title	year	Medical domain	Task	Technique	Position	Intrinsic/ Extrinsic	Motivation	Submotivation
Detecting hallucinations in large language models using semantic entropy.	2024	General Medicine	QA	consistency sampling + semantic entropy threshold	after	extrinsic	uncertainty	low confidence prediction
Identifying Deprescribing Opportunities With Large Language Models in Older Adults: Retrospective Cohort Study.	2025	Geriatrics	clinical decision support	prompting (Verbalized CoT Confidence) + confidence threshold; consistency sampling	before, after	extrinsic	uncertainty	low confidence prediction
Performance of Large Language Models in Numerical Versus Semantic Medical Knowledge: Cross-Sectional Benchmarking Study on Evidence-Based Questions and Answers.	2025	General Medicine	QA	prompting (idk option included in the question)	before	extrinsic	uncertainty	low confidence prediction
Evaluating artificial intelligence bias in nephrology: the role of diversity, equity, and inclusion in AI-driven decision-making and ethical regulation.	2025	Nephrology	QA	- (just evaluates)	-	-	safety	ethical, bias
A Pre-Processing Framework for Securing LLM-RAG Interfaces Against Information Leakage - Practice and Experience in Advanced Research Computing 2025: The Power of Collaboration	2025	Medical Safety	QA	screening model (small, lightweight LLM, prompted to reject off domain inputs)	before	extrinsic	safety	Harmful Content
CARES: Comprehensive Evaluation of Safety and Adversarial Robustness in Medical LLMs	2025	Medical Safety	Alignment Evaluation	a classifier (Qwen2.5-7B-Instruct) to detect the specific type of jailbreaking applied	before	extrinsic	safety	Harmful Content
MediQ: Question-Asking LLMs and a Benchmark for Reliable Interactive Clinical Reasoning - Advances in Neural Information Processing Systems	2024	General Medicine	QA	Prompting Strategies + consistency sampling + rationale generation	before	extrinsic	uncertainty	low confidence prediction
Trustworthy and Practical AI for Healthcare: A Guided Deferral System with Large Language Models	2025	Radiology	report classification	prompting (verbalized confidence) + hidden state prediction + threshold	before, after	extrinsic	uncertainty	low confidence prediction
Evaluating Large Language Models for Health-related Queries with Presuppositions	2024	General Medicine	claim verification	- (just evaluates)	-	-	uncertainty	False or Ambiguous Claims
{LTRC}-{IIITH} at {EHRSQL} 2024: Enhancing Reliability of Text-to-{SQL} Systems through Abstention and Confidence Thresholding - Proceedings of the 6th Clinical Natural Language Processing Workshop	2024	Clinical (EHR)	text-to-SQL	prompting (LLM used as classifier: abstain/not abstain), confidence thresholding	before	extrinsic	uncertainty	Inherently Unanswerable
Semantic Consistency-Based Uncertainty Quantification for Factuality in Radiology Report Generation	2025	Radiology	report generation	sampling + uncertainty thresholding	after	extrinsic	uncertainty	low confidence prediction
MedSafetyBench: Evaluating and Improving the Medical Safety of Large Language Models - Advances in Neural Information Processing Systems	2024	Medical Safety	Alignment Evaluation	fine tuning	within	intrinsic	safety	Adversarial Prompt Harmful Content
Energy Landscapes Enable Reliable Abstention in Retrieval-Augmented Large Language Models for Healthcare	2025	obstetrics and gynaecology	QA	OOD score + trained a MLP to classify -> confidence estimation	before	extrinsic	uncertainty	OOD, outside healthcare
Probing Hidden States for Calibrated, Alignment-Resistant Predictions in LLMs	2025	Medical Safety	Alignment Evaluation	lightweight classifier trained on frozen LLM activations; uses signals from selected layers to estimate calibrated confidence and detect possible misalignment	within	extrinsic	safety	Adversarial Prompt
LMOD: A Large Multimodal Ophthalmology Dataset and Benchmark for Large Vision-Language Models	2024	Ophthalmology	report classification	- (just evaluates)	-	-	uncertainty	Incomplete Information
Smarter Together: Combining Large Language Models and Small Models for Physiological Signals Visual Inspection	2025	cardiology and sleep medicine	clinical decision support	ConMIL (MIL + QTrans-Pooling + conformal calibration) → interpretable, confidence-aware classifier for physiological signals.	before	extrinsic	uncertainty	low confidence prediction
Poison Once, Refuse Forever: Weaponizing Alignment for Injecting Bias in LLMs	2025	Medical Safety	Alignment Evaluation	fine tuning	within	intrinsic	safety	Adversarial Prompt
Trustworthy Agents for Electronic Health Records through Confidence Estimation	2025	Clinical (EHR)	QA	prompting (verbalized confidence) + another LLM that looks at the main's LLM reasoning and computes a score	before, after	extrinsic	uncertainty	low confidence prediction
ASTRID - An Automated and Scalable TRIaD for the Evaluation of RAG-based Clinical Question Answering Systems	2025	Ophthalmology	QA	- (just evaluates)	-	-	uncertainty	outside healthcare
AT-CXR: Uncertainty-Aware Agentic Triage for Chest X-rays	2025	Radiology	image classification	CNN classification probability and OOD score - input into a LLM acting as a router	before	extrinsic	uncertainty	OOD, low confidence prediction
Don't Hallucinate, Abstain: Identifying LLM Knowledge Gaps via Multi-LLM Collaboration	2024	medical genetics	QA	two multi-LLM abstention methods—Cooperate and Compete	after	extrinsic	uncertainty	inherently unanswerable, OOD
AbstentionBench: Reasoning LLMs Fail on Unanswerable Questions	2025	General Medicine	QA	prompting	before	extrinsic	uncertainty	False/Ambiguous Query Inherently Unanswerable Incomplete Information
Characterizing LLM Abstention Behavior in Science QA with Context Perturbations	2024	General Medicine	QA	prompting (removing, replacing, and augmenting provided contexts to control the answerability of questions)	before	extrinsic	uncertainty	False/Ambiguous Query Incomplete Information
Learn to Refuse: Making Large Language Models More Controllable and Reliable through Knowledge Scope Limitation and Refusal Mechanism	2024	General Medicine	QA	Knowledge Base (KB) of verified factual entries that defines what an LLM can safely answer	before	extrinsic	uncertainty	low confidence prediction

Llamas Know What GPTs Don't Show: Surrogate Models for Confidence Estimation	2023	General Medicine	QA	use a surrogate model's confidence and combine them with the target model's	after	extrinsic	uncertainty	low confidence prediction
Keeping LLMs Aligned After Fine-tuning: The Crucial Role of Prompt Templates	2025	Medical Safety	QA	fine tuning	within	intrinsic	safety	Harmful Content
InferAligner: Inference-Time Alignment for Harmlessness through Cross-Model Guidance	2024	Medical Safety	QA	inference-time activation steering	within	extrinsic	safety	Harmful Content
Learning to Say "I Don't Know": A Vision for Abstention in Large Language Models	2025	General Medicine	QA	- (just evaluates)	-	-	uncertainty	low confidence prediction