Too Tight or Too Loose: How VLM Persona Committees Miscalibrate Inter-Rater Ambiguity in Medical Image Segmentation

Md Hasan Al Banna[1 [0000−0003−2444−2017]], Md Maklachur Rahman[1 [0000−0002−1984−767X]], and Tracy Hammond[1 [0000−0001−7272−0507]]

Texas A&M University, College Station, TX 77843, USA
{mdhasanalbanna, maklachur, hammond}@tamu.edu

Abstract. Modeling inter-rater ambiguity, the disagreement among radiologists annotating the same image, is central to trustworthy medical image segmentation. The rise of agentic AI suggests a tempting recipe: prompt a vision language model (VLM) with distinct radiologist personas, let each propose a mask, and treat the resulting spread as a model of clinical disagreement. We assessed whether such training free persona committees actually reproduce human inter-rater ambiguity on lung nodule segmentation using the LIDC dataset. Each persona proposes a bounding box that a promptable segmenter (MedSAM) refines into a mask and the per pixel entropy of the committee forms an ambiguity map. We evaluate three backends, a small general VLM (Qwen), a frontier VLM (Claude), and a medical VLM trained on CT (MedGemma), and we stratify disagreement into boundary cases, where all raters mark the lesion, and existence cases, where some raters see nothing. A parameter free geometric null based on distance to the consensus boundary predicts human disagreement better than every committee, reaching AUROC 0.87 versus 0.53 on boundary cases and 0.64 versus 0.63 on existence cases. Committee diversity is also miscalibrated as Claude personas agree much more than radiologists on existence cases, with Dice 0.70 versus 0.39, while MedGemma personas disagree much more on boundary cases, with Dice 0.47 versus 0.79. These results show that off the shelf persona committees do not capture clinical ambiguity and that effective methods require calibrated diversity and explicit abstention.

Keywords: Medical image segmentation · Multi-agent system · Segmentation ambiguity · MedSAM

1 Introduction

Medical image segmentation is rarely deterministic. Expert radiologists routinely disagree on the extent, and even the presence of a finding, and this disagreement is well documented across modalities and anatomies [1, 6, 11]. In lung computed tomography (CT), for example, the four readers of the LIDC/IDRI corpus frequently delineate the same nodule differently, and often one reader marks a lesion

that another does not annotate at all [1]. Treating a single mask as ground truth therefore discards clinically meaningful uncertainty, and a growing body of work argues that a trustworthy segmentation model should instead capture the full distribution of plausible expert annotations [8, 3, 13].

Prior approaches learn this distribution directly from multiple annotations. The Probabilistic U-Net couples a segmentation backbone with a conditional variational autoencoder to sample diverse hypotheses [8], PHiSeg injects stochasticity hierarchically for greater sample diversity [3], and diffusion formulations model the annotation distribution with a single generative process [13]. All require supervised training on sets of expert masks, which are costly to collect and unavailable for most tasks.

The progress of vision language models (VLMs) and agentic systems suggests a tempting shortcut [12, 16, 9]. Rather than training a distributional model, one may prompt a single VLM to adopt several radiologist "personas," let each propose a region of interest, delegate pixel level delineation to a promptable segmenter such as MedSAM [10, 7], and read the spread of the resulting masks as clinical disagreement. The recipe is appealing as it needs no multi rater training data, adapts to new tasks through prompting alone, and mirrors the collective reasoning of a tumor board. It also inherits a known risk. In natural language processing, persona conditioned models capture only a small fraction of annotator variation and are highly sensitive to prompt and model choice, so the disagreement they simulate need not match that of real annotators [14, 5].

In this work we ask a direct question: do persona prompted VLM committees actually reproduce radiologist inter rater ambiguity? Using lung nodule segmentation on LIDC/IDRI [1], we build committees on three backends spanning the current capability spectrum, a small general VLM [2], a frontier general VLM, and a medical VLM trained on lung-CT [15], and we compare their ambiguity against the four radiologist annotations. To make the comparison precise, we separate two qualitatively different kinds of disagreement: *boundary* cases, in which every reader marks the lesion but its extent is contested, and *existence* cases, in which some readers see a lesion and others see nothing. We find that the agentic recipe does not work, and we characterize exactly how it fails.

The main contributions of this work are summarized below:

- Using a parameter free baseline derived from the distance to the consensus contour as a reference, we show that no evaluated VLM committee surpasses it at localizing radiologist disagreement, reaching an AUROC of at most 0.53 on boundary cases against the baseline's 0.87. This exposes that assembling off the shelf agents does not improve on a trivial geometric reference.
- We distinguish boundary ambiguity, where all radiologists identify the lesion but disagree on its extent, from existence ambiguity, where some radiologists provide no annotation. The results show that geometry explains most boundary disagreement, while existence disagreement remains substantially harder and requires an explicit abstention mechanism.
- We show that persona committees fail in different ways across VLM backends. The frontier model produces overly similar predictions, whereas the

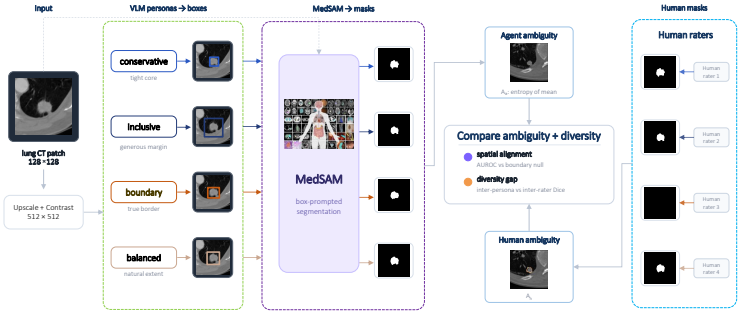

Fig. 1. Overview of the persona committee pipeline. A 128×128 lung CT patch is upsampled and contrast stretched to a legible 512×512 view (affecting only what the agents see, never the annotations). Four persona conditioned VLM agents on a tight to loose axis (*conservative*, *inclusive*, *boundary*, *balanced*) each decide whether a nodule is present and, if so, propose a box, or abstain. MedSAM refines each box into a mask (abstention yields an empty mask, capturing existence disagreement); the mean mask entropy gives the agent ambiguity A_a, and the four radiologist masks give A_h. We compare the two along spatial alignment (AUROC of A_a against a parameter free boundary null) and a diversity gap (inter persona versus inter rater Dice).

medical model generates variation beyond the human annotation spread. Consequently, neither backend reproduces the magnitude or structure of radiologist disagreement.

2 Method

We study whether a committee of persona conditioned vision language agents reproduces the ambiguity expressed by a panel of radiologists. Our pipeline (Figure 1) converts a single image into a set of plausible segmentation masks, aggregates them into an ambiguity map, and compares that map against the ambiguity of the human annotations along two axes: spatial alignment and diversity. Crucially, no component is trained on multiple annotations; the committee is constructed entirely at inference time.

2.1 Problem setting

Let $x \in \mathbb{R}^{H \times W}$ be an image annotated by R experts, giving binary masks $\{y_r\}_{r=1}^{R}$, $y_r \in \{0,1\}^{H \times W}$. The empirical *human probability map* is the per pixel fraction of raters who label the pixel foreground,

$$p_h(x) \;=\; \frac{1}{R} \sum_{r=1}^{R} y_r, \tag{1}$$

and the *human ambiguity map* is its per pixel binary entropy,

$$A_h(x) \;=\; \mathcal{H}\big(p_h(x)\big), \qquad \mathcal{H}(p) = -\,p \log_2 p \;-\; (1-p) \log_2 (1-p), \tag{2}$$

with $\mathcal{H}(0) = \mathcal{H}(1) = 0$. A pixel is a site of disagreement when $0 < p_h < 1$. Given only the image x, an ambiguity model must produce a set of predicted masks $\{\hat{y}_j\}_{j=1}^{J}$ whose induced ambiguity resembles A_h. J is the number of agents in the committee.

2.2 Persona committee

A committee is J agents, each a persona conditioned prompt to a shared vision language model (VLM). Persona π_j is a natural language reading style that shifts how aggressively the agent delineates a finding. We use J=4 personas on a tight to loose axis (*conservative*, *boundary*, *balanced*, *inclusive*; Figure 1, left). Since the lesions are small and low in contrast, the VLM sees a legible view $\tilde{x} = \phi(x)$, where ϕ upsamples the patch to 512×512 and applies a percentile contrast stretch (this never uses the annotations, so it introduces no leakage). Given $\tilde{x}$ and π_j, the VLM returns a presence decision $s_j \in \{0, 1\}$ (it may abstain) and, if present, a box $b_j \in \mathbb{R}^4$ rescaled to the native patch:

$$\big(s_j,\ b_j\big) \ = \ \mathrm{VLM}\big(\tilde{x},\ \pi_j\big). \tag{3}$$

The VLM only localizes; delineation is delegated to MedSAM [10], so an abstaining agent contributes an empty mask, letting the committee express disagreement about lesion *existence* as well as extent:

$$\hat{y}_j = \begin{cases} \mathrm{MedSAM}(x, b_j), & s_j = 1, \\ \mathbf{0}, & s_j = 0. \end{cases} \tag{4}$$

The committee ambiguity mirrors Eqs. (1)–(2) on the predicted masks: $p_a = \frac{1}{J}\sum_j \hat{y}_j$, $A_a = \mathcal{H}(p_a)$. We instantiate the VLM with three backends of increasing specialization, a small general model (Qwen2.5-VL [2]), a frontier general model (Claude Opus 4.8), and a CT trained medical model (MedGemma 1.5 [15]), holding personas, ϕ, and MedSAM fixed so any difference is attributable to the backend.

2.3 The boundary null

Both human and committee disagreement concentrate near object boundaries, so a naive correlation between A_a and A_h can be positive for a trivial geometric reason. We therefore introduce a parameter free *boundary null* that predicts disagreement from geometry alone. Let $C = \{u : p_h(u) \geq \frac{1}{2}\}$ be the majority consensus region (pixels marked foreground by at least half the raters) and $d(u)$ the Euclidean distance from pixel u to its boundary ∂C. The null is a band around the consensus contour,

$$A_{\mathrm{null}}(u) \ = \ \exp\Big(-\tfrac{d(u)^2}{2\sigma^2}\Big). \tag{5}$$

It contains no learned or image derived content beyond the consensus shape, so any committee that fails to exceed it adds nothing over a distance transform, and it is the reference against which every committee is judged.

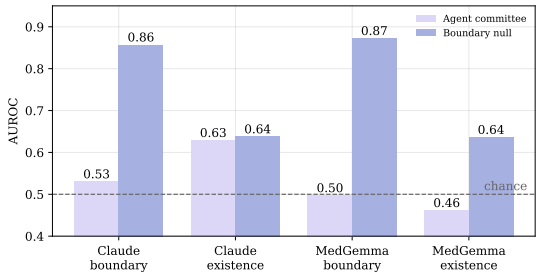

Fig. 2. AUROC of committee ambiguity (light) vs the parameter free boundary null (dark) at predicting rater disagreement. The null dominates on boundary cases and is never surpassed, with the committee only tying it on existence cases.

2.4 Evaluation

Distribution matching. To place our committees in the existing lineage we report generalized energy distance (GED), Collective Insight (CI) [8, 13], and maximum Dice matching $D_{\max}$.

Spatial alignment. Our primary question is *where* a map locates disagreement. Taking the disagreement sites $\mathbf{1}[0 < p_h < 1]$ as labels and $A \in \{A_a, A_{\text{null}}\}$ as scores, we report AUROC within a dilated band around the lesion (so the trivially agreeing background does not inflate it). A committee is informative only if $\text{AUROC}(A_a) > \text{AUROC}(A_{\text{null}})$.

Diversity. We also ask whether the committee disagrees as *much* as the raters, comparing mean pairwise Dice among agents and among raters,

$$\bar{D}_a = \binom{J}{2}^{-1} \sum_{i<j} \text{Dice}(\hat{y}_i, \hat{y}_j), \quad \bar{D}_h = \binom{R}{2}^{-1} \sum_{i<j} \text{Dice}(y_i, y_j). \tag{6}$$

$\bar{D}_a \approx \bar{D}_h$ is calibrated; $\bar{D}_a > \bar{D}_h$ collapses into agreement, $\bar{D}_a < \bar{D}_h$ scatters beyond the human spread.

Stratification. We split two regimes: *boundary* cases, where all R raters mark the lesion ($\min_r |y_r| > 0$), isolate disagreement about extent (where A_{null} is strong); *existence* cases, where at least one rater abstains, isolate disagreement about presence, which no geometric map captures. All metrics are reported per stratum, as this distinction drives our findings.

3 Experiments and Results

3.1 Dataset and setup

We evaluate on LIDC/IDRI [1], the standard benchmark for ambiguous lung nodule segmentation, where each nodule is annotated by up to four radiologists.

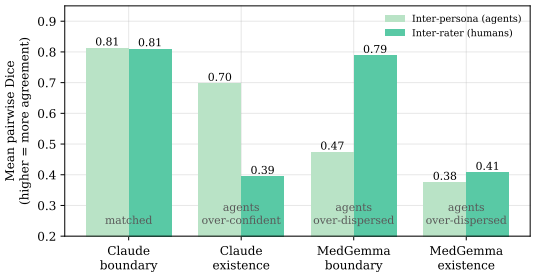

Fig. 3. Mean pairwise Dice among agents (light) vs radiologists (dark); higher means more agreement. Neither backend matches the human spread across strata: Claude collapses on existence, MedGemma over disperses on boundary.

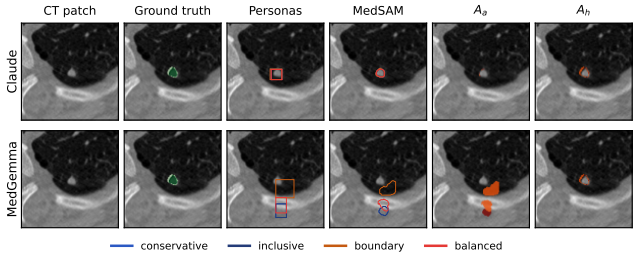

Fig. 4. Committee ambiguity on a random case. Top: Claude; bottom: MedGemma; both share the image and ground truth, so all variation is backend driven. Columns: CT patch; radiologist masks overlaid; persona boxes; MedSAM masks; committee ambiguity A_a; human ambiguity A_h. Claude's personas nearly coincide, so A_a tracks A_h; MedGemma's scatter onto rib and soft tissue, placing A_a where no lesion exists.

We use the preprocessing of Kohl et al. [8] (15,096 axial patches of 128×128, each with four masks; an unmarked rater contributes an empty mask, preserving both extent and existence disagreement) and the 60/20/20 split of [4] (9,058/3,019/3,019), reporting on the test split. The split is patch level, following and shared by all baselines. Committees are training free and run at inference. We use three fixed backends, Qwen2.5-VL [2], Claude, and the CT trained MedGemma [15], with the four personas, ϕ, and MedSAM [10] held fixed. Claude, queried through a paid interface, is evaluated on a random 300 patch subset. MedGemma and the baselines use the full test split. The boundary null uses σ=4 pixels.

3.2 Distribution matching

We first place the committees in the existing lineage using generalized energy distance (GED), Collective Insight (CI), and maximum Dice matching $D_{\max}$ (Table 1). As expected, no training free committee is competitive: the best reaches

Table 1. Comparison of probabilistic segmentation methods and committee-based baselines. Best and second-best results are shown in **bold** and underlined, respectively. ↑/↓ indicate higher/lower is better.

Method	GED ↓	CI ↑	$D_{\max}$ ↑
Prob-U-Net [8]	0.353	0.731	0.892
PHi-Seg [3]	**0.270**	0.736	0.904
Gen. Prob-U-Net [4]	0.299	0.707	0.905
CIMD [13]	0.321	**0.759**	**0.915**
Programmatic committee	0.968	0.299	0.253
Prompt-jitter committee	0.823	0.312	0.264
SAM multi-mask	1.180	0.135	0.106
Claude committee	0.961	0.246	0.339
MedGemma committee	1.020	0.106	0.188

Table 2. Stratified ambiguity results for the Claude and MedGemma committees. Agent and boundary-null performance are reported using AUROC, while persona and radiologist agreement are measured using pairwise Dice.

Backend	Stratum	n	AUROC ↑		Agreement ↑		
			Agent	Boundary null	Persona Dice	Rater Dice	CI
Claude	Boundary	95	0.530	0.857	0.813	0.811	0.441
Claude	Existence	205	0.629	0.638	0.698	0.395	0.156
MedGemma	Boundary	946	0.497	0.872	0.474	0.789	0.028
MedGemma	Existence	2073	0.462	0.636	0.376	0.408	0.142

$D_{\max}$=0.339 (Claude) versus over 0.89 for the trained baselines, unsurprising since committees never observe multiple annotations and MedSAM yields one mask per prompt. These results motivate our central question, which is not whether committees match the reference distribution, but whether their disagreement falls where radiologists actually disagree. Qwen could not generate any bounding boxes, so their evaluation was avoided in the tables and figures.

3.3 Geometry outperforms the committee

Our primary result is that a parameter free boundary null predicts human disagreement better than any committee (Figure 2, Table 2). On boundary cases the null is decisive (AUROC 0.86/0.87 for Claude/MedGemma) while the committees barely exceed chance (0.53/0.50): when all four radiologists mark the lesion, their disagreement is almost entirely explained by distance to the consensus contour. On existence cases the null is weaker (0.64), as there is no clean consensus boundary, and the best committee only matches it (Claude, 0.63), never exceeds it. This is the sole regime where a committee is competitive, and it is the one requiring reasoning about presence rather than extent. Across all four cells the null is never surpassed, our central negative finding: assembling readily available agents does not improve on a trivial geometric baseline at predicting where experts disagree.

3.4 Diversity is miscalibrated in opposite directions

We next ask whether the committee disagrees as *much* as the radiologists, comparing mean inter persona and inter rater Dice (Figure 3, Table 2). The two backends fail in opposite directions. The frontier committee *collapses*: on existence cases its personas agree far more than the raters (0.70 versus 0.39), so it is confident precisely where humans most disagree about whether a lesion exists. The medical committee *scatters*: on boundary cases its personas disagree far more than the raters (0.47 versus 0.79), producing spurious variation where the panel largely agrees. Only one cell is calibrated (Claude, boundary: 0.81 versus 0.81). This double dissociation shows the failure is structural, not a matter of prompt tuning: a committee simultaneously too confident on existence and too scattered on boundaries cannot be fixed by uniformly scaling its diversity. Reproducing radiologist ambiguity is thus not an assembly problem, it requires diversity calibrated to the kind of disagreement present and an explicit abstention mechanism that a box prompted committee lacks. Figure 4 shows both failure directions on one case.

3.5 Qualitative analysis

Figure 4 makes the two failure modes concrete on one case, with a backend per row. For Claude (top), the four personas place nearly coincident boxes and MedSAM returns tightly clustered masks, so A_a closely tracks A_h on this case where the radiologists themselves agree, accurate but lacking spread. For MedGemma (bottom), the personas disagree in the wrong place: two boxes fall below the lesion onto rib and soft tissue, MedSAM faithfully segments these spurious prompts, and A_a places strong ambiguity where no lesion exists. The two rows visualize the quantitative results, agent ambiguity is spatially misplaced and its magnitude miscalibrated in opposite directions across backends.

4 Conclusion

We asked whether persona prompted VLM committees reproduce radiologist inter-rater ambiguity, and found that they do not: across a small general, a frontier general, and a medical VLM, a parameter free boundary null predicts human disagreement better than any committee and is never surpassed, while committee diversity is miscalibrated in opposite directions, collapsing on existence cases for the frontier model and scattering on boundary cases for the medical model. Our study is limited to one dataset (LIDC/IDRI), four hand designed personas, a single segmenter (MedSAM), and a 300 patch subset for the frontier committee, and does not quantify statistical uncertainty around the reported results. Future work should span other modalities, learned persona sets, and stronger segmenters, and move from prompting toward the mechanisms the committee lacks: calibrated diversity and explicit abstention. Our findings suggest reproducing clinical ambiguity is not an assembly problem solved by composing off the shelf agents, but a modeling problem demanding calibrated uncertainty and grounded perception.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., et al.: The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. Medical physics **38**(2), 915–931 (2011)
2. Bai, S., Chen, K., Liu, X., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Baumgartner, C.F., Tezcan, K.C., Chaitanya, K., Hötker, A.M., Muehlematter, U.J., Schawkat, K., Becker, A.S., Donati, O., Konukoglu, E.: Phiseg: Capturing uncertainty in medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 119–127. Springer (2019)
4. Bhat, I., Pluim, J.P., Kuijf, H.J.: Generalized probabilistic u-net for medical image segementation. In: International workshop on uncertainty for safe utilization of machine learning in medical imaging. pp. 113–124. Springer (2022)
5. Hu, T., Collier, N.: Quantifying the persona effect in llm simulations. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 10289–10307 (2024)
6. Joskowicz, L., Cohen, D., Caplan, N., Sosna, J.: Inter-observer variability of manual contour delineation of structures in ct. European radiology **29**(3), 1391–1399 (2019)
7. 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)
8. Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J.R., Maier-Hein, K., Eslami, S., Jimenez Rezende, D., Ronneberger, O.: A probabilistic u-net for segmentation of ambiguous images. Advances in neural information processing systems **31** (2018)
9. Li, B., Yan, T., Pan, Y., Luo, J., Ji, R., Ding, J., Xu, Z., Liu, S., Dong, H., Lin, Z., et al.: Mmedagent: Learning to use medical tools with multi-modal agent. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 8745–8760 (2024)
10. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature communications **15**(1), 654 (2024)
11. Menze, B., 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)
12. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
13. Rahman, A., Valanarasu, J.M.J., Hacihaliloglu, I., Patel, V.M.: Ambiguous medical image segmentation using diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11536–11546 (2023)

14. Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., Hashimoto, T.: Whose opinions do language models reflect? In: International conference on machine learning. pp. 29971–30004. PMLR (2023)
15. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
16. Wu, C., Zhang, X., Zhang, Y., Hui, H., Wang, Y., Xie, W.: Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. Nature Communications **16**(1), 7866 (2025)