VERITAS: A Multi-Agent Co-Scientist for Verifiable Image-Derived Hypothesis Testing

Lucas Stoffl[1,2], Benedikt Wiestler[3,4,5], and Johannes C. Paetzold[1,2]

[1]Department of Radiology, Weill Cornell Medicine, New York, USA
[2]Cornell Tech, New York, USA, [3]TUM University Hospital, Munich, Germany
[4]Technical University of Munich, [5]Munich Center for Machine Learning
lms465@cornell.edu, jpaetzold@med.cornell.edu

Abstract. Scientific research based on multimodal clinical data (including medical imaging) requires coordinating clinical, radiological, programming, and biostatistical expertise, a fragmented process that bottlenecks discovery. We present VERITAS (**V**erifiable **E**pistemic **R**easoning for **I**mage-Derived Hypothesis **T**esting via **A**gentic **S**ystems), a clinical co-scientist: a multi-agent system that **autonomously** tests natural-language hypotheses and produces a fully auditable evidence trail, tracing every conclusion through executable outputs from analysis plan to segmentation masks to statistical code to final verdict. Unlike prior AI-scientist systems, which mainly operate on tabular or text data, VERITAS grounds autonomous discovery directly in medical images. It decomposes the workflow into four phases handled by role-specialized agents, and introduces an *epistemic evidence label framework* that mechanically classifies outcomes as SUPPORTED, REFUTED, UNDERPOWERED, or INVALID by jointly evaluating significance, effect direction, and study power. This distinction is critical in medical imaging, where non-significant results often reflect insufficient sample size rather than absent effects. We construct a tiered benchmark of 64 hypotheses spanning six complexity levels across cardiac and brain glioma MRI datasets. VERITAS reaches 81.4% verdict accuracy with frontier models and 71.2% with locally-hosted open-weight models (8–30B), outperforming all single-model baselines in both classes. It also produces the highest rate of independently verifiable statistical outputs (86.6%), so even its failures remain diagnosable through artifact inspection. Structured multi-agent decomposition thus substitutes for model scale while preserving the verifiability that scientific discovery demands. We release code, hypothesis bank, and evaluation pipeline at https://github.com/LucZot/veritas.

Keywords: AI Co-Scientist · Multi-Agent Systems · Large Language Models · Scientific Discovery · Medical Image Analysis

1 Introduction

Scientific knowledge in medicine advances when clinicians translate bedside observations into formal hypotheses and test them in observational studies. Doing

so requires coordinating expertise across clinical knowledge, patient data curation and analysis, programming, and biostatistics. In the data domain, medical images are indispensable; they encode rich clinical information, including volumetric measurements, morphological features, and their associations with patient outcomes. Yet turning these image data into statistically valid insights, i.e., the gold standard for scientific evidence, remains a particularly fragmented, expert-intensive process. For example, a neurooncologist who wants to study if the more aggressive tumor biology of glioblastoma is reflected in larger contrast-enhancing tumors compared to lower-grade gliomas must coordinate image segmentation, metric extraction, statistical test selection, and result interpretation across multiple software tools and disciplinary boundaries. This pipeline bottlenecks discovery and limits who can participate in image-based research.

Recent "AI scientist" and "co-scientist" efforts apply large language models (LLMs) and multi-agent systems to automate scientific discovery [16, 17, 29]. Extending them to medical imaging hypothesis testing poses three challenges: (i) the pipeline must bridge complex and distinct tasks, including *cohort definition* (selecting the right patients from a given dataset), *vision* (segmenting anatomical structures), *computation* (deriving biomarkers), and *statistics* (selecting, executing and interpreting tests), requiring models to reason jointly across all tasks; (ii) deriving clinical conclusions demands full *auditability*, where every step from segmentation mask to p-value calculation must be inspectable and reproducible; and (iii) small clinical cohorts require *epistemic awareness*, which is the ability to distinguish "no effect detected" from "insufficient power to detect an effect."

Our main contribution is VERITAS (**V**erifiable **E**pistemic **R**easoning for **I**mage-Derived Hypothesis **T**esting via **A**gentic **S**ystems), a multi-agent framework inspired by clinician-scientist workflows. With role-specialized agents (PI, Imaging Specialist, Statistician), VERITAS addresses the above challenges and reaches the following capabilities:

1. **Fully autonomous clinical hypothesis testing.** VERITAS transforms natural language hypotheses into statistically validated conclusions with zero human intervention, managing the full study cycle to generate structured plans, executable code, statistics, and visualizations.
2. **Multimodal clinical data understanding.** The framework integrates medical imaging (e.g., MRI) and clinical metadata using specialized tools, including promptable segmentation models, to derive quantitative biomarkers directly from raw image data.
3. **Auditable and verifiable at each step.** VERITAS produces a complete, transparent execution trail. All results are derived from executable Python scripts and verifiable segmentation masks, allowing clinicians to reproduce and audit both the underlying logic and visual evidence.
4. **High performance in validation.** Validated on 64 hypotheses across cardiac and neuro-oncology domains, VERITAS achieves 71.2% majority-vote verdict accuracy with locally-deployed open-weight models and 81.4% with frontier models, outperforming all baselines in each model class.

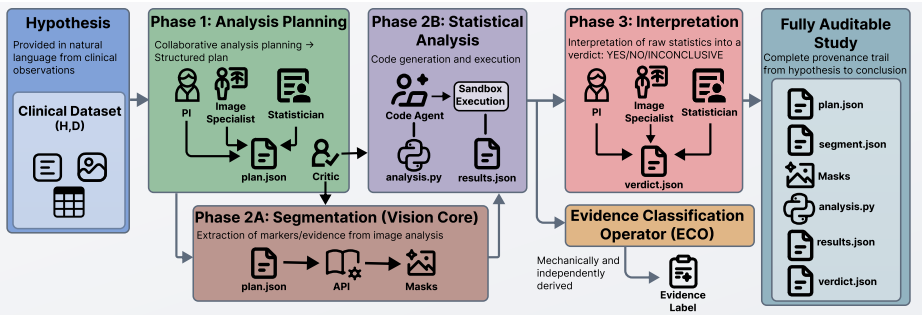

Fig. 1: VERITAS architecture and evidence flow. Given a natural language hypothesis and dataset, three role-specialized agents (PI, Imaging Specialist, Statistician) with a Critic collaborate through four phases: (1) collaborative analysis planning producing a structured plan; (2A) neural segmentation, grounding all evidence in image-derived masks; (2B) sandboxed code generation and execution yielding statistical outputs; and (3) interpretation of the Phase 2B statistical outputs in the context of the accumulated workflow history to produce a verdict. The Evidence Classification Operator (ECO) mechanically derives the evidence label from Phase 2B outputs. Every phase produces a versioned artifact (right), forming a complete provenance trail from hypothesis to conclusion.

5. **Structured decomposition over scale.** By utilizing role-specialized agents, VERITAS enables small, locally-deployed models (8–30B) to perform complex reasoning that typically requires frontier models, facilitating secure, on-site clinical deployment.

Further contributions: To systematize the conclusions our agentic system produces, we propose the **epistemic evidence label framework**. Here, we introduce a four-label classification SUPPORTED, REFUTED, UNDERPOWERED, INVALID that mechanically resolves the epistemic status of a hypothesis from observed statistics and a smallest effect size of interest (SESOI), without relying on agent judgment. This distinction is essential for medical imaging benchmarks, where mistaking underpowered subgroup analyses as refutations inflates apparent accuracy (Sec. 3.3).

A central question is whether a multi-agent system like VERITAS can deliver correct conclusions and verifiable evidence across increasing complexity; to answer it, we contribute a tiered *evaluation benchmark* spanning six complexity levels, from metadata-only queries (L1) to multivariate survival models (L5), and untestable hypotheses (L0). For all hypotheses, the ground truth was established through dataset-derived reference statistics validated by domain experts. The benchmark includes positive, negative, no-effect, underpowered, and untestable controls, thereby probing the full range of the evidence-label framework. We evaluate on two public datasets across distinct clinical domains: cardiology (ACDC [6], cardiac MRI, 150 subjects) and neuro-oncology (UCSF-PDGM [7], brain glioma MRI, 501 subjects).

2 Related Work

Recent advances span medical image foundation models, agentic scientific systems, tool-augmented language models, and statistical best practices. Each addresses a component of the scientific workflow, but they have largely evolved independently; we review them below and position VERITAS as a unifying, auditable hypothesis-testing framework.

Medical image analysis and vision–language models. Foundation models have advanced promptable, generalizable medical segmentation across anatomies and modalities. The Segment Anything Model (SAM) [24] introduced spatial-prompt segmentation, since extended to video and concept prompts [9, 38] and adapted to the medical domain by MedSAM and successors [30, 31, 53]. SAT [52] and VoxTell [39] further allow targets to be specified in natural language, unlike specialist architectures such as nnU-Net [21] that require dedicated training per structure. This makes segmentation programmatically accessible to the code-writing agents our framework relies on.

In parallel, biomedical vision–language models reason jointly over images and clinical text [5, 27, 51]. Closest to an agent-driven imaging pipeline, VoxelPrompt [19] uses language-guided code generation to analyze 3D medical images. Yet these systems operate at the perception level (segmentation, retrieval, question answering) rather than cohort-level inference: none derives statistically validated conclusions from image-derived measurements across patients.

LLMs for scientific discovery and hypothesis testing. Large language models increasingly support scientific workflows such as literature synthesis, hypothesis generation, and experimental planning [2, 12, 14, 36, 46], and several systems automate full research cycles spanning ML experimentation, manuscript writing, and wet-lab validation [16, 17, 29, 32]. Popper [20] is particularly related, validating hypotheses through sequential falsification with formal error control, but operates on tabular and text data without medical imaging or epistemic power analysis. More broadly, programmatic reasoning, where models generate and execute code rather than reasoning in text, improves reliability on quantitative tasks [10, 13, 44], motivating VERITAS's code-first statistical analysis.

In the medical domain, agent-based frameworks apply LLM collaboration to diagnostic reasoning and clinical decision support [11, 22, 23, 28, 34, 43]. Most closely related, MedAgent-Pro [45] pairs segmentation tools with coding agents to compute clinical indicators from images, but targets per-patient diagnosis rather than population-level hypothesis testing. None of these systems address vision-grounded hypothesis testing, which requires bridging image-derived measurements with cohort-level statistical analysis and epistemic classification.

Multi-agent systems for scientific reasoning. Multi-agent systems decompose complex scientific workflows across specialized roles, with teams that collaboratively plan, code, and execute experiments [3, 15, 40–42, 49] built on general-purpose orchestration frameworks [18, 47]; MetaGPT's insight that structured

intermediate artifacts curb cascading hallucinations directly informs our design. A recent independent evaluation, however, found that none of several such frameworks completed an end-to-end research cycle, with agents frequently hallucinating results during implementation [1] (detailed in the supplementary material). Closest to our setting, MESHAgents [50] orchestrates agents for phenotype-wide association studies in cardiovascular imaging, but reasons over pre-extracted tabular phenotypes rather than deriving measurements from images, and provides no epistemic classification of evidence strength.

Statistical evidence and epistemic classification. Equivalence testing via the TOST procedure and the smallest effect size of interest (SESOI) provides a principled way to distinguish "no meaningful effect" from "insufficient power to detect" [25]; Lakens et al. [26] formalized this into a four-outcome decision framework that is the direct conceptual precursor of our evidence label taxonomy (SUPPORTED, REFUTED, UNDERPOWERED, INVALID). VERITAS operationalizes this power-aware classification within an automated pipeline: the Evidence Classification Operator mechanically maps observed statistics to epistemic labels without relying on agent judgment, automating a framework that prior work proposed only as manual analytical guidance (see the supplementary material for an extended discussion).

Across these four directions, no prior system unifies vision-grounded measurement, executable statistical validation, and power-aware evidence calibration within a single auditable pipeline: imaging models stop at perception, scientific and medical agents operate largely on tabular or textual data, and statistical-evidence frameworks remain manual analytical guidance. VERITAS closes this gap by linking segmentation outputs, statistical code, and final conclusions through verifiable artifacts, enabling autonomous, transparent, and reproducible hypothesis testing on medical imaging datasets.

3 Methods

We present VERITAS, a multi-agent framework that transforms natural-language hypotheses about medical imaging datasets and their associated metadata into statistically validated conclusions through autonomous, auditable execution.

3.1 Problem Formulation

Given a dataset $\mathcal{D} = \{(x_i, m_i)\}_{i=1}^{N}$ where x_i represents medical images for patient i and m_i denotes patient metadata (including cohort variables, clinical covariates, and outcomes such as survival endpoints), together with a natural-language hypothesis $\mathcal{H}$, we seek to autonomously determine whether $\mathcal{H}$ is SUPPORTED, REFUTED, UNDERPOWERED, or INVALID given the available evidence.

Unlike traditional computational pipelines that require explicit metric definitions and test specifications, our framework interprets $\mathcal{H}$ semantically, checks feasibility against available data, designs an appropriate analysis plan, and executes it end-to-end with an auditable artifact trail.

3.2 Framework Architecture

VERITAS operates through four sequential phases, each producing structured, inspectable outputs that feed into subsequent phases (Fig. 1). Three role-specialized agents collaborate: a **Principal Investigator** (scientific validity, feasibility), a **Medical Imaging Specialist** (segmentation targets, imaging protocols), and a **Statistician** (test selection, effect sizes, power analysis). A separate **Critic** agent operates by default in execution phases (2A/2B). We evaluate two model deployments with the same workflow and prompts: (i) a local open-weight team (GPT-OSS-20B [37] for PI/imaging, Qwen3-8B [48] for discussions and critic, Qwen3-Coder-30B [8] for coding), and (ii) a frontier team (GPT-5.2 [36] for PI/imaging/coding and GPT-5 Mini [35] for discussions and critic).

Dataset API abstraction. To keep the framework dataset-agnostic and auditable, agents never access raw files directly during analysis. Instead, they interact through a constrained API layer that exposes cohort discovery, per-patient metadata, observation identifiers, segmentation masks, and geometry-aware measurement utilities. This API boundary standardizes provenance and enables automated checks for off-contract data access. Detailed API functions are described in the supplementary material.

Phase 1: Analysis Planning. Given $\mathcal{H}$ and dataset metadata (available groups, observations, patient fields), the agents collaboratively produce a structured analysis plan specifying the target cohorts, population restriction, required anatomical structures and observations, derived measurements, and the planned statistical analysis. The plan includes a feasibility decision (`TESTABLE`/`UNTESTABLE`) with subtype and missing requirements; a programmatic validator checks schema completeness and consistency against the available dataset metadata before execution. If marked untestable, the run terminates as INVALID. The Statistician also computes *a priori* power from the queried cohort sizes under a fixed planning SESOI.

Phase 2A: Segmentation. When the hypothesis requires image-derived measurements, the Imaging Specialist generates a segmentation request specifying the relevant patients, target structures, and observations. A segmentation backend (SAT [52]) processes this request and stores the resulting binary masks in a shared database. This phase is the vision core of the framework: subsequent quantitative analyses are grounded in measurements extracted from these masks. For hypotheses that require no imaging structures, this phase is bypassed.

Phase 2B: Statistical Analysis. The Statistician writes executable code that retrieves the required masks and metadata through the API, computes the planned measurements, performs the statistical test, and reports effect size δ, 95% confidence interval CI_{95}, p-value p, and sample sizes. The code is executed in a sandboxed environment with timeout constraints. Failed or invalid attempts trigger iterative revision using execution traces and validator feedback, with optional Critic input, up to 8 attempts per round.

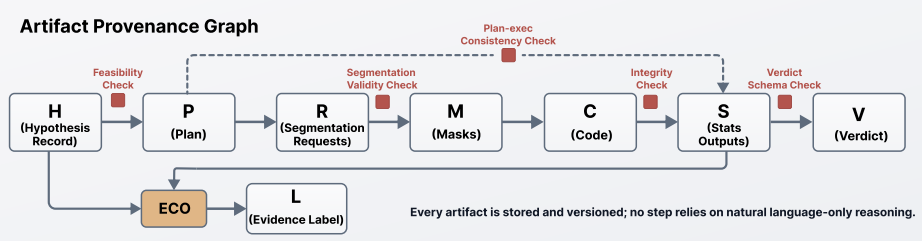

Fig. 2: Auditability through artifact provenance. Every phase produces versioned, inspectable artifacts. Solid checks are in-workflow validity gates; the dashed check is a post-hoc plan–execution consistency test applied by the evaluator. The evidence label is derived mechanically from Phase 2B statistics via the Evidence Classification Operator (ECO; Sec. 3.3), independent of agent judgment.

Table 1: Evidence label definitions. Labels are mechanically derived from Phase 2B statistics with no agent judgment involved.

Label	Condition	Interpretation
SUPPORTED	$p < \alpha \wedge$ direction matches	Evidence confirms hypothesis
REFUTED	$(p \geq \alpha \wedge \pi \geq 0.80)$	Adequate power, no effect
	$\vee\, (p < \alpha \wedge$ opposite dir.)	Significant opposite effect
UNDERPOWERED	$p \geq \alpha \wedge \pi < 0.80$	Insufficient power to conclude
INVALID	Untestable or execution failure	Results unreliable

Phase 3: Interpretation. Agents synthesize the prior workflow outputs through structured discussion, producing a verdict $v \in \{\texttt{YES}, \texttt{NO}, \texttt{INCONCLUSIVE}\}$ with supporting rationale. Phase 3 has access to the relevant planning and statistical outputs from earlier phases, while the evaluator-computed evidence label (see Sec. 3.3) is explicitly withheld.

Auditability. A key design goal is that every analysis is fully auditable. Each phase produces structured artifacts (e.g., JSON plans, segmentation requests, Python code, statistical outputs, verdicts) rather than natural language summaries (Fig. 2). Validity checks enforce consistency at each boundary, and the final statistical output is retained in a form that enables independent re-verification (see the *evidence grounding* metric in Sec. 3.4). The supplementary material provides the full list of per-phase validity checks.

3.3 Evidence Label Framework

A central contribution is our four-label evidence classification that jointly evaluates statistical significance, effect direction, and study power (Table 1).
Here, π denotes statistical power computed at a smallest effect size of interest (SESOI) and $\alpha = 0.05$. This functions as a *decision-theoretic epistemic operator*: given observed statistics (p, δ, n) and a pre-specified SESOI δ_0, the evidence label is a deterministic function, fully removing subjective judgment from the classification. The epistemic distinction between REFUTED and UNDERPOWERED

Table 2: Tiered hypothesis bank. Six complexity tiers probe progressively harder capabilities. **Neg.** includes negative, no-effect, and nonsense controls. **Undp./Unt.** = underpowered ($\pi < 0.80$) or untestable hypotheses.

Tier	Name	n	Capability probed	Pos.	Neg.	Undp./Unt.
L0	Untestable	5	Feasibility detection	–	–	5
L1	Metadata-only	9	Statistical reasoning	3	5	1
L2	Single imaging	23	Vision-to-statistics pipeline	15	8	–
L3	Engineered feat.	6	Feature engineering	6	–	–
L4	Mixed	16	Multi-modal integration	8	3	5
L5	Multivariate	5	Advanced statistical reasoning	5	–	–
	Total	**64**		**37**	**16**	**11**

is critical for medical imaging benchmarks: a non-significant result with power $\pi = 0.38$ (e.g., $n = 13$ patients) provides essentially no evidence against the hypothesis, yet standard binary evaluation would score it as a correct rejection. Details on power computation are in the supplementary material.

3.4 Benchmark & Evaluation

For reproducible benchmarking, we construct a bank of 64 hypotheses (32 per dataset), spanning six complexity tiers (Table 2). Each tier contains a mix of positive controls (37), negative/no-effect controls (16), and underpowered or untestable controls (11). Ground-truth evidence labels are established through canonical reference statistics computed from ground-truth segmentation masks and patient metadata, validated by a domain expert. Full per-hypothesis specifications are in the supplementary material.

Unless stated otherwise, all accuracy metrics are computed on *testable* hypotheses (L1–L5); L0 feasibility accuracy is reported separately. We evaluate with:

- **Evidence-label accuracy**: agreement between evaluator-computed evidence labels (SUPPORTED, REFUTED, UNDERPOWERED) and ground truth, requiring executable statistical output.
- **Verdict accuracy**: agreement between final verdicts (YES, NO, INCONCLUSIVE) and ground truth.
- **Majority-vote accuracy**: hypothesis-level aggregation over repeated runs; the majority verdict/label across runs is compared to ground truth.
- **L0 feasibility accuracy**: fraction of untestable hypotheses correctly classified as INVALID, reported separately from testable hypothesis performance.

For repeated runs, we report both run-level (micro-averaged) and majority-vote (hypothesis-level) metrics. Evidence-label and verdict accuracy are each computed over runs producing valid output for that metric; these pools may differ when a run yields a verdict but incomplete statistical output. We report completion rate separately for VERITAS to quantify execution reliability. We additionally report **diagnostic rates** across three categories: *conclusion quality*

(overclaim, false refutation), *analysis integrity* (hallucinated significance, synthetic data, literal p-value assignment), and *auditability* (evidence grounding rate, measuring whether runs report the four core statistical outputs required for independent verification). Complete metric definitions are in the supplementary material.

4 Experiments

4.1 Experimental Setup

Datasets We evaluate on two public medical imaging datasets from cardiology and neuro-oncology (per-dataset tier coverage in the supplementary material).

ACDC [6] contains 150 subjects (100 training + 50 testing) with short-axis cardiac cine MRI across five diagnostic groups (30 per group): dilated cardiomyopathy (DCM), hypertrophic cardiomyopathy (HCM), myocardial infarction (MINF), right ventricular abnormality (RV), and normal controls (NOR). Each subject includes end-diastolic (ED) and end-systolic (ES) phases with expert-annotated ground-truth segmentations of left ventricle (LV), right ventricle (RV cavity), and myocardium. Patient metadata includes height, weight, and cardiac timing parameters. Hypotheses span L0–L4.

UCSF-PDGM [7] is a preoperative diffuse glioma MRI dataset with 501 subjects across WHO grades 2-4 (2: 56, 3: 43, 4: 396) and molecular markers (IDH mutation, MGMT methylation, 1p/19q codeletion). Ground-truth tumor segmentation masks, following the BraTS conventions [4,33], are available for 495 subjects. Patient metadata includes age, sex, overall survival (395 with survival data; 228 events), and extent of resection. The SAT model [52] segments four brain tumor subregions: whole tumor (WT), necrotic core (NCR), peritumoral edema (ED), and enhancing tumor (ET). Hypotheses span L0–L5, including survival analyses and multivariate regression.

Baselines We compare VERITAS against a structured ablation of five baselines.

Single-model baselines (SMa–SMe). To isolate the contributions of code execution, data access strategy, iterative self-correction, and structured phase guidance, we define our baselines forming an ablation progression:

- **SMa: Direct reasoning.** The model receives per-group summary statistics pre-computed from ground-truth segmentation and produces a verdict without any code.
- **SMb: Code on pre-computed features.** The model receives a per-patient CSV of pre-computed imaging features (volumes, masses, etc.) and writes one Python script executed automatically in the sandbox.
- **SMc: Code via API.** The model receives only the hypothesis and API documentation. It writes a single end-to-end script covering patient discovery, mask loading, metric computation, and statistical testing.
- **SMd: Agentic (iterative).** Same API access as SMc, but the model operates in an iterative loop (up to 3 rounds): write code $\rightarrow$ execute $\rightarrow$ observe output $\rightarrow$ refine or conclude.

– **SMe: Pipeline (structured).** Same API access as SMd, but guided by VERITAS's exact phase agendas (Phase 1 planning, Phase 2 coding, Phase 3 interpretation) as a single-model chain-of-thought.

These baselines isolate four capability dimensions: (i) code execution vs. pure reasoning (SMa→SMb); (ii) pre-extracted features vs. raw API access (SMb→SMc); (iii) iterative self-correction (SMc→SMd); and (iv) structured phase guidance (SMd→SMe). Comparing SMe to VERITAS then directly isolates the contribution of multi-agent discussion over a single model following identical phase instructions.

Model and fairness. We evaluate all five baselines with two model families: **frontier** (GPT-5.2 [36]) and **local/open-weight** (GPT-OSS-20B [37]). For fairness, all baselines that use the API (SMc–SMe) receive *identical* API documentation to VERITAS agents, extracted directly from the same prompt templates. The full dataset-specific structure catalogue (*e.g.*, `"LV"`, `"RV"`, `"Myo"` for ACDC) is also provided, equivalent to what VERITAS's Phase 2A agents discover through tool calls; the model must still determine which structures are relevant to the hypothesis. Code execution uses the same sandboxed environment as VERITAS. For SMd and SMe, a single interpretation step (equivalent to VERITAS's Phase 3) is always invoked after successful code execution; the model's conclusion from this step is used as the final verdict.

4.2 Main Results

Overall performance. Each method was evaluated with 10 repeated runs per hypothesis (64 hypotheses, 640 runs per method), and we report run-level and majority-vote accuracy across both model families (Table 3).

Key findings. VERITAS achieves the highest accuracy in both model families: 71.2% majority-vote verdict accuracy with locally-deployed open-weight models (8–30B) and 81.4% with frontier models, outperforming all single-model baselines in each class. The ablation progression reveals consistent patterns:

Code execution is essential. SMa (direct reasoning) achieves only ∼56% verdict accuracy regardless of model scale, confirming that pure LLM reasoning over summary statistics is insufficient. Adding code execution (SMb) lifts verdict accuracy by 8–15 percentage points.

Iterative self-correction helps selectively. The SMc→SMd gain is modest for verdict accuracy (+2.8 local/open-weight, +3.8 frontier) but more pronounced for evidence accuracy (+7.3, +11.8), suggesting that self-correction primarily improves statistical analysis quality rather than final interpretation.

Multi-agent decomposition adds value. Comparing SMe to VERITAS isolates the multi-agent contribution: +10.2pp verdict majority vote for local/open-weight (71.2% vs. 61.0%) and +20.4pp for frontier (81.4% vs. 61.0%). The larger frontier gap suggests that multi-agent discussion becomes *more* valuable as base model capability increases.

Table 3: Main results on the 64-hypothesis tiered benchmark. Each cell shows accuracy (%). Evidence accuracy requires executable statistical output (not applicable for SMa). Metrics are computed on testable hypotheses (L1–L5); L0 feasibility is reported separately. **Compl.**: fraction of runs completing all workflow phases (Sec. 3.4); baselines always produce output by construction.

Method	Run-level		Majority vote		Compl.
	Evidence	Verdict	Evidence	Verdict	
Local/Open-weight models (GPT-OSS-20B)					
SMa: Direct reasoning	–	56.1	–	55.9	–
SMb: Code on features	58.1	63.8	61.0	62.7	–
SMc: Code via API	39.0	63.2	40.7	59.3	–
SMd: Agentic	46.3	66.0	52.5	66.1	–
SMe: Pipeline	55.1	58.0	57.6	61.0	–
VERITAS (local)	**63.4**	**71.4**	**67.8**	**71.2**	78.1
Frontier models (GPT-5.2)					
SMa: Direct reasoning	–	55.9	–	55.9	–
SMb: Code on features	63.7	70.7	69.5	69.5	–
SMc: Code via API	52.9	72.8	61.0	76.3	–
SMd: Agentic	64.7	76.6	64.4	74.6	–
SMe: Pipeline	60.8	65.5	59.3	61.0	–
VERITAS (frontier)	**70.0**	**81.0**	**76.3**	**81.4**	87.5

VERITAS (local) uses locally-deployed 8–30B models. L0 feasibility: VERITAS 74.0% (local), 100% (frontier); all baselines 100% except SMb-frontier (98%).

Per-dataset gap. Evidence-label accuracy drops sharply from ACDC to UCSF-PDGM across all methods (Table 4), reflecting genuine complexity differences (survival analyses, multivariate regression, molecular-marker stratification). On ACDC, frontier VERITAS achieves 90.0% evidence accuracy and 78.9% verdict at 88.8% completion. Pipeline failures on a subset of runs reduce verdict performance on easier hypotheses where baselines rarely fail. Local VERITAS, despite much smaller model capacities, remains competitive and attains the highest L5-tier scores. VERITAS's advantage emerges on harder UCSF-PDGM hypotheses, where it leads in both evidence (+16pp over SMd) and verdict accuracy (+29pp).

Stratified results. Table 4 provides complementary breakdowns by dataset and by complexity tier, comparing both VERITAS configurations against the three strongest frontier baselines.

Tier behavior. L2 (single imaging metric) is the strongest tier across all methods (71–90%), as these hypotheses require straightforward segmentation feature comparisons. L4 and L5 are hardest, driven by execution fragility in survival analyses and multi-covariate regression. Critically, VERITAS is the *only* method to achieve non-zero L5 evidence accuracy (16–30% across configurations), while all single-model baselines score 0% on L5 despite achieving up to 92% verdict accuracy on these hypotheses. This indicates they produce correct YES/NO guesses without valid statistical backing. The small L5 sample (5 hypotheses) limits precision, but the qualitative gap (non-zero vs. zero) is robust.

Table 4: Stratified run-level accuracy (%) by dataset and complexity tier. Per-dataset: evidence / verdict. Per-tier: evidence-label accuracy. Best non-VERITAS result per tier is underlined. L0 = feasibility detection. Full per-method table in supplementary.

	ACDC		UCSF-PDGM		Evidence by tier						
Method	**Ev.**	**Vd.**	**Ev.**	**Vd.**	**L0**	**L1**	**L2**	**L3**	**L4**	**L5**	**All**
VERITAS (local)	91.0	87.9	36.7	49.0	74	54.4	87.0	51.7	49.4	**30.0**	63.4
SMc: Code (GPT-5.2)	79.0	85.1	27.7	60.9	100	75.6	70.9	43.3	34.4	0.0	52.9
SMd: Agentic (GPT-5.2)	**95.9**	**96.6**	34.7	53.9	100	**94.4**	85.2	55.0	**42.5**	0.0	64.7
SMe: Pipeline (GPT-5.2)	90.3	96.9	32.3	35.1	100	56.7	89.1	56.7	43.1	0.0	60.8
VERITAS (frontier)	90.0	78.9	**50.7**	**83.2**	100	90.0	90.4	**73.3**	45.0	16.0	**70.0**

L1 scaling behavior. Evidence accuracy on L1 (metadata-only) reveals a pronounced model-capacity dependence: local VERITAS drops to 54.4% on L1 versus 87.0% on L2, while frontier VERITAS maintains near-parity (90.0% vs. 90.4%). Although L1 hypotheses require no imaging computation, they demand nontrivial statistical design capabilities (survival model specification, covariate adjustment, and appropriate test selection) that benefit disproportionately from larger model capacity. By contrast, L2 hypotheses (single imaging metric comparisons) rely on more mechanical code generation that local models already handle effectively.

4.3 Diagnostic Analysis

Table 5 reports diagnostic rates that characterize *how* methods fail, beyond aggregate accuracy; these are conditional rates over method-specific pools, not a flat quality ranking. SMa is excluded because it produces no executable statistical output, making code-level diagnostics inapplicable.

One-shot baselines (SMb, SMc) show high overclaim rates (35–65%) and hallucinated significance (9–12%), producing YES verdicts unsupported by their own statistical outputs. Iterative and structured baselines (SMd, SMe) largely eliminate hallucinated significance ($\leq$1.5%) through self-correction, but trade this for higher literal p-value assignment (3–17%), where models hard-code p-values rather than computing them from data.

Local VERITAS shows the lowest overclaim (7.2%) but the highest false refutation (24.2%) and synthetic-data rate (10.9%). Since evidence labels are assigned mechanically by the ECO, false refutation is not a conservative interpretation style but weaker code yielding null or invalid statistics on adequately powered, truly SUPPORTED hypotheses. Frontier VERITAS accordingly cuts false refutation to 12.6% as code improves, while its higher overclaim (20.0%) and literal-p rate (15.0%) follow from completing more, and more complex, analyses. The code-integrity scans also charge VERITAS for engaging the hardest tiers (it alone reaches non-zero L5), where baselines short-circuit. Both configurations keep hallucinated significance low ($\leq$2.7%), the most egregious failure mode the phased pipeline prevents.

The verifiability column (Verif.) quantifies auditability over all runs. Among methods operating on raw data (SMd, SMe, VERITAS), VERITAS achieves the

Table 5: Diagnostic rates (%) for code-producing methods on the combined benchmark. **Conclusion quality:** *Overclaim*: YES verdicts where evidence is not SUPPORTED; *False ref.*: predicted REFUTED labels disagreeing with ground truth. **Analysis integrity:** *Hall. sig.*: YES verdict with $p \geq 0.05$ or missing p-value (output-level check); *Synth.*: code generating mock/random data instead of using real measurements (code-level); *Lit. p*: significance threshold hard-coded in analysis code (code-level). **Auditability:** *Verif.* (↑): fraction of all runs reporting the four core statistical outputs (test type, sample sizes, effect size, p-value) required for independent verification.

	Conclusion quality		Analysis integrity			Audit.
Method	**Overcl.**	**False ref.**	**Hall. sig.**	**Synth.**	**Lit. p**	**Verif. ↑**
Local/Open-weight (GPT-OSS-20B)						
SMb: Features	44.7	7.4	10.9	0.2	0.7	81.9
SMc: Code	65.1	6.7	8.7	0.2	1.4	82.3
SMd: Agentic	25.6	10.2	1.5	0.2	12.2	53.6
SMe: Pipeline	3.8	8.5	0.0	0.9	17.2	57.7
VERITAS	7.2	24.2	2.7	10.9	2.4	77.3
Frontier (GPT-5.2)						
SMb: Features	34.7	15.3	11.7	2.2	4.1	82.8
SMc: Code	45.4	8.6	11.0	1.6	2.7	81.2
SMd: Agentic	14.8	11.7	0.0	0.6	3.3	70.5
SMe: Pipeline	13.0	7.9	0.0	0.5	10.9	68.1
VERITAS	20.0	12.6	0.4	3.6	15.0	86.6

Table 6: Ablation results for VERITAS (local, 10 runs per hypothesis). **Default**: 16k context, critic enabled in execution phases, T=0.2. Per-dataset split shows the ACDC vs. UCSF-PDGM gap across configurations.

	ACDC		UCSF-PDGM		Combined	
Configuration	**Evid.**	**Verd.**	**Evid.**	**Verd.**	**Evid.**	**Verd.**
Default (16k context)	91.0	87.9	36.7	49.0	63.4	71.4
Context 8k	91.4	91.7	29.0	45.7	59.7	72.0
Context 32k	90.0	87.1	31.7	51.8	60.3	72.7
No Critic	90.7	90.1	32.0	50.5	60.8	73.2
Temperature zero	90.3	88.7	27.7	47.5	58.5	72.6

highest verifiability (77.3% local, 86.6% frontier), substantially exceeding SMd (53.6%, 70.5%) and SMe (57.7%, 68.1%). SMb and SMc achieve higher absolute rates (∼82%) but operate on pre-computed features, yielding simpler code with higher completion rates, at the cost of zero L5 evidence accuracy. Thus VERITAS's phased decomposition improves not only accuracy but inspectability, keeping even its failures *recoverable* through artifact inspection.

4.4 Ablation Studies

Table 6 reports ablation results using the local VERITAS configuration (10 runs per hypothesis). We evaluate context window size, critic removal, and temperature zero decoding. Vision source ablations (GT masks, SAT-Pro) are reported in the supplementary material.

ACDC evidence remains stable across all ablations (∼90–91%), indicating that cardiac hypotheses are effectively solved by the base architecture; sensitivity lies almost entirely in UCSF-PDGM evidence, making it the discriminating benchmark. The default 16k context achieves the best UCSF evidence (36.7%): reducing to 8k degrades it by −7.7pp (complex survival/regression analyses no longer fit), while 32k gives no uplift (−5.0pp). Deterministic decoding (T=0) similarly hurts it (−9.0pp), consistent with complex hypotheses benefiting from stochastic code exploration. Critic removal has minimal impact (+1.8pp combined verdict), indicating the iterative code-revision loop already provides sufficient error correction.

4.5 Analysis

Execution vs. reasoning. Across all methods, verdict accuracy consistently exceeds evidence-label accuracy (*e.g.*, VERITAS local: 71.4% vs. 63.4%; frontier: 81.0% vs. 70.0%), indicating that the primary bottleneck is producing correct statistical outputs (code generation), not reasoning about them (interpretation). The gap is most striking for frontier VERITAS on UCSF-PDGM: verdict accuracy (83.2%) *exceeds* ACDC verdict (78.9%) despite much lower evidence accuracy (50.7% vs. 90.0%), identifying code generation for complex analyses (survival models, multivariate regression) as the primary improvement target. Notably, this verdict–evidence gap is much narrower for VERITAS (8–11 pp overall) than for single-model baselines on comparable tiers (*e.g.*, up to 92% verdict vs. 0% evidence accuracy on L5), confirming that VERITAS's verdicts remain substantially grounded in computed statistics rather than prior knowledge alone.

Epistemic label analysis. The four-label evidence framework is essential for fair evaluation. Without the UNDERPOWERED label, several UCSF-PDGM hypotheses with small subgroups would be conflated with true negatives: glioma_31 (1p/19q codeletion, n=13 vs. n=86, power=0.38) is correctly classified as UNDERPOWERED rather than REFUTED; glioma_14 (IDH-mutant Grade IV, n=28 vs. n=367, power=0.72) is similarly borderline. In a binary evaluation, a system that conservatively labels every non-significant result as "refuted" would achieve spuriously high accuracy. The underpowered distinction forces systems to demonstrate genuine epistemic awareness.

5 Discussion and Conclusion

We presented VERITAS, a multi-agent co-scientist for autonomous, auditable hypothesis testing inspired by clinician-scientist workflows. Its design targets three bottlenecks highlighted in the introduction: the fragmented coordination of clinical, imaging, programming, and statistical expertise; the lack of reproducible analysis trails in image-derived studies; and the misinterpretation of underpowered findings as true negatives. By decomposing the workflow into verifiable phases and code, VERITAS makes every step auditable and renders its failures *recoverable* in a way that baseline failures are not. It achieves strong performance even against baselines that receive human assistance (e.g., pre-computed

per-group statistics or per-patient feature tables), and enables small, locally-deployable models to perform complex reasoning that typically requires frontier models.

Recall that VERITAS produces two complementary outcomes: the agents' *verdict* (YES/NO/INCONCLUSIVE), reflecting their synthesized judgment, and the *evidence label* (SUPPORTED/REFUTED/UNDERPOWERED/INVALID), derived mechanically by the Evidence Classification Operator and withheld from the agents. This separation is intentional: comparing the two reveals whether a conclusion is actually warranted by the statistics the agents produced, and distinguishes a true null from a merely underpowered one. Beyond accuracy, the evidence label framework functions as a mechanistic decision rule: given observed statistics (p, δ, π), the label is a deterministic function with no free parameters beyond the significance level α and the smallest effect size of interest (SESOI, δ_0). This eliminates evaluation ambiguities common in agent benchmarks. We view this as a contribution beyond medical imaging: any benchmark involving statistical testing could adopt this framework to avoid conflating negative results with underpowered designs.

Most agent evaluations report only end-to-end accuracy. For scientific applications, *how* a conclusion is reached matters as much as *whether* it is correct. Our evidence grounding metric (Verif., Table 5) quantifies this: among methods operating on raw data, VERITAS produces the highest rate of fully verifiable outputs (86.6% frontier), meaning even its incorrect conclusions can be traced, diagnosed, and corrected through artifact inspection.

By reducing the coordination overhead and execution burden of patient selection, image analysis and biomarker extraction, and statistical testing, VERITAS can accelerate clinical research as an auditable collaborator in the human-led scientific discovery loop. Ultimately, a clinical observational study should always be finalized by a human expert, a process that is substantially facilitated through the availability and auditability of the reasoning steps, artifacts, and code.

Limitations

Segmentation dependency. All imaging-derived conclusions depend on segmentation quality. We quantify the gap between ground-truth and SAT segmentations in the supplement, but the framework treats segmentation as a point estimate, leaving uncertainty propagation to future work.

Benchmark scope. Our evaluation covers two MRI domains analyzed by a specific segmentation model; extending to other modalities (CT, histopathology) requires new but widely available segmentation models. The hypothesis bank also does not yet cover longitudinal or causal analyses.

Ground truth. Our ground truth verdicts are established from the given datasets using expert-validated reference computations. This was strictly necessary to ensure mechanical verifiability in our evaluation, but may reflect dataset-specific rather than population-level effects.

Acknowledgments

We thank Adina Scheinfeld for help with figures and writing.

References

1. Agrawal, S., Anadkat, H.B., Athimoolam, K.K., Bhardwaj, H., Chowdhury, T., Gao, S., Kamat, P.K., Makwana, V., Shariff, M.H., Badkul, A., et al.: Can ai conduct autonomous scientific research? case studies on two real-world tasks. bioRxiv pp. 2026–01 (2026)
2. Anthropic: The claude 3 model family: Opus, sonnet, haiku. Anthropic (2024)
3. Baek, J., Jauhar, S.K., Cucerzan, S., Hwang, S.J.: Researchagent: Iterative research idea generation over scientific literature with large language models. arXiv preprint arXiv:2404.07738 (2024)
4. Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., Ha, S.M., Rozycki, M., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv preprint arXiv:1811.02629 (2018)
5. Bannur, S., Hyland, S., Liu, Q., Perez-Garcia, F., Ilse, M., Castro, D.C., Boecking, B., Sharma, H., Bouzid, K., Thieme, A., et al.: Learning to exploit temporal structure for biomedical vision-language processing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15016–15027 (2023)
6. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al.: Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE transactions on medical imaging **37**(11), 2514–2525 (2018)
7. Calabrese, E., Villanueva-Meyer, J.E., Rudie, J.D., Rauschecker, A.M., Baid, U., Bakas, S., Cha, S., Mongan, J.T., Hess, C.P.: The university of california san francisco preoperative diffuse glioma mri dataset. Radiology: Artificial Intelligence **4**(6), e220058 (2022)
8. Cao, R., Chen, M., Chen, J., Cui, Z., Feng, Y., Hui, B., Jing, Y., Li, K., Li, M., Lin, J., et al.: Qwen3-coder-next technical report. arXiv preprint arXiv:2603.00729 (2026)
9. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
10. Chen, W., Ma, X., Wang, X., Cohen, W.W.: Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. arXiv preprint arXiv:2211.12588 (2022)
11. Chen, W., Dong, Y., Ding, Z., Shi, Y., Zhou, Y., Zeng, F., Luo, Y., Lin, T., Su, Y., Wu, Y., et al.: Radfabric: Agentic ai system with reasoning capability for radiology. arXiv preprint arXiv:2506.14142 (2025)
12. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
13. Gao, L., Madaan, A., Zhou, S., Alon, U., Liu, P., Yang, Y., Callan, J., Neubig, G.: Pal: Program-aided language models. In: International conference on machine learning. pp. 10764–10799. PMLR (2023)
14. Gao, S., Fang, A., Huang, Y., Giunchiglia, V., Noori, A., Schwarz, J.R., Ektefaie, Y., Kondic, J., Zitnik, M.: Empowering biomedical discovery with ai agents. Cell **187**(22), 6125–6151 (2024)
15. Ghafarollahi, A., Buehler, M.J.: Sciagents: automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. Advanced Materials **37**(22), 2413523 (2025)

16. Ghareeb, A.E., Chang, B., Mitchener, L., Yiu, A., Szostkiewicz, C.J., Laurent, J.M., Razzak, M.T., White, A.D., Hinks, M.M., Rodriques, S.G.: Robin: A multi-agent system for automating scientific discovery. arXiv preprint arXiv:2505.13400 (2025)
17. Gottweis, J., Weng, W.H., Daryin, A., Tu, T., Palepu, A., Sirkovic, P., Myaskovsky, A., Weissenberger, F., Rong, K., Tanno, R., et al.: Towards an ai co-scientist. arXiv preprint arXiv:2502.18864 (2025)
18. Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Wang, J., Zhang, C., Wang, Z., Yau, S.K.S., Lin, Z., et al.: Metagpt: Meta programming for a multi-agent collaborative framework. In: The twelfth international conference on learning representations (2023)
19. Hoopes, A., Dey, N., Butoi, V.I., Guttag, J.V., Dalca, A.V.: Voxelprompt: A vision agent for end-to-end medical image analysis. arXiv preprint arXiv:2410.08397 (2024)
20. Huang, K., Jin, Y., Li, R., Li, M.Y., Candès, E., Leskovec, J.: Automated hypothesis validation with agentic sequential falsifications. arXiv preprint arXiv:2502.09858 (2025)
21. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
22. Jin, Q., Wang, Z., Yang, Y., Zhu, Q., Wright, D., Huang, T., Khandekar, N., Wan, N., Ai, X., Wilbur, W.J., et al.: Agentmd: Empowering language agents for risk prediction with large-scale clinical tool learning. Nature Communications **16**(1), 9377 (2025)
23. Kim, Y., Park, C., Jeong, H., Chan, Y.S., Xu, X., McDuff, D., Lee, H., Ghassemi, M., Breazeal, C., Park, H.W.: Mdagents: An adaptive collaboration of llms for medical decision-making. Advances in Neural Information Processing Systems **37**, 79410–79452 (2024)
24. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
25. Lakens, D.: Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. Social psychological and personality science **8**(4), 355–362 (2017)
26. Lakens, D., Scheel, A.M., Isager, P.M.: Equivalence testing for psychological research: A tutorial. Advances in methods and practices in psychological science **1**(2), 259–269 (2018)
27. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)
28. Li, J., Lai, Y., Li, W., Ren, J., Zhang, M., Kang, X., Wang, S., Li, P., Zhang, Y.Q., Ma, W., et al.: Agent hospital: A simulacrum of hospital with evolvable medical agents. arXiv preprint arXiv:2405.02957 (2024)
29. Lu, C., Lu, C., Lange, R.T., Foerster, J., Clune, J., Ha, D.: The ai scientist: Towards fully automated open-ended scientific discovery. arXiv preprint arXiv:2408.06292 (2024)
30. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature communications **15**(1), 654 (2024)
31. Ma, J., Yang, Z., Kim, S., Chen, B., Baharoon, M., Fallahpour, A., Asakereh, R., Lyu, H., Wang, B.: Medsam2: Segment anything in 3d medical images and videos. arXiv preprint arXiv:2504.03600 (2025)

32. Mall, U., Phoo, C.P., Chiquier, M., Hariharan, B., Bala, K., Vondrick, C.: Disciple: Learning interpretable programs for scientific visual discovery. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29258–29267 (2025)
33. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014)
34. Nori, H., Daswani, M., Kelly, C., Lundberg, S., Ribeiro, M.T., Wilson, M., Liu, X., Sounderajah, V., Carlson, J., Lungren, M.P., et al.: Sequential diagnosis with language models. arXiv preprint arXiv:2506.22405 (2025)
35. OpenAI: Gpt-5 mini (1 2025), `https://openai.com/index/gpt-5-mini/`, accessed: 2026-03-03
36. OpenAI: Introducing gpt-5.2 (12 2025), `https://openai.com/index/introducing-gpt-5-2/`, accessed: 2026-03-03
37. OpenAI: Introducing gpt-oss (8 2025), `https://openai.com/index/introducing-gpt-oss/`, accessed: 2026-03-03
38. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
39. Rokuss, M., Langenberg, M., Kirchhoff, Y., Isensee, F., Hamm, B., Ulrich, C., Regnery, S., Bauer, L., Katsigiannopulos, E., Norajitra, T., et al.: Voxtell: Free-text promptable universal 3d medical image segmentation. arXiv preprint arXiv:2511.11450 (2025)
40. Schmidgall, S., Su, Y., Wang, Z., Sun, X., Wu, J., Yu, X., Liu, J., Moor, M., Liu, Z., Barsoum, E.: Agent laboratory: Using llm agents as research assistants. Findings of the Association for Computational Linguistics: EMNLP 2025 pp. 5977–6043 (2025)
41. Su, H., Chen, R., Tang, S., Yin, Z., Zheng, X., Li, J., Qi, B., Wu, Q., Li, H., Ouyang, W., et al.: Many heads are better than one: Improved scientific idea generation by a llm-based multi-agent system. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 28201–28240 (2025)
42. Swanson, K., Wu, W., Bulaong, N.L., Pak, J.E., Zou, J.: The virtual lab of ai agents designs new sars-cov-2 nanobodies. Nature **646**(8085), 716–723 (2025)
43. Tang, X., Zou, A., Zhang, Z., Li, Z., Zhao, Y., Zhang, X., Cohan, A., Gerstein, M.: Medagents: Large language models as collaborators for zero-shot medical reasoning. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 599–621 (2024)
44. Wang, X., Chen, Y., Yuan, L., Zhang, Y., Li, Y., Peng, H., Ji, H.: Executable code actions elicit better llm agents. In: Forty-first International Conference on Machine Learning (2024)
45. Wang, Z., Wu, J., Cai, L., Low, C.H., Yang, X., Li, Q., Jin, Y.: Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. arXiv preprint arXiv:2503.18968 (2025)
46. Wei, J., Yang, Y., Zhang, X., Chen, Y., Zhuang, X., Gao, Z., Zhou, D., Wang, G., Gao, Z., Cao, J., et al.: From ai for science to agentic science: A survey on autonomous scientific discovery. arXiv preprint arXiv:2508.14111 (2025)
47. Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., et al.: Autogen: Enabling next-gen llm applications via multi-agent conversations. In: First conference on language modeling (2024)
48. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)

49. Zhang, H.G., Eckmann, P., Miao, J., Mahon, A.B., Zou, J.: The virtual biotech: A multi-agent ai framework for therapeutic discovery and development. bioRxiv pp. 2026–02 (2026)
50. Zhang, W., Qiao, M., Zang, C., Niederer, S., Matthews, P.M., Bai, W., Kainz, B.: Multi-agent reasoning for cardiovascular imaging phenotype analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 429–439. Springer (2025)
51. Zhao, T., Gu, Y., Yang, J., Usuyama, N., Lee, H.H., Naumann, T., Gao, J., Crabtree, A., Abel, J., Moung-Wen, C., et al.: Biomedparse: a biomedical foundation model for image parsing of everything everywhere all at once. arXiv preprint arXiv:2405.12971 (2024)
52. Zhao, Z., Zhang, Y., Wu, C., Zhang, X., Zhou, X., Zhang, Y., Wang, Y., Xie, W.: Large-vocabulary segmentation for medical images with text prompts. NPJ Digital Medicine **8**(1), 566 (2025)
53. Zhu, J., Hamdi, A., Qi, Y., Jin, Y., Wu, J.: Medical sam 2: Segment medical images as video via segment anything model 2. arXiv preprint arXiv:2408.00874 (2024)