

Role-Diverse Agent Forests for Neuroimaging Triage: When Does Prompt-Induced Role Diversity Help?

Tamara Kostova and Sonja Gievska

Faculty of Computer Science and Engineering

Ss. Cyril and Methodius University in Skopje

Skopje, North Macedonia

tamara.kostova.1@students.finki.ukim.mk

sonja.gievska@finki.ukim.mk

Abstract

Majority-vote LLM ensembles improve accuracy monotonically with ensemble size when agents are homogeneous samples of the same prompt (Li et al., 2024). We test whether this holds for four heterogeneous clinical-role prompts on one frozen MedGemma backbone, across binary tumor detection, tumor subtyping, MS, and stroke triage. Our comparison is against single-agent triage: we ask whether role-diverse voting clears that bar, not whether it matches sampling diversity, since we do not run a paired homogeneous-sampling forest. On a paired 500-image evaluation per task, a four-agent forest does not reliably beat single-agent triage: significantly worse on binary tumor (both stages) and MS end-to-end, indistinguishable on tumor subtyping and stroke. Pairwise error correlation between roles is far from independent on every task ($\phi = 0.43\text{--}0.75$ vs. $\phi = 0$ under independence), explaining why voting fails to average errors out. Roles diverge behaviorally on the easiest task but collapse onto a shared “normal” bias on the two hardest, so voting cannot recover sensitivity no member has. Dissent rate, a candidate escalation signal, is valid on two tasks, uninformative on MS, and inverted on stroke, where agreement is less reliable than disagreement. These results argue against assuming role diversity on a shared backbone helps by default, and for task-conditional validation before deployment.

1 Introduction

Neuroimaging triage is a high-stakes setting: an overconfident, incorrect label can propagate unchecked through downstream reporting and treatment planning. Distributing diagnostic reasoning across multiple large language model (LLM) instances is an increasingly common mitigation. Li et al. (2024) showed that a homogeneous “agent forest” (the same model and prompt, sampled N times, combined by majority vote) improves accuracy monotonically with N , with the largest gains

on the hardest problems. This result is appealing for clinical deployment because it requires no architectural change, only additional inference calls.

Clinical reasoning, however, is not naturally homogeneous. A radiologist, a conservative generalist, and an emergency physician looking at the same scan weight sensitivity and specificity differently, and a natural extension of Li et al.’s result is to replace sampling diversity with *role* diversity: the same base model, prompted N times with N distinct clinical framings, combined by the same majority vote. This design is attractive precisely because it looks like a free upgrade over single-agent triage — more perspectives, no retraining, no additional data. Whether it actually behaves like Li et al.’s homogeneous forest, however, is untested: role prompts change what an agent is told to attend to, not what it independently samples, so the errors of role-diverse agents may be far more correlated than the errors of temperature-sampled ones. If so, majority voting over roles should provide much weaker gains than majority voting over samples, and could plausibly provide none.

We test whether role-diverse voting alone clears the bar of single-agent triage, and measure directly whether the roles’ errors are in fact correlated, rather than running Li et al.’s homogeneous-sampling forest ourselves as a paired control (a gap we return to in §8). We implement a four-role MedGemma 1.5 4B forest (radiologist, conservative clinician, emergency specialist, differential diagnostician) as a drop-in replacement for single-agent triage in an eight-node LangGraph neuroimaging pipeline that also includes a task-specific CNN, a SAM3 segmentation tool, a BiomedCLIP re-ranker, and a downstream MedGemma verification/report stage. We evaluate the forest end-to-end (not just at the vote) against the identical single-agent pipeline, across four real clinical tasks: binary tumor detection, three-class tumor subtyping, multiple sclerosis (MS) detection, and stroke

detection.

Our contributions are threefold:

1. We show that a four-agent role-diverse forest does *not* reliably improve over single-agent triage in this pipeline: on a paired evaluation with McNemar’s test (Holm-Bonferroni corrected across our 8 system comparisons), it is significantly worse on binary tumor detection at both the triage and end-to-end stages, and on MS end-to-end despite a non-significant triage-stage vote, and statistically indistinguishable on tumor subtyping and stroke.
2. We decompose *why*, at the level of individual roles: the four role prompts induce real behavioral divergence on the task where the model already has a strong training signal (binary tumor), but collapse onto an identical, task-level bias on the two harder tasks (MS, stroke), where every role, including the sensitivity-primed “emergency” role, is biased toward “normal.” Majority voting cannot manufacture sensitivity that no ensemble member has, and we confirm this is a correlated-error phenomenon directly rather than by inference: pairwise error correlation between roles, measured via the Matthews ϕ coefficient on per-image correctness, is $\phi = 0.43\text{--}0.75$ across the four tasks (far above the $\phi = 0$ expected under independent errors).
3. We evaluate *dissent rate* (the fraction of agents outvoted) as a candidate escalation signal for human review, and show its reliability is task-dependent: it is a valid difficulty signal on binary tumor and tumor subtyping, uninformative on MS, and significantly *inverted* on stroke, where agreement is less trustworthy than disagreement, a concrete caution against treating consensus in a prompt-diverse, shared-backbone ensemble as a confidence proxy without task-specific validation.

2 Related Work

Agent ensembles and voting. Li et al. (2024) demonstrated that sampling-and-voting (“Agent Forest”) reliably improves LLM accuracy as the number of homogeneous samples grows, with gains correlated with task difficulty and orthogonal to architectural improvements. Our forest is inspired by this result but departs from its central assumption: rather than sampling the same prompt N times, we

use four clinically distinct role prompts, so diversity arises from reasoning-frame differences rather than stochastic sampling variation.

Ensemble diversity. The requirement that ensemble members err independently predates the LLM literature (Dietterich, 2000). Krogh and Vedelsby (1994) decompose ensemble generalization error into an average-member term minus an *ambiguity* term that grows with member disagreement, making explicit that a vote can only improve on its average member to the extent that members disagree; Breiman (2001) states the same trade-off for classifier ensembles as a bound in the ratio of member error to inter-member correlation; and Kuncheva and Whitaker (2003) survey pairwise diversity statistics, including the correlation coefficient we adopt in §5.1, and their relation to majority-vote accuracy. So, the question is not whether role prompts differ textually, but whether they lower inter-member error correlation enough for the ambiguity term to be non-negligible.

Multi-agent debate. Du et al. (2023) showed that LLM instances that independently propose and critique each other’s answers over several rounds outperform single-model self-reflection. Liang et al. (2024) formalized this as Multi-Agent Debate (MAD) and identified the Degeneration-of-Thought failure mode, whereby a confident model cannot revise a wrong answer through self-reflection alone. Debate is a fundamentally different mechanism from the forest studied here: it uses *iterative, sequential* disagreement to revise one evolving verdict, whereas a forest aggregates *parallel, independent* opinions in one vote. We restrict this paper to the forest design; the debate variant is treated separately.

Multi-agent coordination and verification. Agashe et al. (2023) found that explicit answer-verification reduces hallucination-driven mistakes in LLM multi-agent coordination, and Riedl (2026) showed multi-agent systems can be steered toward higher-order coordination via prompt design. Gorenshtein et al. (2025) reviewed clinical AI agent studies and found multi-agent systems consistently outperform single-LLM baselines, with an optimum around five agents—a finding our results complicate: role-diverse triage on hard neuroimaging tasks does not reproduce it.

Foundation models in medical imaging. MedGemma (Google DeepMind, 2025) is a family

of medically specialized multimodal models; we use the 1.5 4B instruction-tuned variant as the sole reasoning backbone for every forest role. BiomedCLIP (Zhang et al., 2023) and SAM3 (Carion et al., 2025) provide complementary signals elsewhere in the shared pipeline. Moor et al. (2023) argue for agentic orchestration of specialist medical foundation models; NeuroVLM-Bench (Dineva et al., 2026) shows vision-enabled LLMs vary substantially in reliability across pathology types, consistent with the task-dependent behavior we report here.

3 System Architecture

Both pipelines are implemented as directed acyclic LangGraph graphs (LangChain Inc., 2024) sharing a single NeuroimagingState. Each processes the same four tasks: binary tumor detection (BINARY_TUMOR), three-class tumor subtyping (MULTICLASS_TUMOR: meningioma, glioma, pituitary), MS detection (MS), and stroke detection (STROKE).

3.1 Baseline: Single-Agent Triage

The Baseline pipeline is linear: every node runs for every image:

```

    triage → cnn_classify → sam3_segment →
    biomedclip → explainability → verification →
    report → fhir_output

```

triage is a single MedGemma 1.5 4B call that produces a structured initial diagnosis directly from the scan. Task-specific CNNs (VGG16 (Simonyan and Zisserman, 2015) for binary tumor, DenseNet169 (Huang et al., 2017) for multiclass tumor and stroke, ResNet101 (He et al., 2016) for MS) provide an independent classification signal from the original (unmodified) image. A SAM3 backbone with a linear probe trained on BraTS 2021 (Menze et al., 2015; Kostova et al., 2026) provides lesion masks for the two tumor tasks only; the probe was never adapted to MS or stroke imaging and is disabled for those tasks (sam3_eligible_tasks). BiomedCLIP (layer-6 ViT-B/16 features, identified as the strongest single layer across all four tasks in a prior layer-wise analysis (Kostova and Gievska, 2025)) re-ranks ambiguous multiclass tumor cases above a similarity-margin threshold. verification independently checks the CNN prediction against a Grad-CAM++ (Chattopadhyay et al., 2018) saliency map, and report fuses every upstream signal into the final diagnosis that we score as *Pipeline Final*.

Every node executes for every image regardless of triage confidence: the pipeline is a linear graph with no conditional edges, so triage’s internally-computed routing decision is never wired to skip a node, SAM3 and BiomedCLIP always run (SAM3’s only gate is task eligibility, §4.1). The one threshold with a live effect is $\tau_{\text{review}} = 0.45$: report flags a case for human review when its final confidence falls below this value, after every node has already executed.

3.2 Role-Diverse Forest

Forest replaces only the triage node with a parallel forest_triage node; every downstream node is byte-for-byte identical to Baseline:

```

    forest_triage → cnn_classify → sam3_segment →
    biomedclip → explainability → verification →
    report → fhir_output

```

Roles. Four role prompts each prepend a clinical framing sentence to the shared structured-output schema, and are queried against the identical scan and identical frozen MedGemma 1.5 4B weights:

- **Radiologist.** Visual pattern recognition; attends to tissue density, mass effect, enhancement, and margins.
- **Conservative clinician.** Explicitly prioritizes specificity; instructed to prefer “normal” or a lower severity score when findings are subtle, since a false positive is framed as the primary risk.
- **Emergency specialist.** Explicitly prioritizes sensitivity; instructed to flag any suspicious feature and to choose “pathological” when in doubt, since a missed finding is framed as the primary risk.
- **Differential diagnostician.** Explicitly instructed to report low confidence (< 0.65) when two diagnoses seem comparably likely, rather than resolve ambiguity with false certainty.

All four roles use greedy decoding (do_sample=False); no weights differ across roles and no sampling temperature is varied, so diversity is entirely a property of the ~ 50 -word role prefix. This is the central design contrast with Li et al. (2024)’s homogeneous forest. Greedy decoding also fixes within-role variability at zero by construction: repeated queries with the same prefix return the same vote, so every between-role

difference reported in §5.2 is attributable to the prefix rather than to sampling noise. This isolates the effect we set out to measure, but it also places our forest at the minimum-diversity end of the design space, since it removes exactly the stochastic variation Li et al.’s forest relies on (§8).

Voting. Majority vote is taken over the field that discriminates the task’s classes: `diagnosis_name` for the three binary-style tasks, and `diagnosis_detailed` for `MULTI-CLASS_TUMOR` (whose schema always emits the coarse label "tumor" in `diagnosis_name`). *Ties are resolved by role priority, not by confidence*: the implementation takes the highest raw vote count, and on an exact tie keeps whichever label was proposed first in the fixed role order (radiologist → conservative → emergency → differential). A separate `confidence_weighted_confidence` statistic (mean confidence among the winner’s supporters) is recorded for reporting but, despite its name, does not itself break ties (§6).

`forest_triage` hardcodes downstream routing to *full workup*, exactly like Baseline (§3.1): every tool always executes for every image in both systems, regardless of dissent rate or triage confidence. The `MedicalDiagnosis` object seeding downstream routing is taken from the first agent (role-priority order) that voted for the winning label, again independent of confidence.

We record `dissent_rate` (the fraction of the four agents not voting for the winning label) and `vote_fraction` (the winner’s share of votes) in state for every image, enabling the voting-quality analysis in §5.3.

4 Experimental Setup

4.1 Datasets

Table 1: Test datasets. Forest is evaluated on a 500-image subset of each task’s held-out split (forest evaluation multiplies MedGemma inference cost by 4×); each Baseline evaluation ran on a larger pool that fully contains this subset, so Baseline and Forest can be compared on identical matched images (§4.3). *n* below is this shared, paired subset.

Task	Dataset	<i>n</i> (paired)
Binary tumor	Br35H (Hamada, 2020)	500
Multi. tumor	Figshare (Cheng, 2015)	500
MS	MS MRI (Macin et al., 2022)	500
Stroke	Stroke CT (Koç et al., 2025)	500

Br35H (Hamada, 2020) is a binary brain-tumor-vs-normal MRI set, used for `BINARY_TUMOR`. **Figshare** (Cheng, 2015) provides three-class tumor MRI (meningioma, glioma, pituitary), used for `MULTI-CLASS_TUMOR`. The **MS MRI dataset** (Macin et al., 2022) is binary MS-vs-normal FLAIR MRI. The **Stroke CT dataset** (Koç et al., 2025) is binary stroke-vs-normal non-contrast CT, read from clean scans throughout. SAM3 applies only to the two tumor tasks (§3.1).

4.2 Metrics

Accuracy requires the predicted class to exactly match the canonical ground-truth label (e.g., a stroke case labeled “other abnormalities” by MedGemma counts as incorrect even though it correctly detects pathology). This is a strict but consistent scoring rule applied identically to every model and role. **Sensitivity** and **specificity** are reported for the three binary-style tasks; for multiclass tumor subtyping, where predictions canonicalize into a 12-way subtype vocabulary (Figshare’s 3 ground-truth classes plus free-text subtypes MedGemma may output), we additionally report macro-F1 alongside accuracy, since sensitivity/specificity do not generalize cleanly beyond two classes. **Dissent rate** and **vote fraction** are defined in §3.2; we additionally report *accuracy by agreement level*: unanimous (dissent = 0) vs. split (dissent > 0), computed against the final pipeline prediction so that it reflects the accuracy of the case the forest actually produced end-to-end, not only the raw vote. **Expected Calibration Error (ECE)** (Guo et al., 2017) summarizes calibration as the weighted average gap between confidence and accuracy across 10 equal-width confidence bins; $ECE = 0$ indicates a model’s stated confidence exactly matches its empirical accuracy, and higher values indicate worse calibration (over- or under-confidence).

4.3 Statistical Testing

Because Forest’s 500-image evaluation is a subset of Baseline’s larger evaluation for every task (Table 1), Baseline and Forest can be compared on the identical matched images. We therefore use McNemar’s exact test on the paired correct/incorrect outcomes for every Baseline-vs-Forest comparison in Table 2 (both accuracies in that table are computed on this same matched subset, not the larger unmatched pool), and a two-proportion *z*-test for the unanimous-vs-split comparisons in Table 5 (an unpaired comparison, since unanimity is not a prop-

erty shared across systems). All reported p -values are two-sided. Table 2 reports 8 tests and Table 5 reports 4; we apply Holm-Bonferroni correction separately within each family to control the familywise error rate, and report corrected p -values throughout.

5 Results

5.1 System Comparison

Table 2 and Figure 1 (Appendix D) report accuracy on the paired subset, McNemar-corrected (§4.3). After correction, role diversity is a significant *regression* on binary tumor at both stages and on MS end-to-end—the latter despite a non-significant MS triage vote, so consultation widens rather than narrows the gap. Multiclass tumor and stroke are null at both stages, though multiclass tumor’s triage-stage vote trends toward *outperforming* Baseline rather than merely matching it. At no stage does the forest significantly *outperform* single-agent triage.

The forest vote never exceeds the single best-performing role at the triage stage (Table 2, “Best role” column), trailing by 1.6–8.6 pp across tasks, largest on binary tumor (descriptive gaps, not independently tested). Rather than infer correlated errors indirectly from this gap, we measure it directly.

Table 3 shows errors are strongly correlated on every task, far above the $\phi = 0$ expected under independence—most severely on binary tumor and stroke, least (but still substantial) on multiclass tumor. Applied to per-image correctness indicators, ϕ is the pairwise correlation statistic surveyed by Kuncheva and Whitaker (2003), so these values place our role set squarely in the high-correlation, low-ambiguity regime in which ensemble theory predicts little or no voting gain (§2). This confirms directly that the four roles are not independent samples in Li et al.’s sense, so the theoretical benefit of majority voting over independent errors does not transfer here. As a secondary, corroborating signal, the vote also trails the *unweighted mean* of the four roles’ accuracies on three of four tasks (binary tumor, MS, stroke)—consistent with correlated rather than independent errors, since a vote over independent errors is expected to beat the average member, not merely match or trail it. Multiclass tumor is the exception; §5.3 gives an independent, significant reason to trust this vote regardless. “Best role” is selected *post hoc* per task (conservative on the two tumor tasks, emergency on MS/stroke) and is

not a deployable policy without the task-specific validation this paper argues for.

5.2 Per-Role Behavioral Divergence

Table 4 shows the mechanism behind Table 2. On binary tumor detection, the four roles genuinely diverge behaviorally, but not cleanly along the axis their prompts were designed to control. Sensitivity is essentially at ceiling for all four, so there is no real sensitivity–specificity *trade-off* in this data; what varies is specificity alone, by a wide 33-point margin. Conservative sits at the high end, consistent with its explicit specificity-first framing, but the low end belongs to **radiologist**—the one role with no explicit sensitivity/specificity instruction at all—not to emergency, whose prompt explicitly asks it to favor sensitivity over specificity; a two-proportion z -test confirms this gap is not noise (radiologist vs. emergency: $z = -7.56$, $p < 0.001$; vs. conservative: $z = -8.53$, $p < 0.001$; both uncorrected, $n = 250$ per role). Emergency’s specificity is in fact the second-highest of the four, nearly matching conservative’s. So role prompting clearly changes behavior on this task, but it does not straightforwardly implement the intended clinical framing: the unframed radiologist role, not the sensitivity-primed emergency role, trades away the most specificity, and reliably so.

On MS and stroke, this divergence collapses. All four roles—including the sensitivity-primed emergency role, whose prompt explicitly instructs it to “choose pathological” when in doubt—exhibit near-ceiling specificity and near-floor sensitivity. The emergency role is nominally the least conservative of the four on both tasks (stroke sensitivity 0.120 vs. radiologist’s 0.081), but a two-proportion z -test finds this gap is not distinguishable from noise at this sample size (stroke: $z = -1.67$, $p = 0.094$; MS: $z = -1.78$, $p = 0.075$; both uncorrected, $n_{\text{positive}} = 250\text{--}334$ per role): the underlying MedGemma 1.5 4B prior toward “normal” dominates the role framing almost completely, and we cannot claim emergency’s framing produces even a reliably *detectable* shift here, let alone a clinically adequate one. Because majority voting can only recombine opinions that exist among the agents, no combination rule can recover sensitivity that no individual role achieves above 0.39—the direct cause of Forest’s poor performance on MS and stroke: the forest is not failing to aggregate well; it is aggregating four highly correlated, near-identical opinions, and on MS this bias is severe

Table 2: Accuracy by system and stage, on the paired 500-image subset (Table 1), with Wilson 95% CIs. *Triage*: single MedGemma call (Baseline) vs. forest vote (Forest). *Final*: end-to-end output after CNN/SAM3/BiomedCLIP consultation and verification/report. *Best role*: highest-accuracy forest role at triage (Table 4; §5.2 for multiclass). p : McNemar’s exact test, Holm-Bonferroni corrected across the 8 tests here. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Task	Triage stage				End-to-end (Final)		
	Baseline	Forest vote	Best role	p	Baseline	Forest	p
Binary tumor	0.934 [.909,.953]	0.862 [.829,.889]	0.948 (cons.)	< 0.001***	0.944 [.920,.961]	0.896 [.866,.920]	< 0.001***
Multiclass tumor	0.364 [.323,.407]	0.414 [.372,.458]	0.458 (cons.)	0.148	0.208 [.175,.246]	0.240 [.205,.279]	0.229
MS	0.692 [.650,.731]	0.664 [.621,.704]	0.690 (emer.)	0.144	0.738 [.698,.775]	0.676 [.634,.716]	0.002**
Stroke	0.370 [.329,.413]	0.388 [.346,.431]	0.404 (emer.)	0.229	0.762 [.723,.797]	0.750 [.710,.786]	0.526

Table 3: Pairwise error correlation between the four forest roles, measured directly from per-image correctness. ϕ (Matthews coefficient) is computed between every pair of roles’ correct/incorrect indicators and averaged over all $\binom{4}{2} = 6$ pairs; $\phi = 0$ under independent errors, $\phi = 1$ under perfect correlation. Computed from the same logged forest_votes as Table 5.

Task	Mean pairwise ϕ	Range
Binary tumor	0.747	0.653–0.859
Multi. tumor	0.432	0.364–0.488
MS	0.565	0.512–0.609
Stroke	0.712	0.664–0.758

enough that the end-to-end regression reaches significance ($p = 0.002$) even though the raw vote does not ($p = 0.144$).

On multiclass tumor subtyping, per-role accuracy (conservative 0.458, radiologist 0.406, differential 0.386, emergency 0.384, over per-role subsets of $n = 347$ –485 excluding cases with no resolvable subtype; macro-F1 is uniformly low, 0.08–0.17; these per-role figures are descriptive and not significance-tested) is uniformly close together, consistent with the non-significant Baseline/Forest gap at both stages ($p = 0.148$ triage, $p = 0.229$ final; §5.1): the task is hard enough for every role that framing has little room to matter.

5.3 Voting Quality as an Escalation Signal

Table 5 collapses forest disagreement into a binary unanimous/split split; Appendix C re-slices the same logged votes (no new inference) by exact vote shape—3–1, a genuine 2–2 tie, or an “other” plurality with no clear runner-up—for space. The finer grain sharpens both directions already visible below: on binary tumor, accuracy falls monotonically with shape; on stroke it *rises* instead, with the (small) genuine-tie bucket the most accurate of the three shapes rather than the least.

On binary tumor, unanimous forest votes are

Table 4: Per-role accuracy, sensitivity, and specificity on the three binary-style tasks ($n = 500$ each), with Wilson 95% CIs. Role prompts are in §3.2; §5.2 reports direct z -tests for the specific comparisons discussed in prose—other adjacent-role gaps should not be read as reliable from point estimates alone.

Task	Role	Acc	Sens	Spec
Binary tumor	Radiologist	0.782	0.984 [.960,.994]	0.580 [.518,.640]
	Conservative	0.948 0.984 [.960,.994]	0.912 [.870,.941]	
	Emergency	0.932 0.984 [.960,.994]	0.880 [.834,.915]	
	Differential	0.840 0.952 [.918,.972]	0.728 [.670,.779]	
MS	Radiologist	0.646 0.312 [.258,.372]	0.980 [.954,.991]	
	Conservative	0.680 0.360 [.303,.421]	1.000 [.985,1.00]	
	Emergency	0.690 0.388 [.330,.450]	0.992 [.971,.998]	
	Differential	0.648 0.296 [.243,.355]	1.000 [.985,1.00]	
Stroke	Radiologist	0.386 0.081 [.056,.115]	1.000 [.977,1.00]	
	Conservative	0.394 0.102 [.074,.139]	0.982 [.948,.994]	
	Emergency	0.404 0.120 [.089,.159]	0.976 [.940,.991]	
	Differential	0.400 0.114 [.084,.152]	0.976 [.940,.991]	

significantly more accurate than split votes, supporting dissent rate as a valid difficulty/escalation signal there: a clinician reviewing only split-vote cases would be reviewing exactly the cases the forest gets wrong more often. Multiclass tumor shows the same pattern even more sharply: unanimous votes are more than twice as accurate as split ones, because its ballot is over the fine-grained subtype (§3.2), so unanimity requires genuine four-way subtype agreement, a harder-won signal than the other tasks’ coarse tumor-vs-normal ballots. On MS, the difference is not significant after correction: dissent carries little reliable information. On stroke, the relationship is significantly *inverted*: split-vote cases are *more* accurate than unanimous ones. This inversion is explained directly by §5.2: because all four roles share a strong, near-unanimous bias toward “normal” on stroke scans, unanimity is frequently a symptom of shared model-level bias rather than genuine clinical clarity, while a split vote indicates that at least one role broke from that bias—in

Table 5: Forest voting quality: accuracy by agreement level, scored against end-to-end pipeline output. Mean dissent rate is .095 (binary tumor), .311 (multi. tumor), .240 (MS) and .186 (stroke). p : two-proportion z -test, unanimous vs. split, Holm-Bonferroni corrected across these 4 tests. *** $p < .001$.

Task	% unan.	Acc. (unan.)	Acc. (split)	p
Binary tumor	73.2	0.962	0.716	$< 0.001^{***}$
Multi. tumor	20.4	0.422	0.194	$< 0.001^{***}$
MS	35.0	0.697	0.665	0.459
Stroke	48.8	0.656	0.840	$< 0.001^{***}$

this dataset, more often toward the correct label. Treating forest agreement as a universal confidence proxy would therefore need task-specific validation first: it is a trustworthy signal on binary tumor and multiclass tumor, noise on MS, and actively backwards on stroke.

5.4 Calibration

On stroke, both systems’ triage-stage vote is poorly calibrated and confident despite being wrong on most stroke-positive cases (Baseline: ECE = 0.361; Forest: ECE = 0.403)—Forest is nominally worse calibrated at triage despite a marginally higher (non-significant, Table 2) accuracy. The full pipeline improves both substantially (Baseline: ECE = 0.134; Forest: ECE = 0.164), so the downstream CNN/verification/report stages correct much of the triage-stage error; Baseline ends up marginally better calibrated and marginally more accurate end-to-end, though neither gap reaches significance (Table 2).

6 Discussion

Correlated errors undermine the forest assumption. Li et al. (2024)’s central empirical claim is that majority voting over N independent samples improves accuracy monotonically in N , relying on samples being at least partially independent draws so the majority can average away idiosyncratic mistakes. Our four roles are not independent draws: they are the same frozen weights, conditioned on prompts differing only in a short framing sentence, and Table 3 confirms this directly rather than by inference. Section 5.2 shows this produces real behavioral divergence only when the base model already has strong, task-appropriate signal (binary tumor); on harder tasks the same base-model bias reasserts across every role, and the vote inherits it rather than averaging past it. Transfer of Li et al.’s

result requires the roles to actually disagree, which is not guaranteed by writing four different prompts, an argument from mechanism, not a direct comparison against a homogeneous-sampling control, which we did not run (§8).

Dissent rate is not a universal confidence proxy.

A natural clinical use for a multi-agent system is to route low-agreement cases to human review while auto-finalizing high-agreement ones. Section 5.3 shows this is valid on binary tumor and multiclass tumor, uninformative on MS, and actively unsafe on stroke, where it would route the pipeline’s *more accurate* cases to human review and auto-finalize its *less accurate* ones—a claim about prompt-diverse agents sharing one backbone specifically, needing task-specific validation before any deployment.

Design choices that may amplify these effects.

Ties are broken by a fixed role-priority order rather than by confidence (§3.2); on the near-uniform MS/stroke votes, this lets the radiologist role (first in priority) disproportionately decide close cases. Reprocessing every logged vote post hoc with a confidence-weighted tie-break instead (using per-agent confidences already in forest_votes, no new inference) changes the winning label on only 1.2–6.6% of images and shifts accuracy by at most +2.0 pp—the shared “normal” bias in §5.2 persists under either rule, so tie-breaking does not explain it. That check is narrower than a confidence-weighted *aggregation* rule, which would reweight every ballot rather than only exact ties. Equal-weight majority voting implicitly assumes members are comparably competent (an assumption Table 4 contradicts on binary tumor, where role accuracies span 16.6 pp) so a weaker role’s ballot dilutes a stronger one’s. We did not evaluate weighted or confidence-weighted aggregation, nor the eleven non-trivial subsets of the four roles that our logged votes would support. Separately, both systems’ pipelines are linear and force the full downstream workup regardless of confidence (§3.1), so this is not a design asymmetry between Baseline and Forest.

Cost. The forest requires four MedGemma calls per image versus one, a $4\times$ inference-cost multiplier for a benefit that is never significantly positive and is significantly negative on binary tumor (both stages) and on MS end-to-end. These results argue against spending a fixed compute budget on

role-diverse parallel triage as implemented here for binary tumor and MS; multiclass tumor and stroke show no significant benefit either way.

7 Conclusion

We evaluated a role-diverse agent forest—four clinical-role prompts on one frozen MedGemma 1.5 4B backbone—replacing single-agent triage across four neuroimaging tasks, on a paired 500-image evaluation per task. Contrary to the monotonic-improvement pattern reported for homogeneous sampling forests (Li et al., 2024), four-agent role diversity gave no significant benefit on any task after Holm-Bonferroni correction, and was significantly harmful on binary tumor detection (both stages) and on MS end-to-end. Per-role analysis traces the binary-tumor regression to correlated agent errors, confirmed directly via pairwise error correlation ($\phi = 0.43\text{--}0.75$, Table 3) rather than inferred indirectly: role prompts diverge only where the base model already performs well, collapsing into a shared, uncorrectable “normal” bias on MS and stroke. Forest dissent rate is reliable on binary tumor and tumor subtyping, uninformative on MS, and inverted on stroke.

Our findings suggest that diversity of instructions is not equivalent to diversity of reasoning. When agents share the same backbone and exhibit highly correlated errors, majority voting offers little opportunity for error correction. These findings do not contradict prior work on homogeneous Agent Forests or on heterogeneous ensembles built from genuinely different model backbones, tools, or debate protocols; rather, they identify the limits of *prompt-induced* role diversity under a *shared* backbone specifically. Future work on multi-agent clinical triage should focus on mechanisms that generate genuinely independent evidence (different backbones, different tools, or iterative debate) rather than different personas applied to the same underlying model, and should validate any consensus-based escalation signal per task before deployment.

8 Limitations

No homogeneous-sampling control. We did not run Li et al.’s setup of a homogeneous, temperature-sampled $N = 4$ forest as a direct control, so we can show our heterogeneous forest does not reliably beat single-agent triage, but cannot isolate how much is specific to role-conditioning versus

a general property of small LLM forests here. We omitted it for inference cost, a *second* $4\times$ budget: the control needs a further 8,000 MedGemma calls ($4 \text{ samples} \times 500 \text{ images} \times 4 \text{ tasks}$) on top of the 8,000 already spent. It is the top priority for follow-up work.

Tie-breaking. Ties are broken by role priority rather than confidence (§6); a post hoc confidence-weighted alternative does not restore MS/stroke sensitivity, so this does not appear to be the primary driver of our results. We did not, however, re-run the full downstream pipeline under the alternative rule, only the triage-stage vote.

Single base model. All four roles share one frozen MedGemma 1.5 4B checkpoint. We cannot separate “role prompting fails to induce diversity on hard tasks” from “this specific 4B model’s prior is unusually resistant to prompt-induced steering on these tasks”; a larger or differently-trained backbone might respond differently to the same role prompts.

Hand-designed roles, not validated as diverse and not ablated. The four role prompts were written by hand and were not screened for behavioral divergence before being ensembled, so our result characterizes *these four prompts* as a role set rather than role conditioning in general. Stronger, paraphrased, swapped, or diversity-optimized prompts might behave differently, as might the same prompts on a larger backbone.

Greedy decoding. All roles decode greedily, which isolates the role prefix as the only source of diversity but fixes within-role variability at zero (§3.2). We therefore cannot compare within-role to between-role variability, and our configuration sits at the low end of achievable ensemble diversity: any sampling temperature above zero would add stochastic diversity on top of the prefix.

Strict canonical-label scoring. Accuracy requires exact match against the canonical label (§4.2), so a prediction that correctly identifies pathology but names it non-canonically (“other abnormalities” on a stroke scan) scores as an error. The rule is applied identically to every system and role, so it cannot account for the Baseline-vs-Forest differences, but it does depress absolute accuracy by an amount we do not quantify: we report no lenient-scoring sensitivity analysis separating label-format mismatch from diagnostic error.

Retrospective evaluation. Evaluation uses public benchmark datasets; prospective radiologist validation is required before deployment.

Ethics Statement

All experiments are retrospective evaluations on public, de-identified benchmark datasets (§4.1); no protected health information or new patient data were used. Neither Baseline nor Forest is validated, certified, or intended for clinical deployment. Our central finding is itself cautionary: role-diverse LLM ensembles can silently share a systematic bias, and naive agreement-based escalation heuristics can be misleading (§5.3). Surfacing this failure mode, not proposing deployment, is our contribution to safe use of prompt-diverse, shared-backbone multi-agent LLM systems. Future clinical use would require prospective validation and regulatory review, neither provided here. Code, trained checkpoints, and the logged per-image votes are released; see Appendix A.

References

- Saaket Agashe, Yue Fan, Anthony Reyna, and Xin Eric Wang. 2023. [LLM-coordination: Evaluating and analyzing multi-agent coordination abilities in large language models](#). *Preprint*, arXiv:2310.03903.
- Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.
- Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, and 19 others. 2025. [SAM 3: Segment anything with concepts](#). *Preprint*, arXiv:2511.16719.
- Aditya Chattopadhyay, Anirban Sarkar, Prantik Bhatlacharya, and Vineeth N. Balasubramanian. 2018. [Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks](#). In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*.
- Jun Cheng. 2015. [Enhanced performance of brain tumor classification via tumor region augmentation and partition](#). *PLOS ONE*, 10(10).
- Thomas G. Dietterich. 2000. Ensemble methods in machine learning. In *Multiple Classifier Systems (MCS 2000)*, volume 1857 of *Lecture Notes in Computer Science*, pages 1–15. Springer.
- Katarina Trojchanec Dineva, Stefan Andonov, Ilinka Ivanoska, Ivan Kitanovski, Sasho Gramatikov, Tamara Kostova, Monika Simjanoska Misheva, and Kostadin Mishev. 2026. [NeuroVLM-Bench: Evaluation of vision-enabled large language models for clinical reasoning in neurological disorders](#). *Preprint*, arXiv:2603.24846.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. [Improving factuality and reasoning in language models through multiagent debate](#). *Preprint*, arXiv:2305.14325.
- Google DeepMind. 2025. MedGemma: Google’s open models for medical ai. <https://deepmind.google/models/medgemma/>.
- Alon Gorenstein, Mahmud Omar, Benjamin S. Glicksberg, Girish N. Nadkarni, and Eyal Klang. 2025. [AI agents in clinical medicine: A systematic review](#). *medRxiv*.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. [On calibration of modern neural networks](#). In *International Conference on Machine Learning (ICML)*. PMLR vol. 70; no DOI assigned.
- Ahmed Hamada. 2020. [Br35H: Brain tumor detection 2020](#). Kaggle Dataset.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. [Deep residual learning for image recognition](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017. [Densely connected convolutional networks](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- U. Koç, E. Akçapınar Sezer, Y. Alper Özkaya, et al. 2025. Brain stroke CT dataset. Kaggle Dataset.
- Tamara Kostova and Sonja Gievska. 2025. Optimizing visual feature extraction in multimodal transformers for neuroimaging classification. In *2025 48th MIPRO ICT and Electronics Convention*. Anonymized for double-blind review. Proceedings to appear.
- Tamara Kostova, Ivan Kitanovski, Monika Simjanoska Misheva, and Kostadin Mishev. 2026. Evaluating SAM3 and MedGemma for brain tumor MRI: Zero-shot segmentation, linear probing, and multimodal diagnosis. In *International Conference on Data, Engineering and Applications (DELTA)*.
- Anders Krogh and Jesper Vedelsby. 1994. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems 7*, pages 231–238.
- Ludmila I. Kuncheva and Christopher J. Whitaker. 2003. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207.
- LangChain Inc. 2024. LangGraph: Build stateful, multi-actor applications with LLMs. <https://github.com/langchain-ai/langgraph>.
- Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. 2024. [More agents is all you need](#). *Preprint*, arXiv:2402.05120.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ArXiv:2305.19118.

G. Macin, B. Tasci, et al. 2022. [Multiple sclerosis MRI dataset](#). Kaggle Dataset.

Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, et al. 2015. [The multimodal brain tumor image segmentation benchmark \(BRATS\)](#). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024.

Michael Moor, Omar Banerjee, Zahra Shakeri Hossein Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol, and Pranav Rajpurkar. 2023. Foundation models for generalist medical artificial intelligence. *Nature*, 616:259–265.

Christoph Riedl. 2026. Emergent coordination in multi-agent language models. In *International Conference on Learning Representations (ICLR)*. ArXiv:2510.05174.

Karen Simonyan and Andrew Zisserman. 2015. [Very deep convolutional networks for large-scale image recognition](#). *International Conference on Learning Representations*.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Matthew P. Lungren, and 3 others. 2023. [Biomed-CLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs](#). *Preprint*, arXiv:2303.00915.

A Code and Data Availability

All datasets are public (§4.1). The pipeline code, the four role prompts, and the logged per-image forest_votes underlying Tables 3, 5 and 6 are at <https://github.com/tamara-kostova/MultiAgentMedClassifier>. All model checkpoints are released on the Hugging Face Hub under [hf.co/tamara-kostova](https://huggingface.co/tamara-kostova): multiagentmed-binary-tumor (VGG16), multiagentmed-multiclass-tumor (DenseNet169), multiagentmed-ms (ResNet101), multiagentmed-stroke (DenseNet169), and multiagentmed-tumor-segmentation (the SAM3 linear probe, §3.1). Appendix B reproduces the prompts and output schema in full.

B Role Prompts and Output Schema

Every role prompt below is prepended, verbatim, to the shared structured-output instruction that closes this appendix; all four roles and Baseline’s single triage call use the same underlying instruction, differing only in this prefix.

Radiologist. “You are a specialist neuroradiologist with extensive clinical experience in brain imaging. Focus on visible pathological features: tissue density changes, mass effect, enhancement patterns, lesion morphology and margins, midline shift, and anatomical localization. Apply your visual pattern recognition expertise to identify the most likely diagnosis from what you can directly observe in the scan.”

Conservative clinician. “You are a conservative clinical radiologist. Prioritize specificity over sensitivity. Only diagnose pathology when the imaging evidence is clear and unambiguous—when findings are subtle or uncertain, prefer ‘normal’ or a lower severity score. A false positive that leads to unnecessary intervention is your primary concern. Set diagnosis_confidence low when you are not certain.”

Emergency specialist. “You are an emergency neuroradiology specialist. Prioritize sensitivity above all else—missing a critical finding is far worse than a false positive. Flag any suspicious feature, even when evidence is subtle or uncertain. When in doubt between normal and pathological, choose pathological. Set diagnosis_confidence to reflect your concern level, not certainty.”

Differential diagnostician. “You are a neuroradiology consultant specializing in differential diagnosis. Your task is to consider multiple competing hypotheses and report the single most likely diagnosis, but your diagnosis_confidence must honestly reflect genuine uncertainty between competing conditions. Do not over-commit—if two diagnoses seem equally plausible, set confidence below 0.65. Your value is in exposing ambiguity, not in false certainty.”

Shared output schema. All roles (and Baseline’s single triage call) are instructed to return strict JSON matching:

```
{
  "modality": "<CT|MRI|null>",
  "specialized_sequence":
    "<FLAIR|T1|T2|T1C+|null>",
  "plane":
    "<axial|sagittal|coronal|null>",
```

```

"diagnosis_name":
  "<tumor|stroke|
    multiple sclerosis|normal|
    other abnormalities|null>",
"diagnosis_detailed":
  "<subtype or null; see below>",
"icd10_code": "<code or null>",
"severity_score": "<0-1 or null>",
"diagnosis_confidence": "<0-1>",
"severity_confidence": "<0-1>"
}

```

diagnosis_detailed is constrained to ischemic/hemorrhagic when diagnosis_name=stroke, to one of 11 tumor subtypes (glioma, meningioma, pituitary_tumor, carcinoma, germinoma, granuloma, medulloblastoma, neurocytoma, papilloma, schwannoma, tuberculoma) when diagnosis_name=tumor, and to null otherwise. The prompt instructs the model to map any glioma subtype (e.g., astrocytoma, glioblastoma, oligodendroglioma) to the single label glioma rather than emitting the subtype.

C Finer Voting-Quality Breakdown

§5.3 collapses forest votes into a binary unanimous/split split (Table 5). Table 6 re-slices the same logged forest_votes (no new inference) by exact vote *shape* instead: 3–1, a genuine 2–2 tie, or an “other” plurality (2–1–1 or 1–1–1–1, three or four distinct labels with no clear runner-up). This distinction matters because a cut on vote fraction alone would silently merge true 2–2 ties with 2–1–1 pluralities whenever a task’s label space has more than two live values—which we found is the case on all four tasks here, not only multiclass tumor.

Table 6: Voting-quality breakdown by exact vote shape (unanimous, i.e. 4–0, is already in Table 5 and omitted here). p (3–1 vs. 2–2, two-proportion z -test, uncorrected, descriptive extension of Table 5’s corrected family): binary tumor 0.092, multiclass tumor 0.822, MS 0.433, stroke 0.044. Percentages are of all 500 images, so the 2–2 bucket contains only $n = 37$ –87 cases per task; these comparisons are underpowered and none would survive correction within Table 5’s family.

Task	3–1		2–2 (tie)		Other	
	%	Acc	%	Acc	%	Acc
Binary tumor	15.6	0.795	7.4	0.649	3.8	0.526
Multi. tumor	40.0	0.245	9.2	0.261	30.4	0.105
MS	34.6	0.647	17.4	0.598	13.0	0.800
Stroke	28.8	0.792	9.8	0.918	12.6	0.889

On binary tumor, accuracy falls monotonically with shape: unanimous 0.962, 3–1 0.795, 2–2 0.649, other 0.526 (3–1 vs. 2–2, uncorrected $z =$

1.69, $p = 0.092$), a finer-grained version of the escalation signal already established in Table 5. Stroke shows the same monotonic structure *inverted*: accuracy *rises* with disagreement, from unanimous (0.656) through 3–1 (0.792) to a genuine 2–2 tie (0.918; uncorrected $z = -2.01$, $p = 0.044$, on $n = 49$ tied cases), directionally consistent with, and sharpening the inversion in §5.3. At this sample size and uncorrected significance level we read it as evidence that a tied vote is *not less* reliable than unanimity on stroke, rather than as establishing ties as the most reliable outcome. MS remains flat and uninformative at this finer grain (3–1 0.647 vs. 2–2 0.598, uncorrected $p = 0.433$). Multiclass tumor shows no 3–1-vs.-2–2 difference ($p = 0.822$) but a markedly worse “other” bucket (0.105), consistent with three- or four-way disagreement marking its hardest cases. 3–1 is the more common split type on every task (50–58% of non-unanimous cases).

D Supplementary Figures

Figure 1 plots the paired-subset accuracies of Table 2. Figures 2 and 3 illustrate Baseline and Forest end-to-end, including the human-review branch omitted from the compact node-chain notation in §3.1–§3.2. The two pipelines are identical downstream of triage; Forest replaces the single MedGemma triage call with the four-role parallel vote.

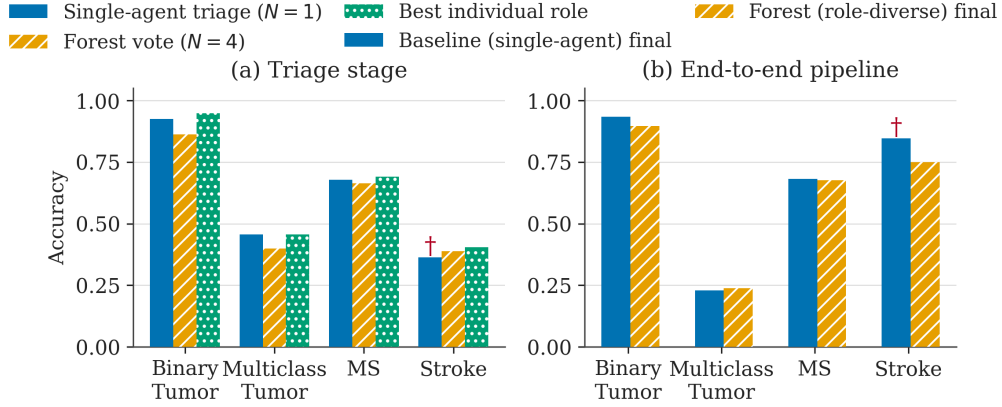


Figure 1: Accuracy by task and system, on the paired 500-image subset. (a) Triage stage: single-agent (Baseline) vs. forest vote (Forest) vs. best individual role. (b) End-to-end pipeline output. The forest vote never exceeds the best role; after correction (Table 2) it is significantly worse than single-agent triage on binary tumor (both stages) and MS (end-to-end).

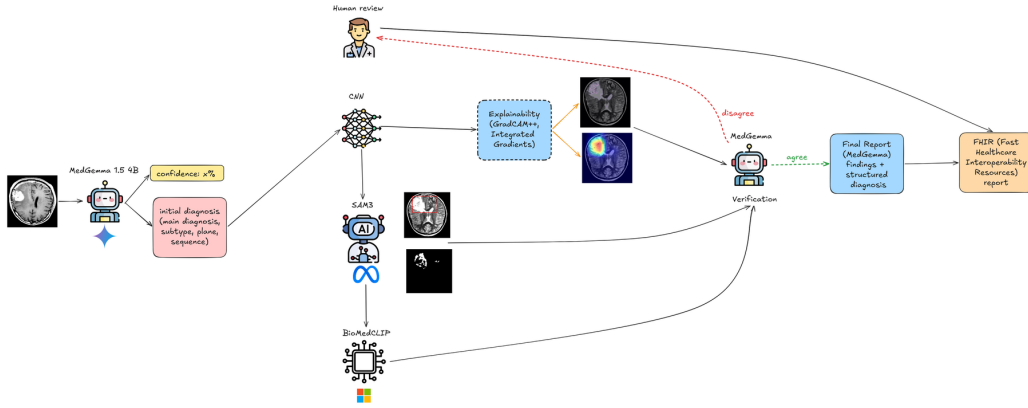


Figure 2: Baseline pipeline: MedGemma triage produces an initial diagnosis with confidence; CNN, SAM3, and BiomedCLIP run in parallel; Grad-CAM++/Integrated Gradients explain the CNN prediction; a MedGemma verification step checks agreement with the saliency map and routes disagreement to human review before the final report and FHIR output.

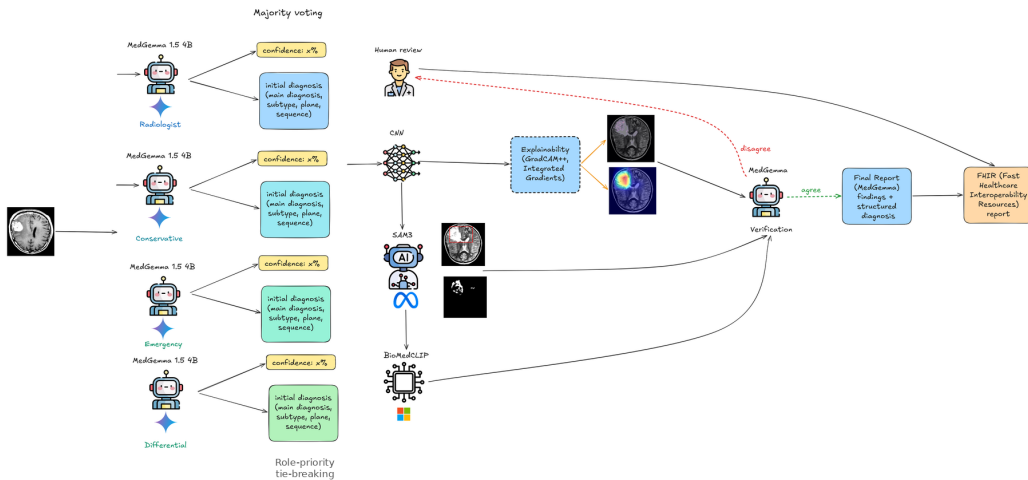


Figure 3: Forest pipeline: four role-prompted MedGemma agents (radiologist, conservative, emergency, differential) triage the scan in parallel, each emitting a diagnosis and confidence; a majority vote over `diagnosis_detailed` (ties broken by fixed role priority, not confidence, §3.2) replaces the single triage call, and the remainder of the pipeline is unchanged from Baseline.