



Thesis Proposals in Artificial Intelligence for Cutaneous Lymphoma Research

Overview

These nine topics connect applied artificial intelligence with research on cutaneous lymphoma and its treatments. They cover clinical images, language models, published research and approved medicine information. Each combines a **theoretical component** with a **practical component**, and every dissertation can be completed using public data and documents without access to hospital records.

The work is supervised from an AI and computing perspective, in collaboration with researchers from medicine who help define clinically meaningful questions and interpret the findings. Clinical judgements are reviewed by appropriate medical or pharmacy experts. Students do not work with identifiable patient records.

Topics 1-8 include a Python implementation, reproducible experiments and a comparison between a baseline and a tested method. Some involve classifier training or limited fine tuning; others focus on retrieval, verification or evaluation algorithms. Topic 9 is a literature-based review of AI research and requires no programming or model training. Each topic is self-contained, with its sources, resources, expert input and final scope agreed before the main work begins.

To choose a topic, consider whether you prefer working with images, analysing research papers or evaluating generated text. The descriptions below explain the research question, theoretical background, practical tasks, expected project outputs and required skills. Optional extensions can be discussed once the core project is complete.

Links and Resources

Open Science Framework	https://osf.io/
Reporting guidelines	EQUATOR Network ; PRISMA ; TRIPOD
Risk of bias	PROBAST and current AI-specific guidance
Public image data	Fitzpatrick17k , DDI and SkinCon, under their respective terms
Medicine information	European Medicines Agency: approved product information
Trials and publications	ClinicalTrials.gov ; PubMed Central , checking article reuse conditions
Public evaluation resources	NLI4CT ; RoBBR , as reference tasks rather than CTCL-specific data



Thesis Topics

Topics 1 and 8 include bounded neural-network fine tuning; topics 2 and 3 train lightweight classifiers. Topics 4-7 develop and evaluate software pipelines without requiring language-model training. Topic 9 focuses on reading, organising and critically comparing the research literature, without coding. Resources and workload are agreed during the initial feasibility check.

Each topic is offered to one student only. Topics marked ALLOCATED have been assigned and are no longer available.

1. Reliable Deep Learning for Recognising Mycosis Fungoides Among Its Inflammatory Mimics

ALLOCATED

This topic evaluates whether image models distinguish mycosis fungoides from selected inflammatory skin diseases when the image source changes. The student implements and compares frozen-feature classifiers with limited fine tuning of a pretrained encoder. The central question is whether adaptation improves independent-source reliability, and how its gains depend on the available training data.

Theoretical component

The student studies transfer learning, frozen visual representations, dataset shift, calibration and selective prediction, where a model can defer uncertain cases. The study covers leakage through duplicates, source artefacts and related images, and the uncertainty of estimates from small rare-disease datasets.

Practical component

The student implements a Python data pipeline with duplicate groups and source-aware splits. A frozen encoder with a linear classifier provides the baseline. A bounded PyTorch experiment then trains the classification head and fine tunes the final block of one pretrained encoder. Encoder choice, preprocessing and the evaluation protocol remain matched. Learning curves, repeated seeds and frozen-versus-adapted comparisons test whether additional training helps on an independent source.

Training, preprocessing and model selection are separated from the final test source. Duplicate groups stay together, and source-disjoint evaluation is used where the available classes support it. Calibration and any abstention threshold are selected on development data. Evaluation includes sensitivity, specificity, uncertainty intervals, calibration and the proportion of cases on which the model answers. Error analysis examines disease groups, source and, where sufficiently supported, skin tone.

If patient identifiers are unavailable, the student reports this limitation and does not describe deduplication as patient-level separation. Images without verified disease stage cannot establish early-stage diagnostic performance. Saliency maps support exploratory analysis rather than prove that a model reasons clinically.

Main outcomes and further work

A versioned Python/PyTorch implementation, trained classifier and encoder-adaptation checkpoints, reproducible experiment configurations, learning curves and a calibrated error analysis. An optional extension compares an additional encoder or abstention method.

Required skills: good Python, PyTorch and machine learning evaluation. GPU or cloud access is needed for the bounded fine-tuning experiment. No model is pretrained from scratch.



2. Vision-Language Models for Mycosis Fungoides: Diagnostic Evidence, Repeatability and Uncertainty

This topic tests existing vision-language models on public photographs of mycosis fungoides and its inflammatory mimics. It is a bounded evaluation project: a small number of models, a fixed task and predefined prompts. Its contribution lies in a carefully controlled disease-specific evaluation, not in training a new large model.

Theoretical component

The student studies image-text representations, zero-shot classification, generated differential diagnoses, prompt sensitivity and uncertainty. The review examines prior dermatology benchmarks, possible exposure of models to public test images, and the difference between a plausible explanation and evidence actually visible in a photograph.

Practical component

The student builds a Python evaluation harness for two accessible VLMs, with batched inference, fixed prompt templates, output parsing and versioned results. A linear classifier trained on frozen image embeddings provides the baseline. The protocol fixes the image manifest, source groups and primary task before testing. A small prompt ablation and repeated-run subset measure which changes affect results; a top-three differential diagnosis is optional.

The evaluation records sensitivity and specificity for mycosis fungoides, repeatability on a predefined subset, prompt-related changes, refusal or abstention frequency, and inference cost. If probabilities are unavailable, the study does not treat a model's verbal confidence as a calibrated probability. Performance on answered cases is reported alongside the fraction of images answered.

Medical collaborators review a predefined sample of explanations without seeing model identity and assess whether the described features are visible. Differences by image source and comparison disease are examined. Skin-tone subgroup estimates are included only when supported by enough cases, with uncertainty intervals and limitations.

Public-atlas contamination cannot normally be excluded. The student records known training-data overlap, tests a less familiar source if suitable data exist and describes the remaining uncertainty. The final test set is not used to select prompts, models or demonstrations.

Main outcomes and further work

A reusable Python evaluation harness, a trained lightweight baseline, prompt-comparison and repeatability results, an error taxonomy and an expert-reviewed explanation subset. The code records invalid outputs, failures and cost. A few-shot extension uses development examples only.

Required skills: good Python, model libraries or APIs, and statistics. Only the lightweight classifier is trained; VLM fine tuning is not required. Model access costs and permitted image uses are checked first.



3. An AI-Assisted Audit of Public Images for Cutaneous Lymphoma Research

This topic creates a reliable, documented image manifest for evaluating AI models on cutaneous lymphoma photographs. It uses pretrained visual embeddings and duplicate-detection methods to investigate image overlap, label consistency, source composition and representation. It suits a student who prefers data analysis and reproducible engineering to neural-network training.

Theoretical component

The student studies dataset documentation, perceptual hashing, embedding similarity, label noise and evaluation leakage. The review covers published audits and cleaned versions of dermatology datasets so that the dissertation adds a disease-specific contribution rather than repeating a general cleaning exercise. Skin-tone labels are considered with their measurement limitations.

Practical component

The student inventories publicly accessible collections with relevant lymphoma and comparison images. For each source, the inventory records terms of access and reuse, available identifiers, diagnosis labels, provenance and other annotations. Availability of metadata is distinguished from permission or ability to download and reuse the photographs.

The student implements a Python audit pipeline with perceptual hashes, pretrained embeddings and similarity-based grouping. Hash-only and embedding-assisted duplicate detection are compared against manually reviewed image pairs, with thresholds fixed on development pairs and images kept separate across splits. The experiment reports duplicate-detection precision and recall, runtime and the effect on source and diagnosis distributions.

Medical collaborators independently review a bounded subset for diagnostic plausibility, photograph quality and relevant visible characteristics. Agreement is reported and disagreements remain traceable. Labels are not described as histologically verified unless the source supplies that evidence. Patient identities or early-disease status are not inferred from visual similarity alone.

The student trains a small classifier on frozen features and automates comparisons of ordinary, duplicate-group and, where feasible, source-disjoint splits. Repeated seeds and uncertainty intervals quantify the effect of each audit step. Removing one cleaning component at a time tests which steps change the measured performance.

Main outcomes and further work

A Python audit tool, reviewed duplicate groups, a data card, a trained lightweight benchmark and a reproducible cleaning-ablation report. Identifiers and annotations can be shared under appropriate terms; images are redistributed only when permitted.

Required skills: good Python, data handling, image embeddings and statistics. A lightweight classifier is trained; encoder fine tuning is not required.

4. Evidence-Grounded AI Information on Cutaneous Lymphoma Treatments in English and Greek

This topic measures how accurately language models answer treatment-information questions in English and Greek, and whether providing approved evidence reduces errors and omissions. It is an evaluation of information fidelity, not a system for recommending a treatment to an individual patient. The initial study uses a bounded set of medicines and factual questions.

Theoretical component

The student studies retrieval-augmented generation, cross-language retrieval, factual support, answer completeness and abstention when evidence is insufficient. The literature review distinguishes this work from existing drug-information and medical question-answering benchmarks. It identifies which combination of language, evidence handling and error analysis adds a defensible contribution.

Practical component

Public EMA product information and approved patient leaflets provide dated English and Greek source documents for a clinically agreed treatment scope. Disease-information sources may support a separately defined question subset. Each question has an independently checked reference answer, required facts, supporting passages and an explicit indication when the supplied sources cannot answer it.

The student implements document parsing, chunking, indexing and answer generation in Python. A BM25 retrieval baseline is compared with dense or hybrid retrieval using matched documents and model settings. No-document and correct-passage conditions distinguish retrieval failures from generation failures. Retrieval recall and answer fidelity are evaluated separately. A focused chunking or reranking ablation uses development data only. Translations are checked by a bilingual reviewer.

Primary outcomes are factual accuracy, omission of predefined critical facts, support by the cited passage and appropriate abstention. Medical or pharmacy reviewers assess clinical significance without seeing model identity. Agreement between reviewers and the handling of disagreements are reported. Readability is secondary and is not treated as evidence of correctness.

Questions based on the same source fact or medicine are grouped during analysis. Repeated model answers do not count as independent questions. If models are adapted or prompts tuned, evaluation questions remain separate. Model versions, document dates, settings, cost and repeated-run variability are recorded.

Main outcomes and further work

A bilingual question-and-evidence resource, a runnable retrieval-augmented generation pipeline, controlled retrieval and answer comparisons, and an error analysis. Experiments can be rerun from saved configurations and document manifests. An optional extension adapts a reranker if enough independent relevance annotations exist.

Required skills: good Python, text retrieval, basic NLP and statistics, and fluent Greek and English. LLM training is not required. Medical or pharmacy review is needed for clinical appraisal.



5. Can Language Models Help Appraise the Methodological Quality of AI Studies in Cutaneous Lymphoma? **ALLOCATED**

This topic evaluates language models as assistants for finding and checking methodological evidence in published prediction-model studies. It combines a focused literature map with an empirical AI benchmark. The project focuses on how accurately models identify reported methods and support their assessments with evidence from the paper.

Theoretical component

The student studies internal and external validation, overfitting, patient-level splitting, calibration and data leakage. The review covers the current appropriate versions of PROBAST and TRIPOD, the distinction between risk of bias and completeness of reporting, and existing AI appraisal benchmarks such as RoBBR. A missing report is not automatically evidence that a method was not performed.

Practical component

The student writes a protocol and searches for eligible CTCL prediction-model studies. A short literature map supplies context. The number and accessibility of suitable papers are checked early. If necessary, the supervisor and student agree a clearly defined related dermatology prediction task before model experiments begin.

Two independent reviewers create a reference set of explicit methodological facts and selected appraisal judgements, with source passages and documented adjudication. Examples include the reported split unit, use of external validation, calibration assessment, and availability of code. The supervisor arranges appropriate expertise for risk-of-bias judgements.

The student implements a Python pipeline for full-text parsing, section-aware retrieval, structured appraisal fields and source-span validation. Direct full-text prompting provides the baseline. Removing section-aware retrieval or the evidence-checking stage tests each component's contribution. Outputs distinguish reported evidence, insufficient information and uncertainty. An LLM-generated judgement is never used as its own reference standard.

Results include agreement with adjudicated reviewers, evidence support, false claims about unreported methods, and error types. A predefined human-only versus AI-assisted subset can measure review time and correction burden if staffing permits. Papers used for prompt development remain separate from final evaluation.

Main outcomes and further work

A runnable appraisal-assistance pipeline, an expert-annotated benchmark, component-ablation results and a source-linked error report. The code exports structured annotations and calculates agreement and evidence-support metrics. Human correction effort is measured on a bounded subset if reviewers are available.

Required skills: good Python, text processing, machine learning evaluation and academic English. Model training is not required. Expert-reference annotations remain the main practical dependency.

6. AI-Assisted Extraction of Treatment Outcomes from Public Cutaneous Lymphoma Studies

This topic evaluates whether language models can extract treatment results with the correct outcome definition, population, time point and supporting evidence. It addresses an evidence-synthesis task relevant to pharmaceutical research, using public trial records and reusable publications. It does not estimate treatment effects from individual patient data.

Theoretical component

The student studies information extraction, structured generation, evidence retrieval and the interpretation of trial outcomes. Medical collaborators explain the distinction between skin and global response, relevant analysis populations, confirmation of response and duration. The student reviews existing numerical extraction tools and evidence-synthesis systems, including TrialMind, to define the remaining task-specific question.

Practical component

A feasibility inventory identifies CTCL studies with accessible result records or full-text reports that permit the intended use. The initial task uses one or two well-defined outcome families and a manageable set of fields. Expert-checked reference entries link each value to the treatment arm, population, outcome definition, denominator, time point and exact source passage. An unreported value is left unknown.

The student implements ingestion of public trial records and accessible article text, evidence retrieval, schema-constrained extraction and Python validation rules. Direct prompting provides the baseline. Comparisons remove retrieval or validation to measure their effects. Checks flag impossible counts, inconsistent percentages and missing population or time-point fields; flagged records are reviewed or left unresolved rather than silently corrected. Different clinical endpoints are not treated as equivalent.

Accuracy is evaluated both by field and by complete outcome record. Error analysis distinguishes a wrong value, wrong arm, wrong denominator, wrong time point and unsupported extraction. The system can abstain instead of filling a missing field. Review effort and computation cost are recorded where directly measurable.

All records and publications from the same trial stay together across development and test sets. Trial registration identifiers help prevent overlap with secondary reports. The number of suitable CTCL results is checked in a pilot so that the supervisor and student can agree a feasible scope before the main data collection.

Main outcomes and further work

An annotated outcome resource, a runnable extraction-and-validation tool, component-ablation results and a numerical error report. The implementation includes checks on representative valid and invalid records and exports evidence-linked results. Any time saving requires a timed comparison.

Required skills: good Python, structured data handling, document processing, NLP and statistics. LLM training is not required; domain-expert reference review is needed.

7. Faithful AI Simplification of Approved Treatment Information in English and Greek

This topic asks whether language models can make approved medicine information easier to read while preserving its critical meaning. The task is rewriting a supplied passage rather than answering a question. It studies factual omissions and changes to numbers, conditions and warnings, alongside readability.

Theoretical component

The student studies controlled text generation, plain-language rewriting, factual consistency and cross-language evaluation. The review examines existing package-leaflet simplification work and readability tools, including the limited validity of directly applying English formulas to Greek. Lower reading difficulty is considered together with retention of required information.

Practical component

The student selects matched English and Greek passages from public EMA product information for the agreed treatment scope. Suitable passages contain clearly identifiable facts, such as an administration condition, a monitoring requirement or a warning. A medical or pharmacy reviewer checks a structured checklist of facts that every acceptable rewrite must preserve. The source version and date are recorded.

The student builds a Python rewriting-and-verification pipeline. Plain-language prompting provides the baseline; checklist-guided generation and a verify-then-revise method are compared with it. The verifier checks numbers, units and medicine names, and uses a pretrained entailment model to flag possible semantic changes. Its flags are evaluated against human annotations. Development passages guide method choices; final passages remain held out by document or medicine.

Independent reviewers, blinded to model and strategy, assess preservation of facts, omissions, changed negations and conditions, unsupported additions and clarity. Automatic checks record changes to numbers, units and medicine names, but do not replace judgement about the meaning of a warning. A predefined subset is repeated to assess output stability.

The analysis compares readability and factual retention jointly. It reports whether a method achieves clearer text at the cost of lost meaning, with uncertainty intervals clustered by source document. Patient comprehension or clinical safety is not inferred from readability scores; those would require a separate appropriately designed user study.

Main outcomes and further work

A bilingual passage-and-fact resource, a runnable rewriting-and-verification tool, a controlled method comparison and a reusable error taxonomy. Removing the verifier measures its effect on factual retention, false alarms and inference cost. Human evaluation remains the reference for meaning and clarity.

Required skills: good Python, text processing, basic NLP and statistics, and fluent Greek and English. Model training is not required. Medical or pharmacy review supports the reference annotations.

8. Verifying Numerical Treatment Claims with Language Models and Executable Checks

This topic investigates how AI systems can verify a treatment claim against supplied study evidence. The student checks whether a statement is supported, contradicted or left unresolved by the source. The emphasis is on numbers, populations, time points and appropriate uncertainty.

Theoretical component

The student studies natural language inference, numerical reasoning, evidence attribution and selective prediction. The review includes NLI4CT and its perturbed clinical-trial tasks, which already evaluate logical and numerical consistency. The dissertation must identify an additional contribution, such as a tested verification method or a well-defined CTCL extension, rather than only rerunning an established benchmark on newer models.

Practical component

The initial experiments use a public annotated clinical-trial inference dataset with existing reference labels. This is a general reference task, not a CTCL cohort. A separate, expert-checked CTCL set is built from public treatment reports. Claims and annotations are created independently of the model under evaluation.

The student fine tunes one compact pretrained inference encoder on public training data, using trial-grouped splits and preserving the benchmark's original labels. The trained encoder is compared with a general LLM and a hybrid verifier combining evidence retrieval with Python arithmetic and unit checks. A separate CTCL evaluation includes insufficient-evidence cases through a development-tuned abstention gate; missing evidence is not relabelled as contradiction.

Controlled test cases alter a denominator, outcome definition, treatment arm or time point. Their reference labels are manually checked; automatically changing a sentence does not by itself establish the correct label. Claims derived from the same trial, source passage or perturbation family stay in the same split.

Evaluation includes error rates by reasoning type, evidence correctness, abstention coverage and the cost per verified claim. Any threshold for confidence or deferral is selected on development data. A final independent evaluation checks whether gains persist beyond the examples used to design the rules.

Main outcomes and further work

A compact fine-tuned model, a runnable hybrid verifier, a reviewed CTCL extension and numerical error analysis. Removing the arithmetic checks or abstention gate tests their individual effects. Agreement with arithmetic does not establish clinical comparability between trials.

Required skills: good Python, PyTorch or a transformer training library, NLP and numerical checks. A modest GPU or cloud budget supports compact-encoder fine tuning. No large model is pretrained from scratch.



9. Artificial Intelligence in Cutaneous Lymphoma: A Structured Scoping Review

This is a literature-based dissertation for a student who prefers reading research, comparing methods and building a clear account of a field. It investigates what AI has been used for in cutaneous lymphoma, which data and methods have been studied, what the results establish and which questions remain unresolved. No programming, model training or new clinical dataset is required.

Theoretical component

The student studies the distinction between narrative, systematic and scoping reviews, and the principles of reproducible literature searching and evidence synthesis. The AI background covers classification, transfer learning, foundation models, validation, data leakage and performance metrics. The [PRISMA-ScR reporting checklist](#) guides the description of the review.

Practical component

The student first reads existing reviews and agrees a focused question with the supervisor. A short pilot checks the volume and accessibility of suitable studies and helps avoid duplicating an existing review. The scope may focus on a particular data modality, AI approach or evaluation issue where a clear question can be answered from the available literature.

A written protocol defines the sources to search, search terms, dates and inclusion criteria. The student records the search, removes duplicate references, screens titles and abstracts, and reads the eligible full texts. A second reviewer checks selection and extraction as agreed in the protocol, with disagreements documented and resolved.

For each study, the student records the task, data source and size, AI method, comparator, validation design, reported metrics and stated limitations. The synthesis compares findings, identifies contradictory results and distinguishes evidence demonstrated by the studies from claims that require further testing. Study quality and missing information are discussed using criteria appropriate to the question.

Reference-management software, spreadsheets and word-processing tools are sufficient for the work. The student prepares a study-selection flow diagram, structured comparison tables and a critical narrative synthesis. The review reports its search and access limitations. A statistical meta-analysis, expert survey and software implementation are not required.

Main outcomes and further work

A documented search and selection process, an organised evidence table, a critical comparison of AI approaches and a reasoned account of the field's limitations and open questions. An optional extension examines one narrower methodological issue in more detail, if the available studies support it.

Required skills: good academic English, an understanding of basic AI concepts, careful reading, organised work and spreadsheet use. No Python, GPU or model-training experience is required.