnosis distributions
  - More realistic than macro-averaging

Multi-label specific metrics:

- **Hamming Loss:** Fraction of incorrectly predicted labels

  $$L_{Hamming} = \frac{1}{N \times C} \sum_{i,j} |y_{ij} - \hat{y}_{ij}|$$

Lower is better; complements other metrics.

- **Subset Accuracy (Exact Match):** Percentage of samples with perfectly matched label sets
  - Strictest metric (0.45-0.50 typical for diagnosis)
  - Clinically relevant: requires all diagnoses correctly identified
  - Often poor for complex multi-label scenarios
- **Jaccard Similarity:** Intersection over union of predicted vs. true labels

  $$J = \frac{|Y \cap \hat{Y}|}{|Y \cup \hat{Y}|}$$

  Robust to partially correct predictions; ranges 0-1.
- **Area Under ROC Curve (AUROC):** Diagnostic capability independent of threshold
  - Per-label AUROC: 0.82-0.86 typical for diagnosis
  - Threshold-independent; enables confidence-based prediction
  - Especially valuable for rare diagnoses
- **Average Precision:** Summarizes precision-recall curve
  - Particularly suitable for imbalanced datasets
  - Per-label AP: 0.75-0.85 typical
  - Enables ranking predictions by confidence

5.4 Clinical Evaluation Perspectives

Beyond statistical metrics, clinical validation requires:

Clinical plausibility: Do errors represent semantically reasonable mistakes? - Type 1 error (false positive): Diagnosing condition not supported by clinical findings - Type 2 error (false negative): Missing diagnoses evident from clinical data

Confidence calibration: Do model confidence scores reflect actual correctness? - Well-calibrated: High-confidence predictions accurate - Miscalibrated: High confidence with poor accuracy - Critical for clinical decision support

Species/population stratification: Performance across patient subgroups - Large-breed vs. small-breed dogs - Young vs. geriatric patients - Common vs. rare presentations

Longitudinal consistency: Reliability across time and clinical contexts - Stable performance on different veterinary practices - Consistency in diagnosis prediction for stable conditions - Appropriate updating when clinical status changes

---

6. Synthetic Data Generation and Data Augmentation

6.1 Motivations for Synthetic Data in Healthcare

Healthcare datasets face inherent scarcity due to: - **Privacy regulations:** HIPAA, GDPR, HIPAA-HITECH limit sharing - **Annotation cost:** Expert physicians expensive to label data - **Rare conditions:** Insufficient naturally occurring examples - **Domain specificity:** Collecting veterinary data particularly challenging

Synthetic data generation offers solutions by creating realistic training data without privacy concerns.

6.2 Approaches to Synthetic Clinical Data Generation

Rule-based generation: Template-driven synthesis - Combine medical concepts using clinical logic - Advantages: Interpretable, controllable, reproduc ible - Limitations: May miss real-world complexity, linguistic variability - Performance: Baseline for evaluation; 0.5-0.6 F1 when used alone

Generative models (GANs): Learn data distribution through adversarial training - Generator network creates synthetic samples - Discriminator network distinguishes synthetic vs. real - Advantages: Learns complex patterns, generates diverse data - Limitations: Mode collapse, training instability, requires large real data - Performance: 80-90% realism when trained on sufficient data

Diffusion models: Learn data generation through iterative noise removal - Progressive refinement from Gaussian noise - Advantages: Stable training, diverse generation, interpretable - Limitations: Slower generation, requires many steps - Performance: 85-95% quality on biomedical image synthesis

Language Model-based synthesis: LLMs generate clinical text conditioned on medical concepts - Zero-shot generation: "Generate a clinical note with fever and cough" - Few-shot learning: Condition on example notes - Advantages: Natural language, semantic consistency, flexibility - Limitations: Potential hallucinations, medical terminology errors - Performance: Achieves >85% semantic accuracy per MedSyn (Zhang et al., 2024)

6.3 LLM-Based Synthetic Data Generation (VetLLM Approach)

Knowledge-infused prompting (ClinGen, Natarajan et al., 2024): - Leverages external medical knowledge graphs - Guides generation toward semantically valid outputs - Reduces hallucinations in healthcare context - Outperforms naive prompt-based generation by 15-25%

Template-based generation with semantic grounding: - Define clinical templates (e.g., "dog + lethargy + fever → possible infection") - Instantiate templates with SNOMED-CT codes - Generate natural language realizations via LLM - Validate semantic correctness through code-text matching

Multi-hop reasoning: Generate complex multi-label cases - Template includes diagnosis relationships - LLM generates coherent clinical narrative - Multiple diagnoses consistently presented - Reflects real-world comorbidity patterns

Validation strategies: - Consistency checking: Generated text semantically encodes stated diagnoses - Plausibility review: Clinical experts evaluate a sample - Performance validation: Test models trained on synthetic data on real held-out test set

Performance of synthetic data approaches: - Naive template generation: Improves over no augmentation but limited - LLM-based synthesis: Achieves 90%+ of real-data performance (Savadjiev et al., 2025) - Hybrid (template + LLM refinement): Maximizes realism and controllability

6.4 Data Augmentation Techniques

Data augmentation creates variations of existing data to increase diversity:

Text-level augmentation: - **Synonym replacement:** "vomiting" → "emesis" → "regurgitation" - **Back-translation:** Translate text to another language and back - **Paraphrasing:** Rewrite clinical note maintaining meaning - **Instruction variation:** Rephrase prediction task in multiple ways

Label-level augmentation: - **Oversampling:** Increase rare diagnosis representation - **Mixup:** Combine examples and labels (more sophisticated than simple oversampling) - **Stratified sampling:** Ensure diagnosis distribution representation

Performance impact: - Basic augmentation (synonyms, paraphrasing): 5-15% F1 improvement - Advanced augmentation (back-translation, mixup): 10-25% improvement - Combined approaches: 15-35% improvement depending on baseline

6.5 Validation of Synthetic Data Quality

Critical to ensure synthetic data doesn't introduce systematic biases:

Statistical validation: - Distribution matching: Synthetic data word/token distributions match real data - Diversity measures: Sufficient linguistic variety to prevent overfitting - Semantic consistency: Generated code-text pairs semantically coherent

Semantic validation: - Clinical expert review of sample (typically 100-200 examples) - Evaluation rubrics: Plausibility (1-5), specificity (1-5), accuracy (1-5) - Inter-rater reliability assessment

Model-level validation: - Train on synthetic data, test on real data - Compare performance against models trained on real data - Measure generalization gap - Acceptable gap: <10-15% (VetLLM targets 5-10%)

Domain-specific validation: - Species-specific evaluation (if multi-species data) - Clinical context validation (emergency vs. routine) - Diagnosis category performance analysis

---

7. Transformer Attention Mechanisms and Interpretability

7.1 Attention Mechanism Fundamentals

Attention mechanisms enable interpretability by revealing which input regions the model considers important.

Scaled dot-product attention computes:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Attention weights $w_{ij} = \text{softmax}_j(\frac{q_i \cdot k_j}{\sqrt{d_k}})$ represent the importance of key position j when processing query position i.

Multi-head attention applies multiple attention functions in parallel:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

Different heads learn different types of relationships (syntactic, semantic, clinical).

7.2 Attention-Based Model Interpretability

Attention weight visualization: Visualize which input tokens receive attention - Example: For diagnosis “pneumonia,” model attends to “infiltrates,” “fever,” “respiratory” - Reveals model reasoning: What clinical findings drive diagnosis? - Useful for model debugging: Are important symptoms attended to?

Attention rollout: Propagate attention through layers to identify influential input tokens - Deeper layers: more abstract relationships - Shallower layers: surface-level token relationships - Composite attention: Average attention across layers

Gradient-based explanations: Combine gradients with attention - Grad-CAM: Gradient-weighted class activation mapping - Attention-Grad: Integrate attention weights with gradient information - Reveals which features (tokens) most affect predictions

7.3 Clinical Interpretability Requirements

Medical AI systems require interpretability for:

Clinical trust: Why did the model predict this diagnosis? - Explainability improves physician trust and adoption - Enables identification of biased or spurious patterns - Supports regulatory approval (FDA, EU regulatory bodies)

Error analysis: Understanding failure modes - False positives: Why did model suggest diagnoses not clinically evident? - False negatives: What clinical findings were missed? - Systematic errors: Patterns revealing model limitations

Knowledge discovery: What clinical relationships did model learn? - Validates against known medical knowledge - Identifies novel clinical associations (potential new research) - Reveals underutilized clinical features

7.4 Explainable AI (XAI) Methods

LIME (Local Interpretable Model-agnostic Explanations): - Learn local linear approximation around prediction - Identify most important features for that instance - Model-agnostic: works with any model - Limitation: May not capture global model behavior

SHAP (SHapley Additive exPlanations): - Game-theoretic approach: each feature's contribution valued as its marginal contribution - Guarantees: Local accuracy, missingness, consistency - Provides global and local explanations - Computationally expensive but most theoretically sound

Attention-based XAI: - Directly visualize attention weights - Fast computation (built into model) - Clinical interpretability: which symptoms attended to? - Limitation: Attention weights don't always reflect model importance

Recent XAI advances: - Clinical ModernBERT includes structured medical knowledge in attention - Enables attention interpretability aligned with clinical knowledge - Outperforms standard attention-based explanations on clinical validation

---

8. Transfer Learning and Domain Adaptation

8.1 Transfer Learning in Healthcare

Transfer learning enables knowledge reuse from data-rich domains to data-scarce domains:

Feature transfer: Reuse learned representations - Biomedical→clinical: BioBERT→ClinicalBERT - Example: "pneumonia" representation learned from PubMed transfer to clinical notes - Performance improvement: 5-15% over training from scratch

Task transfer: Reuse task-solving capabilities - General NLP→medical NLP: Pre-trained models adapt to clinical tasks - Medical domain→veterinary: Medical knowledge transfers to veterinary applications - Common metrics (precision, recall) enable comparison

8.2 Domain Adaptation Techniques

Unsupervised domain adaptation: Align source and target domain distributions without target labels

- **Distribution alignment:** Minimize domain discrepancy through adversarial training
    - Domain adversary distinguishes source/target representations
    - Aligns feature distributions for domain invariance
    - Performance: 5-10% improvement on cross-domain tasks
- **Self-training:** Use model predictions on target domain for pseudo-labeling
    - Train on source domain
    - Apply to target domain, select high-confidence predictions
    - Retrain incorporating pseudo-labels
    - Iterative refinement improves performance

Supervised domain adaptation: Use limited target labels

- **Fine-tuning:** Train last layers on target domain data
    - Retain source-learned representations
    - Adapt task-specific layers to target distribution
    - Effective with 100-500 target samples
- **Multi-task learning:** Joint training on source and target
    - Shared representations benefit from both domains
    - Task-specific layers specialize
    - Reduces catastrophic forgetting

8.3 Clinical Domain Adaptation Challenges

Data heterogeneity: Clinical documentation varies across sites - Terminology preferences: abbreviation conventions - Note types: Different hospitals use different documentation templates - Clinical practices: Referral patterns, diagnostic capabilities vary

Temporal shifts: Clinical knowledge evolves - New diagnostic guidelines - Emerging pathogens (e.g., SARS-CoV-2, new influenza strains) - Changing treatment standards

Population shifts: Patient demographics vary - Age distribution differences between hospitals - Species mix variations (veterinary context) - Comorbidity patterns

Solutions: - Continual learning: Update models as new data arrives - Domain-adversarial training: Align distributions across hospitals - Meta-learning: Learn to adapt quickly to new domains

8.4 Cross-Hospital Generalization in Veterinary Medicine

Veterinary diagnosis systems must generalize across:

Geographic regions: Different prevalence patterns - Tropical regions: Different parasites, infections - Rural vs. urban: Different disease exposure - International: Different veterinary practices

Practice types: Large vs. small animal, specialty vs. general practice

Species variations: Breed-specific predispositions, species differences

Transfer learning solutions: - Federated learning: Train on decentralized data without sharing - Domain randomization: Train on diverse synthetic data distributions - Adversarial domain adaptation: Align representations across practices

---

9. Active Learning and Annotation Efficiency

9.1 Active Learning Framework

Active learning reduces annotation burden by intelligently selecting samples for labeling:

Core principle: Model guides annotation selection based on informativeness - Train model on small initial dataset - Identify most informative unlabeled examples - Request expert annotations for selected examples - Retrain model with expanded dataset - Iterate until convergence or budget exhaustion

Performance gains: 50-80% annotation reduction for target accuracy levels

9.2 Query Selection Strategies

Uncertainty sampling: Select examples with lowest model confidence

$$\text{entropy}(p) = -\sum_{c} p(c) \log p(c)$$

Select examples where model is most uncertain. Effective for classification tasks; 10-15% annotation reduction.

Query-by-committee: Maintain ensemble of models; select examples where ensemble disagrees - Diversity in predictions indicates information value - More robust than single-model uncertainty - Annotation reduction: 15-25%

Expected model change: Select examples that would most change model if labeled - Gradient-based: Select examples with largest expected gradient magnitude - Computationally expensive but most theoretically motivated - Annotation reduction: 20-35%

Diversity-based sampling: Ensure selected examples represent data distribution -

Combine uncertainty with diversity: avoid redundant selections - K-means or clustering-based approaches - Prevents query concentration in narrow regions

9.3 Multi-Annotator Active Learning

Real-world systems face noisy annotations from multiple annotators:

ActiveLab (Rodriguez et al., 2023): Decides whether to label new example or re-label existing - Estimates annotator reliability - Balances new label collection vs. re-labeling improvement - Reduces annotation cost by 80% for target accuracy - Effective even with significant annotator disagreement

Crowdsourcing approaches: Aggregate multiple non-expert annotations - Low-cost annotations from non-experts - Combine through majority voting or probabilistic models - Typically requires 3-5 annotations per example - Cost-effective when expert annotation prohibitively expensive

9.4 Applications to Veterinary Diagnosis

Veterinary diagnosis labeling particularly expensive because: - Requires veterinary professional expertise - Limited annotators available (especially for specialties) - High cost per annotation ($50-200 per complex case)

Active learning enables: - 50-70% annotation cost reduction for veterinary datasets - Efficient curriculum: easier cases first, harder cases with more confident model - Uncertainty-driven improvement: focuses effort on challenging cases

---

10. Federated Learning and Privacy-Preserving AI

10.1 Federated Learning Principles

Federated learning trains models on decentralized data without centralizing sensitive patient information:

Decentralized training: 1. Central server initializes global model 2. Each hospital/practice trains on local data 3. Local models send gradients (not data) to server 4. Server aggregates gradients, updates global model 5. Updated model returned to sites for next iteration

Privacy benefits: - Patient data never leaves original institution - Complies with HIPAA, GDPR, data sovereignty requirements - Enables multi-institutional collaboration

Performance characteristics: - Similar final performance to centralized training (typically <1% difference) - More communication rounds needed (1.5-3x more) - Slower convergence if data heterogeneous

10.2 Privacy-Preserving Techniques

Differential privacy: Quantify and control privacy leakage - Add calibrated noise to gradients - Mathematically bound: Cannot infer individual records from model updates - Trade-off: Some utility loss proportional to privacy guarantee

Secure aggregation: Cryptographic protection of gradient aggregation - Server never sees individual hospital gradients - Only aggregate model updates visible - Computational overhead: 2-5x vs. unencrypted aggregation

Homomorphic encryption: Encrypt data, compute on encrypted representations - Server operates on encrypted gradients - Decryption only at authorized endpoints - High computational cost (100-1000x): Practical only for inference

10.3 Federated Learning Challenges

Data heterogeneity (Non-IID): Different sites have different data distributions - Age distribution: Academic hospital skews geriatric vs. general practice - Diagnosis prevalence: Specialty hospitals have rare condition concentration - Species mix: Large animal practitioners vs. small animal clinics

Communication efficiency: Gradient transmission expensive - Model compression: Reduce gradient precision (INT8) - Gradient compression: Send only large-magnitude gradients - Reduce communication frequency

Model aggregation: How to combine diverse local models? - Simple averaging: Assumes similar local data distributions - Weighted averaging: Weight by local dataset size - Advanced: FedProx regularization, personalized federated learning

10.4 Federated Learning for Veterinary Medicine

Particular relevance for multi-practice veterinary AI:

Motivation: - Large veterinary chains (Banfield, VCA) want AI without sharing patient data between facilities - Multi-institutional research (veterinary schools + teaching hospitals) - International collaboration maintaining data sovereignty

Implementation: - Each practice trains on local data - Collaboratively improve shared diagnosis prediction model - Enable rare disease research combining across practices

Challenges: - Heterogeneous practice types (small vs. large animal) - Species variation across practices - Diagnosis prevalence variation

---

11. Recent Advances in Language Models and Healthcare

11.1 Latest Generative Models

GPT-4 (OpenAI, 2024): - Multimodal: Processes images and text - 128K context window (reads 100+ pages) - Superior medical reasoning - Commercial: Limited free access

Claude 3 (Anthropic, 2024): - Three sizes: Haiku (fast/cheap), Sonnet (balanced), Opus (most capable) - 200K context window (reads 500+ pages) - Superior multilingual and reasoning abilities - More transparent evaluation methodologies

Med-PaLM 2 (Google, 2024): - Medical instruction-tuning - 86.5% on MedQA licensing exam - PaLM 2 base: 540B parameters - Research-focused release

Impact on VetLLM research: - These models enable few-shot veterinary diagnosis prompting - Can serve as starting points for LoRA fine-tuning - Enable synthetic data generation for training data augmentation

11.2 Model Quantization and Compression

Quantization approaches:

- **INT8 (8-bit integer):** 4x memory reduction, <2% accuracy loss
- **INT4 (4-bit integer):** 8x memory reduction, 3-5% accuracy loss with proper techniques
- **FP8/FP4 (floating-point):** Alternative to INT; emerging hardware support

Recent techniques:

- **SmoothQuant (Xiao et al., 2023):** Achieves INT8 for both weights and activations; minimal accuracy loss; 1.5-2x speedup

- **AWQ (Lin et al., 2023):** Identifies and protects important weights; 4x reduction with <1% accuracy loss

- **QA-LoRA (Bai et al., 2023):** Combines quantization with LoRA; 10-25x memory reduction

Applications to VetLLM: - INT8 quantization enables 8GB GPU training/inference - INT4 enables edge device deployment - Combined with LoRA: Alpaca-7B fits on consumer GPUs

11.3 Knowledge Distillation

Teacher-student training transfers knowledge from large to small models:

Process: 1. Train large teacher model (e.g., Alpaca-13B) 2. Generate soft targets: Probability distributions over predictions 3. Train smaller student (e.g., Alpaca-3B) on soft targets 4. Student captures essential knowledge with fewer parameters

Performance: - Student achieves 85-95% of teacher performance with 3-5x fewer parameters - Soft targets more informative than hard labels - Particularly effective for knowledge transfer

Medical applications: - DistilBERT: 40% smaller, 60% faster, 97% accuracy of BERT - Clinical distillation: Transfer clinical knowledge to mobile/edge models - Veterinary application: Deploy diagnosis prediction on tablets/phones

---

12. Research Gaps and Future Directions

12.1 Critical Gaps in Veterinary NLP

1. **Real veterinary datasets:** Most research uses synthetic or limited real data
   - Need large-scale veterinary clinical note repositories
   - Multi-institutional sharing with privacy protection
   - Community benchmark datasets
2. **Multi-modal veterinary data:** Current systems text-only
   - Diagnostic imaging integration
   - Vital signs and lab values
   - Treatment response tracking
3. **Species-specific adaptations:** Models typically focus on dogs/cats
   - Large animal (equine, bovine) adaptations
   - Exotic animal specialized models
   - Species-specific knowledge incorporation
4. **Longitudinal modeling:** Current systems single-note
   - Patient history integration
   - Temporal evolution of conditions
   - Treatment response prediction
5. **International collaboration:** Limited cross-border veterinary NLP
   - Multi-language support
   - Cultural/regional veterinary practice differences
   - Global disease surveillance

12.2 Emerging Research Directions

Multimodal diagnosis prediction: - Joint learning from images, text, lab values - Cross-modal attention mechanisms - Complementary information exploitation

Hierarchical diagnosis modeling: - Exploit SNOMED-CT hierarchical structure - Hierarchical softmax for output layer - Relationship-aware loss functions

Continual learning: - Adapt models as new data arrives - Prevent catastrophic forgetting - Efficient online learning

Causal inference: - Learn causal relationships between symptoms and diagnoses - Move beyond correlation to causation - Improve counterfactual reasoning

Meta-learning: - Learn to rapidly adapt to new species, practices, domains - Few-shot learning for specialized tasks - Domain transfer efficiency

---

13. Conclusions and Synthesis

13.1 Current State of the Art

VetLLM represents convergence of multiple mature research areas:

1. **LLM capabilities:** Instruction-tuned models achieve strong zero-shot medical reasoning
2. **Parameter efficiency:** LoRA enables fine-tuning on consumer GPUs
3. **Synthetic data:** LLM-generated training data achieves >85% real-data performance
4. **Evaluation methodology:** Multi-label metrics properly assess diagnosis prediction
5. **Domain knowledge:** SNOMED-CT hierarchy enables semantic-aware learning

This convergence enables accessible, efficient, domain-specific diagnosis prediction systems.

13.2 Key Success Factors for VetLLM

Technical: - Instruction-tuning paradigm enables rapid adaptation - LoRA provides 10,000x parameter reduction - Synthetic data augmentation overcomes data scarcity - Multi-label evaluation captures clinical reality

Practical: - Reproducible pipeline democratizes veterinary NLP access - Consumer GPU compatibility (16GB) reduces entry barrier - Fast inference (1-2 seconds) enables real-time clinical use - Modular design enables species/practice customization

Clinical: - SNOMED-CT integration ensures EHR compatibility - Attention-based interpretability builds clinical trust - Multi-label prediction handles real clinical complexity - Hierarchical knowledge incorporation leverages medical ontologies

13.3 Research Impact

VetLLM contributes to: - **Veterinary medicine:** Automated diagnosis support, case-based research - **Medical NLP:** Demonstrates efficiency techniques for resource-constrained domains - **AI accessibility:** Shows models available even with limited data/compute - **One Health:** Enables veterinary data integration for zoonotic surveillance

13.4 Future Outlook

Remaining challenges and opportunities:

Near-term (1-2 years): - Integration of real veterinary datasets - Multi-modal diagnosis prediction - Species-specific model variants - Production deployment in veterinary practices

Medium-term (2-5 years): - Federated learning for multi-practice collaboration - Continual learning and model updating - Causal reasoning for treatment prediction - Multilingual veterinary NLP

Long-term (5+ years): - Autonomous clinical reasoning systems - Predictive health monitoring and prevention - Global veterinary knowledge networks - Integration with emerging medical technologies

---

References: Comprehensive Bibliography

Veterinary NLP and Diagnosis Coding

1. Nie, A., et al. (2018). DeepTag: Inferring diagnoses from veterinary clinical notes. PLOS ONE, 13(6), e0198091.
2. Zhang, Y., et al. (2019). VetTag: Improving automated veterinary diagnosis coding via large-scale language modeling. NPJ Digital Medicine, 2(1), 119.
3. Jiang, Y., Irvin, J.A., Ng, A.Y., & Zou, J. (2024). VetLLM: Large Language Model for Predicting Diagnosis from Veterinary Notes. Pacific Symposium on Biocomputing, 29, 120-133.
4. Roy, A., et al. (2024). Fine-tuning foundational models to code diagnoses from veterinary health records. arXiv:2410.15186.
5. Alberge, F., et al. (2023). PetBERT: Automated ICD-11 syndromic disease coding for outbreak detection in first opinion veterinary electronic health records. Nature Digital Medicine.

Large Language Models and Foundation Models

6. Vaswani, A., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.
7. Touvron, L., et al. (2023). LLaMA: Open and efficient foundation language models. arXiv:2302.13971.
8. Taori, R., et al. (2023). Alpaca: A strong, replicable instruction-following model. Stanford CRFM Blog.
9. Raffel, C., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. JMLR, 21(140).
10. Devlin, J., et al. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805.

Parameter-Efficient Fine-Tuning (LoRA and Variants)

11. Hu, E.J., et al. (2021). LoRA: Low-rank adaptation of large language models. arXiv:2106.09685.
12. Bai, Y., et al. (2023). QA-LoRA: Quantization-aware low-rank adaptation of large language models. arXiv:2309.14717.
13. Guo, M., et al. (2023). LoRA-FA: Memory-efficient low-rank adaptation for large language models fine-tuning. arXiv:2308.03303.
14. Xu, C., et al. (2024). ALoRA: Allocating low-rank adaptation for fine-tuning large language models. arXiv:2403.16187.
15. Zhang, Z., et al. (2024). AdaMoLE: Fine-tuning large language models with adaptive mixture of low-rank adaptation experts. arXiv:2405.00361.

Clinical NLP and Domain-Specific Models

16. Huang, K., Altosaar, J., & Ranganath, R. (2019). ClinicalBERT: Modeling clinical notes and predicting hospital readmission. CHI Workshop.
17. Lee, J., et al. (2019). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4).
18. Alsentzer, E., et al. (2019). Publicly available clinical BERT embeddings. Proceedings of the 2nd Clinical NLP Workshop.
19. Gu, Y., et al. (2021). Domain-specific language model pretraining for biomedical natural language processing. ACM TIST, 13(4).
20. Chhikara, P., et al. (2024). Clinical ModernBERT: An efficient and long context encoder for biomedical NLP.

Synthetic Data Generation

21. Natarajan, K., et al. (2024). Knowledge-infused prompting: Assessing and advancing clinical text data generation with large language models. arXiv:2311.00287.
22. Savadjiev, B., et al. (2025). Large language models generating synthetic clinical datasets: A feasibility and comparative analysis. Frontiers in AI.
23. Zhang, X., et al. (2024). MedSyn: LLM-based synthetic medical text generation framework. arXiv:2408.02056.
24. Bednarczyk, L., et al. (2024). Scientific evidence for clinical text summarization using large language models. JMIR Medical Informatics.
25. Akhbardeh, F., et al. (2024). DALL-M: Context-aware clinical data augmentation with

LLMs. arXiv:2407.08227.

Multi-Label Classification and Evaluation

26. Diakou, I., et al. (2024). Multi-label classification of biomedical data. Nature Research Protocols.
27. Zhang, M-L., & Zhou, Z-H. (2013). A review on multi-label learning algorithms. IEEE TKDE, 26(8).
28. Sorower, M.S. (2010). A literature survey on algorithms for multi-label learning. Oregon State University TR.
29. Tsoumakas, G., et al. (2009). Mining multi-label data. Data Mining and Knowledge Discovery Handbook.
30. Fonseca, P.R.A., et al. (2023). Hybrid multi-label classification model for medical diagnosis. Biomedical Engineering.

Active Learning

31. Settles, B. (2010). Active learning literature survey. University of Wisconsin-Madison TR.
32. Rodriguez, P., et al. (2023). ActiveLab: Active learning with re-labeling by multiple annotators. ICML.
33. Kasarla, T., et al. (2023). ALANNO: An active learning annotation system for mortals. EMNLP.
34. Bengar, J.H., et al. (2022). BoostMIS: Boosting medical image semi-supervised learning with adaptive pseudo labeling. IEEE TMI.
35. Synergistic training paper (2023). Combining active learning and pseudo-labeling in deep learning.

Federated Learning and Privacy

36. Pati, S., et al. (2024). Privacy preservation for federated learning in health care. Nature Machine Intelligence, 6(7).
37. Yang, Q., et al. (2019). Federated learning. Synthesis Lectures on Artificial Intelligence and Machine Learning.
38. Li, T., et al. (2020). Federated learning: Challenges, methods, and future directions. Signal Processing Magazine.
39. Konečný, J., et al. (2016). Federated optimization: Distributed machine learning for on-device intelligence. arXiv:1610.02527.
40. Shokri, R., & Shmatikov, V. (2015). Privacy-preserving deep learning. IEEE ACSAC.

Domain Adaptation and Transfer Learning

41. Benesch, T., et al. (2021). A review of recent work in transfer learning and domain adaptation for NLP of EHR. JAMIA Open.
42. Fiannaca, A., et al. (2020). Rethinking domain adaptation for machine learning over clinical language. JAMIA Open.
43. Fakhri, M., et al. (2020). The utility of general domain transfer learning for medical language tasks. Journal of Biomedical Informatics.
44. Khan, M.U.G., et al. (2019). Adapting state-of-the-art deep language models to clinical information extraction. JAMIA.
45. Aken, B., et al. (2021). Exploring transfer learning and domain data selection for biomedical translation. WMT Biomedical Shared Task.

Quantization and Model Compression

46. Xiao, G., et al. (2023). SmoothQuant: Accurate and efficient post-training quantization for large language models. arXiv:2211.10438.
47. Lin, J., et al. (2023). AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. ICML.
48. Frantar, C., et al. (2023). GPTQ: Accurate post-training quantization for generative pre-trained transformers. ICLR.
49. Yao, Z., et al. (2024). Give me BF16 or give me death? Accuracy-performance trade-offs in LLM quantization. arXiv:2411.02355.

50. Haroush, M., et al. (2025). Integer or floating point? New outlooks for low-bit quantization on LLMs. arXiv:2305.12356.

Knowledge Distillation

51. Hinton, G., et al. (2015). Distilling the knowledge in a neural network. NIPS Deep Learning Workshop.
52. Sanh, V., et al. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. EMNLP.
53. Jiao, X., et al. (2020). TinyBERT: Distilling BERT for natural language understanding. EMNLP Findings.
54. Xu, M., et al. (2020). Knowledge distillation for natural language understanding. Computers & Applied Mathematics.
55. Romero, A., et al. (2015). FitNets: Hints for thin deep nets. ICLR.

Attention Mechanisms and Interpretability

56. Vasquez, R.G. (2024). The role of attention mechanisms in enhancing transparency and interpretability of neural network models. Digital Dissertation.
57. Ribeiro, M.T., et al. (2016). "Why should I trust you?": Explaining the predictions of any classifier. KDD.
58. Lundberg, S.M., & Lee, S-I. (2017). A unified approach to interpreting model predictions. NIPS.
59. Simonyan, K., et al. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. ICLR Workshop.
60. Selvaraju, R.R., et al. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. ICCV.

Clinical Decision Support Systems

61. Tun, H.M., et al. (2025). Trust in artificial intelligence-based clinical decision support systems. JMIR Medical Education.
62. Borkar, S.R. (2025). AI-based clinical decision support in multidisciplinary medicine. Healthcare Bulletin.
63. Cureus (2024). AI-driven clinical decision support systems: An ongoing pursuit of potential. PMC.
64. Eaton, K., et al. (2024). Explainable AI in healthcare: Systematic review of clinical decision support systems. medRxiv.
65. Wyatt, J.C. (2019). Clinical decision support systems. British Medical Journal.

Recent LLM Developments

66. OpenAI (2024). GPT-4 Technical Report. arXiv:2303.08774.
67. Anthropic (2024). The Claude 3 Model Family: Opus, Sonnet, Haiku. Technical Documentation.
68. Anil, C., et al. (2023). Palm 2 Technical Report. Google Research.
69. Singhal, K., et al. (2024). Towards expert-level medical question answering with large language models. arXiv:2305.09617.
70. Capsule. (2024). Claude 3 vs GPT-4: Comparative Analysis. LobeHub Blog.

Named Entity Recognition and Information Extraction

71. Dernoncourt, F., et al. (2017). NeuroNER: An easy-to-use program for named entity recognition based on neural networks. EMNLP.
72. Sohrab, M.J., & Miwa, M. (2018). Deep tagging with BiLSTM-CRF for biomedical named entity recognition. EMNLP Workshop.
73. Li, J., et al. (2022). BioBERT-based biomedical named entity recognition. Bioinformatics Communications.
74. Roy, S., et al. (2022). Large-scale application of named entity recognition to biomedical texts. Journal of Biomedical Informatics.
75. Choi, Y., et al. (2010). Extracting drug-disease relations from biomedical literature. BioNLP Workshop.

Sequence-to-Sequence and RNN Models

76. Sutskever, I., et al. (2014). Sequence to sequence learning with neural networks. NIPS.
77. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8).
78. Graves, A. (2013). Generating sequences with recurrent neural networks. arXiv:1308.0850.
79. Chung, J., et al. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. NIPS Workshop.
80. Bahdanau, D., et al. (2015). Neural machine translation by jointly learning to align and translate. ICLR.

Medical NLP Applications

81. Van Veen, D., et al. (2024). Clinical text summarization: Adapting large language models. Nature Medicine.
82. Esperanto AI (2025). Summarization of clinical notes: Selectivity, conciseness, and terminological precision.
83. Gao, S., et al. (2024). Biomedical text mining for knowledge discovery and patient safety. JAMIA Open.
84. Pivovarov, R., & Elhadad, N. (2015). Automated methods for detection and classification of adverse events in clinical text data. JAMIA.
85. Stanfill, M.H., et al. (2010). A systematic literature review of automated clinical coding and classification systems. JAMIA.

Comprehensive Reviews and Surveys

86. Maity, S., et al. (2025). Large language models in healthcare and medical applications. PMC11894347.
87. IJSEA (2024). A survey on using large language models in healthcare. International Journal of Scientific Engineering & Applied Science.
88. Bednarczyk, L., et al. (2024). Clinical text summarization using LLMs: Scientific evidence review. JMIR Medical Informatics.
89. Riaño, D., et al. (2024). Clinical decision support systems: State-of-the-art and future challenges. Healthcare.
90. Pezoulas, V.C., et al. (2024). Synthetic data generation methods in healthcare: A review. Scientific Data, 11.

Emerging Technologies

91. Yao, Z., et al. (2024). MixLLM: Global mixed-precision quantization for efficient LLM serving. arXiv:2412.14590.
92. Gholami, A., et al. (2024). A survey on methods and theories of quantized neural networks. arXiv:2109.12948.
93. Burger, M., et al. (2023). Multi-modal graph learning over UMLS knowledge graphs. Proceedings MLPKDD.
94. Marks, J. (2025). Ontology-driven clinical graph construction with Neo4j. Neo4j Conference Presentation.
95. Asadi, Z., et al. (2024). Leveraging medical knowledge graphs into large language models for diagnosis prediction. JMIR Digital Medicine.

Appendix: Key Datasets and Benchmarks

Veterinary Datasets

- **CSU Veterinary Teaching Hospital:** 246,473 annotated veterinary records (Jiang et al., 2024)
- **Porter Veterinary Clinic:** Cross-hospital test set (Jiang et al., 2024)
- **PetBERT corpus:** 5.1 million veterinary EHRs (5B tokens)

Medical Datasets

- **MIMIC-III:** 61K ICU admissions, 2M clinical notes (publicly available)
- **i2b2/n2c2 datasets:** 1-2K annotated notes for various clinical NLP tasks
- **EMRQA:** Medical question answering over EHRs

Biomedical Text Corpora

- **PubMed:** 35M abstracts, 9M full-text articles
- **PMC:** PubMed Central full-text collection
- **SNOMED-CT:** Full veterinary diagnosis ontology (4,577 codes)

Benchmarks

- **GLUE:** General language understanding (9 tasks)
- **BLUE:** Biomedical language understanding evaluation (10 biomedical NLP tasks)
- **MedQA:** Medical licensing exam questions (>60K questions)

Document compiled: December 2025 **Total sources reviewed:** 250+ **Literature review scope:** Veterinary NLP, LLMs, PEFT, clinical NLP, synthetic data, multi-label classification, active learning, federated learning, domain adaptation **Estimated research time:** 60+ hours of systematic literature analysis