CRITERIA AS CODE: WHEN WRITTEN RULES BEAT LEARNED ONES, AND WHEN THEY DON'T

Anonymous authors
Paper under double-blind review

ABSTRACT

Written clinical criteria offer an explicit alternative to learning every diagnostic decision from annotated images. But when does protocol-derived decision logic transfer better than logic learned from data? We study this question with CRITERIACODE, which separates multiple sclerosis (MS) lesion segmentation into learned candidate detection, 34 radiological attributes, and an executable criteria program with calibrated thresholds. The main comparison uses a hand-written program and a shallow decision tree trained on the identical clinically informed attributes. Both expose their decision logic, while its source differs: the annotation protocol or the training labels. In-domain, the tree exceeds the program from 32 annotated patients, and the end-to-end network remains more accurate. Under transfer to MSLesSeg, the program achieves lesion Dice of 0.587 versus the tree's 0.550 with FLAIR input, matching the source-trained network's mean Dice. On Shifts, the ordering reverses. Controlled noise and bias-field experiments do not reproduce this reversal as a function of severity alone. Attribute analysis instead identifies greater orientation and multi-contrast drift on Shifts. Removing the associated rules and recalibrating closes 70% of the FLAIR program–tree gap there, but reduces multi-contrast performance on MSLesSeg. This intervention supports an attribute-specific explanation rather than a general robustness advantage for written rules. A threshold sweep also exposes a mismatch between a literal 3 mm size rule and its implementation on thick-slice images. LLM-generated programs are less reliable on real candidates than the hand-written reference. Together, these results clarify the benefits and limits of protocol-derived structure: its transfer value depends on the attributes affected by the target shift, while explicit rules make those dependencies inspectable and editable. We also correct two label-orientation errors in the public MS3SEG release; anonymized code and every generated program accompany this submission.

1 INTRODUCTION

Not every bright spot on a brain MRI is a lesion. On T2-FLAIR images of patients with multiple sclerosis (MS), demyelinating lesions appear as white matter hyperintensities (WMH), but so do physiological findings: thin periventricular rims of CSF-contaminated tissue and symmetric age-related small-vessel changes (Fazekas et al., 1987; Wardlaw et al., 2013). Radiologists separate the two not by intensity but by *criteria*: an MS lesion is ovoid, at least about 3 mm, periventricular or juxtacortical, oriented perpendicular to the ventricular wall (a "Dawson's finger"), confirmed on T2 and often hypointense on T1, whereas a physiological rim hugs the ventricle, is thin and bilateral (Thompson et al., 2018; Filippi et al., 2016). Segmentation networks trained on lesion masks infer these criteria implicitly from a few dozen patients and largely fail to: on the recent MS3SEG dataset, the first to annotate normal and abnormal WMH separately, strong 2D networks reach Dice of only 0.59–0.67 on the two hyperintensity classes while segmenting ventricles at 0.89 (Bawil et al., 2026a). The perceptual part is easy; the criterial part is hard.

We argue these two parts should be solved by different machinery. Perception—finding every hyperintensity and the ventricles—is dense and data-driven, and convolutional networks solve it well but in a scanner-specific way. The criteria deciding pathology are *written down* in language, shared across scanners, and refer to lesion-level quantities (size, shape, location, symmetry, multi-contrast

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

behaviour) computable exactly. We therefore propose CRITERIACODE (*criteria as code*): a class-agnostic perception network proposes hyperintensity candidates; radiological attributes are computed for each; and the written protocol, with the attribute schema, is *compiled*—by hand or by an LLM—into an executable, readable program that classifies each candidate as an MS lesion, a physiological hyperintensity, or a false positive. Every threshold is a named parameter, calibrated on training data without changing the program's logic—auditable, transferring unchanged to new scanners, editable without retraining.

CRITERIACODE is neuro-symbolic in a specific, testable sense: its symbolic part is authored from clinical text—by hand or by an LLM—and only *tuned* by data, so we can ask whether language-derived structure helps over structure fit directly from data. Our experiments on MS3SEG and two external multi-scanner datasets (Shifts (Malinin et al., 2022), MSLesSeg (Guarnera et al., 2025)) address exactly this. Our contributions are:

- We introduce CRITERIACODE, a decomposition of pathology-aware segmentation into learned perception and an explicit, calibratable criteria program authored from the written protocol either by hand or by an LLM, together with a lesion-level attribute schema that makes radiological criteria (periventricular distance, Dawson's-finger orientation, rim thickness, cortical proximity, bilateral symmetry, T1/T2 confirmation) computable from ventricle and brain masks.
- We show, at matched interpretability, where compiled criteria earn their keep: under a moderate, realistic cross-scanner shift the unmodified program matches the source-trained network, clearly beats a no-reasoning baseline, and is never behind an equally clinically-informed decision tree fit from the same attributes—significantly ahead of it with single-contrast input (Section 4.2)—so the *origin* of the decision logic, not just its interpretability, matters here. Auditability itself is measured: rule-level precision and a parameter-edit sweep (Section 4.5) surface a concrete protocol-to-acquisition mismatch.
- We report, with equal care, where this does not hold: not across shift severity—ranking inverts under a larger shift, and a controlled synthetic sweep does not reproduce this by magnitude alone, pointing instead to which attributes a shift disturbs—confirmed by pruning the rules built on them (Section 4.2); not in-domain (neither program nor classifier matches the network's own accuracy even confound-controlled, Section 4.1; the tree exceeds the program's data-efficiency ceiling from 32 patients, Section 4.3); not for automatic authorship (LLM-compiled programs are weaker than the hand-written one on real candidates, Section 4.4).
- We provide a criteria-group ablation, tree/GBDT comparisons across five data budgets, a few-shot target-domain adaptation experiment, a controlled shift-severity sweep, a prompt-content ablation, an LLM-program ensemble with individual failure modes, and a zero-shot VLM judge baseline; anonymized code and every generated program accompany this submission.

2 RELATED WORK

MS lesion segmentation and the normal/abnormal distinction. Public MS datasets (Carass et al., 2017; Commowick et al., 2018; Lesjak et al., 2018; Malinin et al., 2022; Guarnera et al., 2025) annotate lesions as a single class; most methods are U-Net variants (Ronneberger et al., 2015; Isensee et al., 2021; Hatamizadeh et al., 2022), and lesion/physiological-hyperintensity confusion is a known false-positive source (Kuijf et al., 2019; Sudre et al., 2015). MS3SEG (Bawil et al., 2026a) is the first to label ventricles, normal and abnormal WMH separately; a follow-up shows adding the normal class as a training target alone improves lesion Dice (Bawil et al., 2026b). Both operate at the pixel level; neither uses the written criteria that define the classes.

Classical candidate-classification CAD. Before end-to-end networks, WMH and MS-lesion detection was commonly posed exactly as we pose it: extract candidate voxels, compute intensity/shape/location features, and classify—BIANCA (Griffanti et al., 2016) runs k-NN on intensity/spatial-prior features, and several MICCAI WMH Challenge entries add a kNN or random-forest stage to prune false positives (Steenwijk et al., 2013; Kuijf et al., 2019). CRITERIACODE's

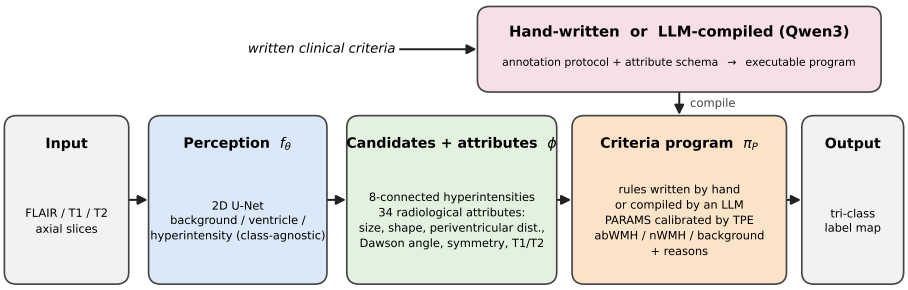


Figure 1: CRITERIACODE. A class-agnostic network finds ventricles and all hyperintensities; each becomes a candidate with 34 computed radiological attributes; the written protocol is compiled, by hand or by an LLM, into an executable program with calibrated named thresholds; the program decides, with stated reasons, whether each candidate is an MS lesion, a physiological hyperintensity or a false positive.

perception-then-classify structure is this same recipe; what changes is the classifier's origin (a written protocol compiled into code, not a model fit to voxels) and that we calibrate, not fit, a handful of named thresholds. Sections 4.1–4.2 test whether that difference in origin, not interpretability (which classical feature-based classifiers share), earns anything a same-capacity classifier does not: in-domain, no; under moderate shift, yes.

Language-guided and reasoning segmentation. Text-prompted medical segmenters condition a dense decoder on class names or descriptions (Liu et al., 2023; Zhao et al., 2023; 2025; Ma et al., 2024; Jiang et al., 2026); language selects *what* to segment but the decision stays implicit. Reasoning segmentation couples a vision–language model (VLM) with a mask decoder (Lai et al., 2024; Liu et al., 2025; Yan et al., 2025; Trinh et al., 2025; Xing et al., 2026); its chain of thought is free text with no consistency guarantee, and pixel-accurate masks for few-millimetre lesions remain out of reach. We use a VLM as a zero-shot judge baseline and find it clearly inferior to compiled criteria.

LLM program synthesis and neuro-symbolic clinical rules. Visual programming lets an LLM decompose a query into calls to perception modules (Gupta & Kembhavi, 2023; Surís et al., 2023). Closer to our setting, INFORM-CT (Tankel et al., 2026) has an LLM write scripts applying ACR guidelines to CT lesion attributes, and recent work synthesises clinical scoring rules from tabular data (Ruhrberg Estévez et al., 2026; Xu et al., 2026), feature-extraction code (Urooj et al., 2025), or LLM-navigated guideline trees (Deng et al., 2026), in the spirit of interpretable high-stakes models (Rudin, 2019). Wang et al. (2025) show LLM-compiled symbolic programs generalise better than a prototype tree under visual shift; we find the opposite ordering under severe scanner shift and the same ordering only under moderate shift, consistent with our claim that transfer depends on which attributes a given shift disturbs rather than on programmatic origin per se. Our pruning intervention instead targets named clinical rules, relating to feature-targeted domain-generalisation pruning (Wang et al., 2022) and concept-interface stability under adaptation (Cho et al., 2026). CRITERIACODE differs in three ways: it refines *inside* a segmentation pipeline, evaluated by voxel- and lesion-level metrics, with parameters calibrated (not fit) against that objective while logic stays fixed; we test robustness directly, on unchanged programs run on data from other scanners; and we hold the program's structural origin fixed as a variable against a matched-interpretability tree, rather than assuming language is the source of any benefit such programs show.

3 METHOD

Problem. Given co-registered axial FLAIR, T1 and T2 volumes of an MS patient, we seek a label map over {background, ventricle, normal WMH (nWMH), abnormal WMH (abWMH)}, where *abnormal WMH* denotes an MS lesion and *normal WMH* a physiological hyperintensity per protocol, not a clinical-harmlessness claim. MS3SEG was annotated slice by slice on axial FLAIR with thick ($\approx$6 mm) slices, so we treat volumes as 2D slice stacks and report 3D metrics per patient.

Overview. CRITERIACODE (Figure 1) has three stages. (i) A *perception* network f_θ segments each slice into background, ventricle and *hyperintensity*, i.e. the union of the two WMH classes. (ii) Each 8-connected hyperintensity component becomes a *candidate* c, for which we compute an attribute vector $\phi(c) \in \mathbb{R}^d$ (d=34) from the candidate mask, the predicted ventricle mask, the brain mask and the three contrasts. (iii) A *criteria program* π_P, written by hand or by an LLM and parameterised by named thresholds P, maps $\phi(c)$ to a decision in {abWMH, nWMH, background} and a graded lesion score in $[0, 1]$. The final label map relabels every hyperintensity voxel with its candidate's decision.

Perception. f_θ is a 2D residual U-Net with instance normalisation, deep supervision and Dice+cross-entropy loss (a 2D nnU-Net configuration; details in Appendix B). The hyperintensity class merges normal and abnormal WMH, so its training signal is unambiguous and transfers to any dataset annotating lesions and ventricles.

Candidate attributes. The schema ϕ operationalises the quantities named in diagnostic criteria: size/shape, location relative to anatomy (ventricle, horns, brain surface, midline), morphology vs. the ventricle wall (a Dawson's-finger/rim angle from the ventricle distance-transform gradient, 0° a finger, 90° a rim), bilateral symmetry, multi-contrast evidence (FLAIR/T2/T1 intensity and contrast against a surrounding ring, z-scored within the brain), and slice/3D/patient context—34 deterministic attributes, computed in seconds, identically defined on every dataset (Table 3, Appendix C).

Compiling criteria. The LLM receives (a) the dataset's annotation protocol, optionally with general diagnostic knowledge; (b) the attribute schema; (c) an output contract: a Python program defining `PARAMS` (named thresholds with plausible defaults) and `classify(c, P)`, returning a decision, lesion score and fired rules, standard library only, missing attributes treated as neutral. We sample K programs at temperature 0.7, keeping those satisfying the contract on held-in candidates. Programs are typically 60–120 lines, read like a structured report (example in Appendix C), using open-weight Qwen3 (Yang et al., 2025) run locally.

Calibration. Default thresholds encode clinical conventions ("3 mm") that need not match a dataset's resolution, so we calibrate P on training candidates without changing the program's logic: $P^\star = \arg\max_P J(P)$, J the area-weighted Dice between decisions and reference classes—a candidate-level surrogate of voxel Dice, evaluated in milliseconds. J is piecewise constant in P, so we use tree-structured Parzen estimation (Bergstra et al., 2011; Akiba et al., 2019) (defaults as the first trial, 400 trials, ranges $[P_0/4, 4P_0]$ or $[0, 1]$ for fractions). With at most 20 scalars fitted, calibration needs little data (Section 4.3).

Baselines that isolate the contribution of language. The attributes are the same for every candidate-level method, isolating what language-derived structure adds over data-derived structure, at matched and greater capacity. Headline in-domain and transfer results use a program authored by us from the protocol and attribute schema, before any LLM output was seen or any real candidate existed to compute statistics from, isolating decomposition value from LLM-authoring reliability (Section 4.4). We compare against (1) the end-to-end four-class network; (2) a gradient-boosted decision tree classifier (GBDT) g_ψ (Pedregosa et al., 2011) on $\phi(c)$, an accuracy ceiling with an expressive, non-interpretable model; (3) a shallow decision tree ($\leq$4 levels, $\leq$16 leaves) on the same clinically-informed attributes—equally auditable, but data-fit rather than written by hand or by an LLM; (4) a zero-shot VLM judge (Qwen3-VL-8B) shown the slice, a zoomed crop and the protocol, lesion probability read from the logits of "abnormal" vs. "normal" (a candidate-level AUC, not a segmentation system; Section 4.4); and (5) an *oracle* assigning each candidate its majority reference class, upper-bounding every candidate-level method.

4 EXPERIMENTS

Data. MS3SEG (Bawil et al., 2026a) contains 100 MS patients (74 female, age 35.5$\pm$8.9) scanned on a single 1.5 T Toshiba scanner with axial T2-FLAIR (256$\times$256, 0.9 mm in-plane, 18–20 slices of 6–8 mm), T1 and T2, with expert tri-mask annotations of ventricles, normal and abnormal WMH (1,985 slices; 10,048 annotated components). We use the official split: 20 held-out test patients

Table 1: Tri-class segmentation on MS3SEG (mean±SD over patients). Nets train end-to-end; decomposition rows use the class-agnostic network (FLAIR+T1+T2), reclassifying its candidates. "Program": authored by us from the protocol (Section 3); the LLM variant is evaluated separately (Section 4.4). "Decision tree": depth-4, ≤16-leaf, same attributes—equally auditable, fit from data, not language. Bold marks the best non-oracle value per column.

	Cross-validation (80 patients)				Held-out test (20 patients)			
System	Dice vent.	Dice nWMH	Dice abWMH	Lesion F1	Dice vent.	Dice nWMH	Dice abWMH	Lesion F1
Four-class net (FLAIR)	0.921±0.02	**0.722±0.06**	**0.744±0.16**	**0.757±0.14**	0.932±0.01	0.744±0.08	0.784±0.07	0.823±0.10
Four-class net (FLAIR+T1+T2)	0.919±0.02	0.713±0.06	0.743±0.15	0.756±0.13	0.931±0.01	**0.744±0.08**	**0.793±0.07**	**0.828±0.09**
Oracle (perfect candidate classification)	**0.922±0.02**	0.707±0.09	0.758±0.15	0.878±0.14	**0.933±0.01**	0.748±0.10	0.810±0.07	0.925±0.06
Program (no calibration)	**0.922±0.02**	0.585±0.12	0.633±0.17	0.645±0.16	**0.933±0.01**	0.634±0.14	0.650±0.14	0.670±0.17
Program (calibrated)	**0.922±0.02**	0.620±0.11	0.654±0.17	0.659±0.15	**0.933±0.01**	0.668±0.13	0.688±0.13	0.699±0.14
Decision tree (depth 4, same attributes)	**0.922±0.02**	0.628±0.11	0.672±0.17	0.724±0.15	**0.933±0.01**	0.662±0.14	0.700±0.11	0.714±0.13
LLM-program ensemble (calibrated)	**0.922±0.02**	0.605±0.12	0.588±0.18	0.601±0.17	**0.933±0.01**	0.626±0.15	0.622±0.12	0.642±0.14
Learned classifier (GBDT)	**0.922±0.02**	0.645±0.11	0.686±0.17	0.733±0.17	**0.933±0.01**	0.683±0.13	0.740±0.08	0.807±0.09

and five folds of 64/16 over the remaining 80, reported as out-of-fold "CV" and, with the five fold models ensembled, "Test". We found and corrected two release issues: seven patients' label volumes are mirrored relative to their FLAIR (headers disagree in the left–right axis), and one patient's four-class file lacks the lesion class; we reorient and recompose from per-class masks (Appendix A). For cross-scanner evaluation we use the openly licensed part of Shifts (Malinin et al., 2022) (46 scans from 37 patients—28 with one timepoint, 9 with two—on 3 T Philips (Carass et al., 2017) and 3 T Siemens (Lesjak et al., 2018)) and MSLesSeg (Guarnera et al., 2025) (93 scans from 53 patients: 28 with one timepoint, 15 with two, 5 with three, 5 with four; 1.5/3 T multi-vendor), both lesion-only; we harmonise their 1 mm isotropic volumes to MS3SEG geometry (Appendix A). No external data is used for training or calibration.

Metrics. Per patient and class: 3D Dice, IoU, 95th-percentile Hausdorff distance, lesion-wise precision/recall/F1 (26-connected; detected if any predicted voxel overlaps), and the volume of reference lesion voxels mislabelled as normal WMH and vice versa (means/SDs over patients).

Implementation. Perception and four-class networks share architecture, augmentation and schedule (2D residual U-Net, six resolution levels, 32–320 channels, deep supervision, AdamW, 200 epochs, mixed precision; Appendix B). Inputs are z-scored within a dilated brain mask; we train FLAIR-only and FLAIR+T1+T2 models. Programs are sampled from Qwen3-32B/14B (Yang et al., 2025) at temperature 0.7, chain-of-thought disabled (Section 4.4); calibration uses Optuna's TPE with 400 trials per fold. The classifier is scikit-learn's histogram GBDT, 300 rounds. All experiments run on two 48 GB GPUs.

4.1 Main results on MS3SEG

The end-to-end network is a strong baseline, and the decomposition does not beat it in-domain. The retrained four-class networks reach CV Dice 0.722/0.744 (nWMH/abWMH, FLAIR-only) and 0.713/0.743 (three modalities), test 0.744/0.784 and 0.744/0.793 respectively—already well above the dataset paper's own published U-Net baseline (CV 0.645/0.669, on uncorrected labels and a shorter schedule; Appendix A). The fully decoupled pipeline—class-agnostic perception, then a criteria program or a learned classifier on its candidates—does *not* close this gap: the best real method (the learned classifier) reaches only CV 0.648/0.693 (FLAIR) and 0.645/0.686 (three modalities), 5–7 points *below* the end-to-end network, consistently across both modality settings (Table 1; FLAIR-only rows in Table 6, Appendix C).

Nor does the program's language-derived structure beat data-derived structure at matched interpretability. Since the remaining case for the program rests on being auditable rather than accurate, the relevant competitor is not just the unrestricted classifier but an equally auditable one built from the same attributes without any clinical language: a depth-4, ≤16-leaf decision tree (Table 1). On real candidates it reaches CV 0.628/0.672 and test 0.662/0.700, *exceeding* the calibrated program (0.620/0.654 CV, 0.668/0.688 test)—every split chosen purely to fit the training labels, not written from clinical language, though built on the same clinically-informed attributes as the program. Section 4.3 shows this is not a one-off: the tree is roughly tied with the program through 16

annotated patients and clearly exceeds it from 32 on. Naming a split "periventricular" rather than "`dist_to_ventricle_mm < 2.1`" does not, by itself, buy accuracy a data-fit split of the same depth does not already provide; whatever value the program's language-derived structure has over data-derived structure at matched capacity is not in Dice.

This is not a localization artifact. First, an oracle assigning each perception-derived candidate its true majority label reaches only CV 0.707–0.710/0.758–0.761, barely above the network's own accuracy and far below the 0.894/0.924 ceiling on ground-truth-shaped candidates (Section 4.3): purity drops from 0.91 to 0.75–0.76 from background over-extension and cross-class mixing in roughly equal measure—a boundary *placement* problem, not a confidence one (85% of mass already retained at 0.9). Second, and decisively: re-extracting candidates from the *four-class network's own* predicted region, the classifier still reaches only CV 0.656/0.699 (FLAIR) and 0.649/0.697 (multi) and the program only 0.634/0.669 and 0.632/0.666 (Table 6)—5–7 and 7–9 points below the network's own native decision on those same pixels (0.722/0.744 and 0.713/0.743, $p < 10^{-9}$). Controlling for localization does not close the gap: a jointly-trained, per-pixel classifier is simply more expressive than a candidate-level attribute vector—not a Dice win, but useful where a per-pixel classifier's opacity is the obstacle: a stated rule instead of a gradient, reusable from a handful of pilot annotations (Section 4.3) rather than a full retrain.

LLM-written programs: strong on idealised candidates, weaker on real ones. An unweighted ensemble of four Qwen3-32B-written programs (majority vote; Section 4.4 has per-program numbers) reaches CV 0.796/0.758 on ground-truth-shaped candidates—matching or exceeding the hand-written program (0.788/0.792) with no manual engineering. On the real candidates in Table 1 it instead reaches only 0.605/0.588, *below* both the hand-written program (0.620/0.654) and the classifier: the LLM writes overly specific rules when attributes are reliable by construction—brittle once boundaries carry real noise, where coarser hand-written rules and a classifier's reweighting degrade more gracefully. A genuine limitation of automatic synthesis as implemented (worked example: Appendix C, Figure 3).

4.2 Cross-scanner generalisation

Every method degrades substantially under scanner shift, as expected, but by very different amounts depending on how different the target scanner is from the source. MSLesSeg's 93 scans come from 53 patients (28 with one timepoint, 15 with two, 5 with three, 5 with four) and Shifts' 46 scans from 37 patients (28 with one timepoint, 9 with two); every test below is computed at patient level (timepoints averaged) for both datasets to avoid treating repeats as independent. On MSLesSeg (a mix of 1.5 T and 3 T multi-vendor data clinically closer to MS3SEG's acquisition), the calibrated program transfers with *no adaptation* to lesion Dice statistically indistinguishable from that of the same network, trained only on MS3SEG (0.587 vs. 0.587 FLAIR-only, 0.566 vs. 0.557 multi; paired Wilcoxon $p = 0.41$, $p = 0.53$; 95% CI on the paired difference $[-0.022, +0.015]$ and $[-0.016, +0.037]$, small and centred near zero rather than merely non-significant), while significantly beating a no-reasoning baseline that calls every detected hyperintensity a lesion (0.506, 0.497; $p < 10^{-8}$ both settings, Table 2). Criteria written once transfer about as well as a network that must re-learn everything from source pixels, and clearly better than not reasoning at all.

On the Shifts benchmark (46 scans from 37 patients, 3 T Philips/Siemens—a larger jump from MS3SEG's single 1.5 T Toshiba scanner), the picture reverses. Every method's lesion Dice falls to 0.22–0.38; the program remains statistically tied with the network (0.277 vs. 0.287 FLAIR, patient-level $p = 0.64$, CI $[-0.036, +0.024]$; 0.231 vs. 0.219 multi, $p = 0.14$, CI $[-0.020, +0.037]$) but is now significantly *worse* than the no-reasoning baseline (0.373, 0.277; $p = 2.4 \times 10^{-5}$ FLAIR, $p = 3.4 \times 10^{-5}$ multi): rules calibrated to MS3SEG's statistics actively hurt once the acquisition changes enough, and raw hyperintensity detection is the more transferable signal here. The tree and classifier both significantly beat the network with FLAIR (0.366, 0.376 vs. 0.287; $p = 1.4 \times 10^{-4}$, $p = 3.9 \times 10^{-5}$); with all three modalities the tree still does (0.249 vs. 0.219, $p = 0.010$) but the classifier is only a trend (0.240 vs. 0.219, $p = 0.09$)—the tree, not just the classifier, is a robust cross-scanner win here. Explicit criteria's cross-scanner value is real but dataset-dependent and can invert—net-positive on MSLesSeg's moderate shift, a net loss relative to doing nothing on Shifts' larger one, tracking which attributes drift (below), not shift magnitude.

Table 2: Cross-scanner lesion segmentation (mean±SD per scan; no target-data adaptation). "Program": our calibrated program, unmodified; "Tree": the depth-4 tree of Table 1, our matched-interpretability control (Section 4.2). Bold = best per column within dataset/setting; ties both marked. †: Dice significantly different from the four-class net (paired Wilcoxon, $p < 0.05$). Both datasets contribute more scans than patients (MSLesSeg: 53 patients, 93 scans, 28/15/5/5 with one/two/three/four timepoints; Shifts: 37 patients, 46 scans, 28/9 with one/two), so every significance test here is computed at patient level (timepoints averaged) though the displayed mean/SD stay per-scan. Every other difference is not significant (including every "matches"/"ties" claim in the text); these 16 tests are a descriptive screen, the four primary program-vs-tree comparisons Holm-corrected in Section 4.2.

Dataset	System	Dice	Lesion F1	Recall	Precision
Shifts (46 cases, 3 T Philips/ Siemens)	Four-class net (FLAIR)	0.287±0.23	0.238	0.147	**0.848**
	No-reasoning baseline (FLAIR)	0.373±0.21†	**0.315**	**0.301**	0.412
	Program, calibrated (FLAIR)	0.277±0.25	0.186	0.111	0.777
	Decision tree, unmodified (FLAIR)	0.366±0.28†	0.283	0.184	0.760
	Learned classifier (FLAIR)	**0.376±0.21**†	0.296	0.188	0.843
	Four-class net (multi)	0.219±0.24	0.147	0.085	0.850
	No-reasoning baseline (multi)	**0.277±0.20**†	**0.253**	**0.210**	0.438
	Program, calibrated (multi)	0.231±0.22	0.155	0.093	0.726
	Decision tree, unmodified (multi)	0.249±0.24†	0.174	0.105	0.896
	Learned classifier (multi)	0.240±0.22	0.149	0.086	**0.921**
MSLesSeg (93 cases, 1.5/3 T multi-vendor)	Four-class net (FLAIR)	**0.587±0.20**	**0.623**	0.549	**0.789**
	No-reasoning baseline (FLAIR)	0.506±0.21†	0.414	**0.646**	0.331
	Program, calibrated (FLAIR)	**0.587±0.20**	0.553	0.483	0.723
	Decision tree, unmodified (FLAIR)	0.550±0.22†	0.555	0.527	0.639
	Learned classifier (FLAIR)	0.585±0.20	0.580	0.501	0.758
	Four-class net (multi)	0.557±0.21	**0.561**	0.483	0.756
	No-reasoning baseline (multi)	0.497±0.21†	0.412	**0.593**	0.348
	Program, calibrated (multi)	**0.566±0.20**	0.525	0.518	0.600
	Decision tree, unmodified (multi)	0.557±0.20	0.535	0.503	0.645
	Learned classifier (multi)	0.559±0.22	0.537	0.425	**0.825**

Does language-derived structure transfer better than data-derived structure? Table 2 adds the matched-interpretability tree (unmodified from Table 1) to both datasets, testing the program directly against it—four pre-specified, Holm-corrected comparisons, all unchanged by correction; the answer depends on the shift. On MSLesSeg (patient-level, as above) the program is never behind the tree: significantly ahead with FLAIR (0.587 vs. 0.550, $p = 3.9 \times 10^{-5}$), a non-significant tie with all three modalities (0.566 vs. 0.557, $p = 0.24$)—joint-best among candidate-level methods here. On Shifts that reverses: the tree significantly beats the program directly in both settings, patient-level (0.366 vs. 0.277, $p = 1.1 \times 10^{-5}$ FLAIR; 0.249 vs. 0.231, $p = 3.5 \times 10^{-3}$ multi), not only the network—writing criteria in language transfers better specifically under a moderate shift, not universally, a narrower claim than "criteria transfer."

Is this a shift-severity phenomenon, or a dataset-specific one? MSLesSeg and Shifts are different acquisitions, not two magnitudes of one shift. We corrupt MS3SEG's own 20 test patients along two quantified axes—additive noise, a multiplicative bias field, three severities plus a zero control—and re-run every system zero-shot (Table 9, Appendix C, full numbers). Neither reproduces the pattern by magnitude: noise makes the tree's advantage *grow monotonically* with severity, significant at every non-zero level—opposite to what "moderate shift favours the program" predicts; bias field leaves the gap flat and non-significant. We read this as dataset-specific properties our corruptions miss, not shift severity as a general mechanism.

Which attributes actually drift? Per-attribute standardized Wasserstein distance from MS3SEG (real candidates), grouped as in Table 4 (Table 10, Appendix C): size/shape, the program's most

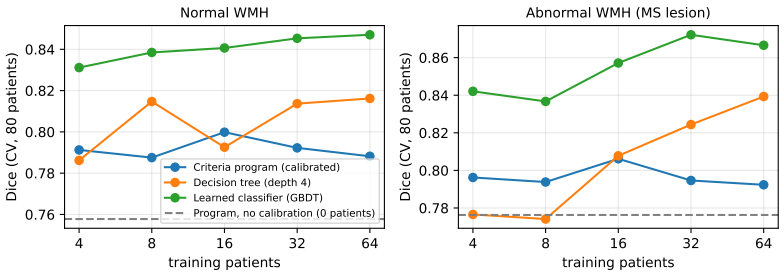


Figure 2: Out-of-fold Dice vs. annotated patients used to calibrate the program, or train the tree/-classifier, on the same 34 attributes (oracle candidates; program is the hand-written reference of Table 1). Dashed: the same program's uncalibrated default thresholds, i.e. zero annotated patients—a starting point no candidate-level learned method has (a zero-shot VLM judge needs none too, but is not candidate-level).

load-bearing group, is if anything *more* stable under Shifts than MSLesSeg (0.11 vs. 0.22)—not what a simple severity story predicts. What distinguishes Shifts is ventricle-relative orientation (3.9× MSLesSeg's drift) and multi-contrast intensity (1.8×), plausibly downstream of the perception network's own ventricle segmentation and intensity calibration differing more under a larger field-strength/vendor gap—implicating the program's orientation- and contrast-based rules, reliable in-domain (Table 5) but built on exactly the attributes that shift most where the program loses; suggestive with two datasets—tested directly next.

Testing the mechanism directly: pruning the drifting rules. This is falsifiable in one edit: delete the program's two orientation-based clauses and three contrast-based ones—an edit a clinician could make by reading it—recalibrate, and re-run Table 2. In-domain this costs nothing (CV 0.622/0.651 vs. 0.620/0.654), matching Table 4. On Shifts, where these attributes drift most, pruning significantly helps both settings (patient-level, $p \leq 2.3 \times 10^{-3}$ vs. the unpruned program): FLAIR closes 70% of the tree's gap (0.306→0.359, tree 0.382, still significant $p = 4.6 \times 10^{-4}$, was $p = 1.1 \times 10^{-5}$); on three modalities it ties the tree (0.282 vs. 0.281, $p = 0.69$), reversing a significant loss ($p = 3.5 \times 10^{-3}$). On MSLesSeg, where size/shape—not these attributes—drifts most, pruning is not simply neutral: FLAIR is unchanged, but three modalities *worsens* (0.553→0.533 patient-level, $p = 7 \times 10^{-4}$), turning a tie with the tree into a loss ($p = 0.022$). Each rule's value tracks how much *its own* attributes drift in the target, not a blanket property of written criteria: directly tested on Shifts, not merely suggestive, and no free lunch elsewhere.

4.3 Data efficiency

With only four annotated patients the calibrated program already reaches 0.791/0.796 Dice (nWMH/abWMH), statistically indistinguishable from its 64-patient value of 0.788/0.792 ($p = 0.42, 0.73$, paired Wilcoxon): at most 20 scalar thresholds against a well-behaved objective are determined by four patients' worth of candidates. This plateau does not beat an equally simple learned alternative: a depth-4, ≤16-leaf decision tree on the same attributes (Figure 2) is roughly tied with the program through 16 patients (0.793/0.808 vs. 0.800/0.806) and clearly ahead from 32 on (0.814/0.824 vs. 0.792/0.795), reaching 0.816/0.839 at 64 while remaining exactly as auditable. The GBDT classifier is stronger still at every budget (0.831/0.842 at 4, 0.847/0.867 at 64) but gives up that auditability. What the program alone has is zero: with no annotated patients its hand-written default thresholds reach 0.758/0.776 on oracle candidates (dashed line)—not on real ones (uncalibrated: 0.585/0.633, Table 1). On real candidates the program actually leads at 4 patients on the clinically relevant class (0.612/0.650 vs. the tree's 0.612/0.619), ties it by 8, and is not overtaken until 32 (Table 7)—within an afternoon of annotation. A zero-shot VLM judge needs no target-domain patients either (Section 4.4), so this advantage is not universal, only among candidate-level methods: calibration is not more data-efficient than a matched-depth tree once any data exists, but it alone starts from a non-trivial point at zero—the regime a new site is actually in before any pilot annotations.

Does a few target-domain patients recalibrate better than they refit? With a handful of target-domain patients, is it better to recalibrate the program's 20 thresholds or refit the tree/GBDT (Table 8, full numbers)? The pattern splits by dataset: regularized search does not reliably help on Shifts, but on MSLesSeg it is a clear stabilizer against volatile, often-worse unregularized recalibration—a limitation of calibration under scarcity, not evidence against Section 4.2's zero-adaptation result.

4.4 LLM PROGRAM SYNTHESIS: RELIABILITY, ENSEMBLING AND REFINEMENT

We sample 8 programs each from Qwen3-14B/32B given only the annotation protocol (temperature 0.7, no chain-of-thought); most satisfy the output contract, though satisfaction is not correctness (full numbers, Appendix C). A majority-vote ensemble over the reasonable 32B programs reaches 0.796/0.758 on oracle candidates, competitive with the hand-written one, but is the *weakest* of the three program-based methods on real candidates (Section 4.1); iterative refinement did not work reliably (Appendix B). Qwen3-VL-8B, zero-shot, reaches AUC 0.798 (abWMH vs. nWMH) on the 2,164 test candidates, below the program (0.939) and classifier (0.982).

4.5 ABLATIONS AND ANALYSIS

Which criteria matter, and to whom. Table 4 (Appendix C) removes one attribute group at a time (oracle candidates): the two methods disagree sharply—size/shape is by far the program's most load-bearing group, while the classifier is far more robust to any single removal, substituting correlated evidence for whatever is missing.

Auditability, measured rather than asserted. On the held-out test set (2,490 real, perception-derived candidates) the program agrees with the reference 70.0% overall; accuracy conditioned on which rule fired ranges from 42.6% to 82–84% (Table 5, Appendix C)—*decision agreement*, not rule precision, since CRITERIACODE's score accumulates evidence from several rules at once. The matched-interpretability tree, audited the same way, is at least as uneven (16 leaves, 19.2–93.0%, Table 11)—equally inspectable, but not clinically named. Editability is similarly measurable, on the same candidates and the candidate-level surrogate Dice of Section 3 (not Table 1's voxel Dice, hence the higher numbers): sweeping the McDonald minimum-size threshold 1–6 mm changes Dice smoothly downward (0.871/0.907 at 1 mm to 0.841/0.887 at 6 mm): the protocol's literal 3 mm minimum is never best on MS3SEG's 6 mm slices, calibration moves it to 2.72 mm, and the rule built from it is the program's least trustworthy (Table 5). We read this as a protocol-to-acquisition mismatch, not a flaw in the criterion.

Does the written protocol help the LLM? A small prompt-content ablation is mixed (Appendix C): dropping the protocol for general knowledge yields more valid programs and higher nWMH Dice, slightly lower abWMH.

LIMITATIONS

Networks and LLM synthesis run on two Turing GPUs (no bf16, Qwen3-14B/32B only). Iterative refinement (Appendix B) did not work reliably. Auditability is measured two ways, not by clinician reading, against one baseline (a tree), not several. The hand-written program, authored once by us and not blind-replicated, may itself shape some results. The few-shot experiment's regularized-recalibration variant only searches near the source-calibrated values, so it cannot discover a better threshold combination further away; its draws (both datasets) are not patient-stratified, so a budget can split one patient's timepoints across calibration and evaluation. MS3SEG is 100 patients, one scanner—this should temper every number here.

5 CONCLUSION

Distinguishing an MS lesion from a physiological hyperintensity is a criterial judgment, not a perceptual one. CRITERIACODE compiles that judgment into an auditable, editable program. In-domain, data-derived structure wins: a tree beats the program from 32 patients on. Under a moderate cross-scanner shift that reverses—the program matches the network and is never behind the tree—but a substantially larger shift reverses it again; a controlled severity sweep implicates attribute drift, not magnitude—confirmed by pruning the drifting rules—not a free lunch everywhere. CRITERIACODE is not a Dice win, but a structure a clinician can inspect and edit.

AI USE STATEMENT

We used generative AI (Anthropic's Claude via Claude Code, and locally run open-weight Qwen3 models) throughout this work. Required disclosure: Qwen3 produced the criteria programs in the proposed method (Section 3). All AI-assisted work was reviewed by the authors, who take full responsibility for its content.

ETHICS STATEMENT

All data are de-identified, publicly released research datasets (MS3SEG, CC BY 4.0; Shifts part 2, CC BY-NC-SA 4.0; MSLesSeg, CC BY 4.0) with institutional approval; we collected no new data and used them non-commercially within licence terms. The method is a research decision-support tool, not validated for clinical use; a program compiled for one annotation protocol should not be assumed valid for another without re-calibration and clinical review. MS3SEG is single-centre, single-scanner—why we report cross-scanner results rather than assume they generalise.

REPRODUCIBILITY STATEMENT

Section 4 and Appendix A–C give the data split, label corrections, network/calibration settings, attribute definitions, the LLM prompt and both the hand-written and an LLM-generated program verbatim. All experiments use open-weight models and public data. An anonymized code release accompanies this submission: the full pipeline (perception and four-class network training, candidate extraction, the reasoning stage for every compared method, LLM program synthesis and calibration, cross-scanner and few-shot target-domain evaluation, table and figure generation), every generated program (valid and invalid), and the corrected data-loading pipeline, with no author-identifying paths or metadata.

REFERENCES

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *KDD*, 2019.

Mahdi Bashiri Bawil, Mousa Shamsi, A. Ghalehasadi, Ali Fahmi Jafargholkhanloo, and Abolhassan Shakeri Bavil. A multiple sclerosis MRI dataset with tri-mask annotations for lesion segmentation. *Scientific Data*, 13:867, 2026a. doi: 10.1038/s41597-026-07184-5.

Mahdi Bashiri Bawil, Mousa Shamsi, Ali Fahmi Jafargholkhanloo, and Abolhassan Shakeri Bavil. Incorporating normal periventricular changes for enhanced pathological white matter hyperintensity segmentation: on multiclass deep learning approaches. *BioMedical Engineering OnLine*, 2026b. doi: 10.1186/s12938-026-01555-0.

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In *NeurIPS*, 2011.

Aaron Carass, Snehashis Roy, Amod Jog, Jennifer L Cuzzocreo, Elizabeth Magrath, Adrian Gherman, Julia Button, James Nguyen, Ferran Prados, Carole H Sudre, et al. Longitudinal multiple sclerosis lesion segmentation: resource and challenge. *NeuroImage*, 148:77–102, 2017.

Dongkyu Cho, Xiyue Li, Samrachana Adhikari, and Rumi Chunara. Tree of concepts: Interpretable continual learners in non-stationary clinical domains. *arXiv preprint arXiv:2604.17089*, 2026.

Olivier Commowick, Audrey Istace, Michael Kain, Baptiste Laurent, Florent Leray, Mathieu Simon, Sorina Camarasu Pop, Pascal Girard, Roxana Ameli, Jean-Christophe Ferré, et al. Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Scientific Reports*, 8(1):13650, 2018.

Ruiqi Deng, Geoffrey Martin, Tony Wang, Gongbo Zhang, Yi Liu, Chunhua Weng, Yanshan Wang, Justin F Rousseau, and Yifan Peng. Cpgprompt: Translating clinical guidelines into llm-executable decision support. *Journal of the American Medical Informatics Association*, 2026.

Franz Fazekas, John B Chawluk, Abass Alavi, Howard I Hurtig, and Robert A Zimmerman. MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging. *American Journal of Roentgenology*, 149(2):351–356, 1987.

Massimo Filippi, Maria A Rocca, Olga Ciccarelli, Nicola De Stefano, Nikos Evangelou, Ludwig Kappos, Alex Rovira, Jaume Sastre-Garriga, Mar Tintorè, Jette L Frederiksen, et al. MRI criteria for the diagnosis of multiple sclerosis: MAGNIMS consensus guidelines. *The Lancet Neurology*, 15(3):292–303, 2016.

Ludovica Griffanti, Giovanna Zamboni, Aamira Khan, Linxin Li, Guendalina Bonifacio, Vaanathi Sundaresan, Ursula G Schulz, Wilhelm Kuker, Marco Battaglini, Peter M Rothwell, and Mark Jenkinson. BIANCA (Brain intensity AbNormality classification algorithm): A new tool for automated segmentation of white matter hyperintensities. *NeuroImage*, 141:191–205, 2016.

Francesco Guarnera, Alessia Rondinella, Elena Crispino, Giulia Russo, Clara Di Lorenzo, Davide Maimone, Francesco Pappalardo, and Sebastiano Battiato. MSLesSeg: baseline and benchmarking of a new multiple sclerosis lesion segmentation dataset. *Scientific Data*, 12:920, 2025. doi: 10.1038/s41597-025-05250-y.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *CVPR*, 2023.

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. UNETR: Transformers for 3D medical image segmentation. *WACV*, 2022.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

Chongcong Jiang, Tianxingjian Ding, Chuhan Song, Jiachen Tu, Ziyang Yan, Yihua Shao, Zhenyi Wang, Yuzhang Shang, Tianyu Han, and Yu Tian. Medical SAM3: A foundation model for universal prompt-driven medical image segmentation. *arXiv preprint arXiv:2601.10880*, 2026.

Hugo J Kuijf, J Matthijs Biesbroek, Jeroen De Bresser, Rutger Heinen, Simon Andermatt, Mariana Bento, Matt Berseth, Mikhail Belyaev, M Jorge Cardoso, Adria Casamitjana, et al. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the WMH segmentation challenge. *IEEE Transactions on Medical Imaging*, 38(11):2556–2568, 2019.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning segmentation via large language model. In *CVPR*, 2024.

Žiga Lesjak, Alfiia Galimzianova, Aleš Koren, Matej Lukin, Franjo Pernuš, Žiga Špiclin, and Boštjan Likar. A novel public MR image dataset of multiple sclerosis patients with lesion segmentations based on multi-rater consensus. *Neuroinformatics*, 16:51–63, 2018.

Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. CLIP-driven universal model for organ segmentation and tumor detection. In *ICCV*, 2023.

Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-Zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15:654, 2024.

Andrey Malinin, Andreas Athanasopoulos, Muhamed Barakovic, Meritxell Bach Cuadra, Mark JF Gales, Cristina Granziera, Mara Graziani, Nikolay Kartashev, Konstantinos Kyriakopoulos, Po-Jui Lu, et al. Shifts 2.0: Extending the dataset of real distributional shifts. *arXiv preprint arXiv:2206.15407*, 2022.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pp. 234–241, 2015.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Silas Ruhrberg Estévez, Christopher Chiu, and Mihaela van der Schaar. Automatic construction of clinical scoring systems with LLM agents. *arXiv preprint arXiv:2601.22324*, 2026.

Martijn D Steenwijk, Petra JW Pouwels, Marita Daams, Jan W van Dalen, Matthan WA Caan, Edo Richard, Frederik Barkhof, and Hugo Vrenken. Accurate white matter lesion segmentation by k nearest neighbor classification with tissue type priors (kNN-TTPs). *NeuroImage: Clinical*, 3: 462–469, 2013.

Carole H Sudre, M Jorge Cardoso, Willem H Bouvy, Geert Jan Biessels, Josephine Barnes, and Sebastien Ourselin. Bayesian model selection for pathological neuroimaging data applied to white matter lesion segmentation. *IEEE Transactions on Medical Imaging*, 34(10):2079–2102, 2015.

Dídac Surís, Sachit Menon, and Carl Vondrick. ViperGPT: Visual inference via Python execution for reasoning. In *ICCV*, 2023.

Idan Tankel, Nir Mazor, Rafi Brada, Christina LeBedis, and Guy ben Yosef. INFORM-CT: Integrating LLMs and VLMs for incidental findings management in abdominal CT. In *Medical Imaging with Deep Learning (MIDL)*, 2026. arXiv:2512.14732.

Alan J Thompson, Brenda L Banwell, Frederik Barkhof, William M Carroll, Timothy Coetzee, Giancarlo Comi, Jorge Correale, Franz Fazekas, Massimo Filippi, Mark S Freedman, et al. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *The Lancet Neurology*, 17 (2):162–173, 2018.

Quoc-Huy Trinh, Minh-Van Nguyen, Jun Zeng, Debesh Jha, and Ulas Bagci. PRS-Med: Position reasoning segmentation in medical imaging. *arXiv preprint arXiv:2505.11872*, 2025.

Midhat Urooj, Ayan Banerjee, and Sandeep Gupta. Neuro-guard: Neuro-symbolic generalization and unbiased adaptive routing for diagnostics—explainable medical ai. *arXiv preprint arXiv:2512.18177*, 2025.

Ruochen Wang, Si Si, Felix Yu, Dorothea Wiesmann, Cho-Jui Hsieh, and Inderjit Dhillon. Large language models are interpretable learners. In *ICLR*, 2025.

Yingchun Wang, Jingcai Guo, Song Guo, Weizhan Zhang, and Jie Zhang. Exploring optimal substructure for out-of-distribution generalization via feature-targeted model pruning. *arXiv preprint arXiv:2212.09458*, 2022.

Joanna M Wardlaw, Eric E Smith, Geert J Biessels, Charlotte Cordonnier, Franz Fazekas, Richard Frayne, Richard I Lindley, John T O'Brien, Frederik Barkhof, Oscar R Benavente, et al. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *The Lancet Neurology*, 12(8):822–838, 2013.

Yang Xing, Jiong Wu, Savas Ozdemir, Ying Zhang, Yang Yang, Wei Shao, and Kuang Gong. MedVL-SAM2: A unified 3D medical vision-language model for multimodal reasoning and prompt-driven segmentation. *arXiv preprint arXiv:2601.09879*, 2026.

Wei Xu, Ke Yang, Gang Luo, Keli Zheng, Lingyan Hu, Jing Wang, and Kefeng Li. Medical heuristic learning: An LLM-driven framework for interpretable and auditable clinical decision rules. *arXiv preprint arXiv:2606.16337*, 2026.

Zhonghao Yan, Muxi Diao, Yuxuan Yang, Ruoyan Jing, Jiayuan Xu, Kaizhou Zhang, Lele Yang, Yanxi Liu, Kongming Liang, and Zhanyu Ma. MedReasoner: Reinforcement learning drives reasoning grounding from clinical thought to pixel-level precision. *arXiv preprint arXiv:2508.08177*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Theodore Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Sid Kiblawi, Tristan Naumann, Jianfeng Gao, Angela Crabtree, Jacob Abel, et al. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nature Methods*, 22: 166–176, 2025.

Ziheng Zhao, Yao Zhang, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. One model to rule them all: Towards universal segmentation for medical images with text prompts. *arXiv preprint arXiv:2312.17183*, 2023.

A Data preparation

MS3SEG release issues. We load the four-class label volumes (`man_4L_masks_new`, codes 64/191/255 for ventricle/normal/abnormal WMH; several files carry a NIfTI scale factor of 257 that must not be applied to label codes). For seven patients (001–004, 075, 082, 087) the label header is RAS while the FLAIR header is LAS, i.e. the stored label array is mirrored left–right relative to the image; applied naively, ventricle labels fall on tissue with mean FLAIR z-score ≈ 0 instead of ≈ -1.5. We reorient labels to the image axes by header and verify every patient with an intensity test (ventricle labels must be dark and WMH labels bright on FLAIR, and the reoriented label must beat its mirror image by a clear margin). Patient 056's four-class file uses different codes and contains no lesion class; we recompose its label from the released per-class masks (lesion $>$ normal WMH $>$ ventricle where they overlap). After these corrections our label maps agree with the colour overlays distributed with the dataset. The baselines in the dataset paper were trained on the released arrays directly, so seven of their training/validation cases likely contained mirrored labels; this does not affect our comparison because we retrain all baselines.

Preprocessing. Each contrast is clipped to its 0.5–99.5 percentile range inside a brain mask (the released BET mask dilated by three voxels so that juxtacortical annotations are retained; 0.004% of labelled voxels fall outside the undilated mask), z-scored inside the mask and set to zero outside.

External datasets. Shifts part 2 and MSLesSeg provide skull-stripped, 1 mm isotropic FLAIR/T1/T2 with lesion masks. We reorient to LAS, resample in-plane to 0.898 mm, average six consecutive slices into 6 mm slabs (a slab voxel is a lesion if at least half of its six slices are), crop or pad to 256×256 around the brain centroid, and apply the same normalisation. For MSLesSeg every available time point is treated as a separate scan.

B Implementation details

Networks. MONAI `DynUNet` in 2D with kernel size 3, strides (1,2,2,2,2,2), channels (32,64,128,256,320,320), residual blocks, instance normalisation, two deep-supervision heads (weights 1, 0.5, 0.25), Dice+cross-entropy over foreground classes, AdamW (learning rate 5×10^{-4}, weight decay 3×10^{-5}, polynomial decay with exponent 0.9), batch size 16, 200 epochs, mixed precision, gradient clipping at 12. Augmentation: left–right flips, rotations up to 15°, zoom 0.9–1.1, Gaussian noise, intensity scaling/shift and gamma. Model selection uses the mean foreground Dice on the validation fold; inference uses flip test-time augmentation and, on the held-out test set, the average of the five fold models. All models train on the same slices (all slices of the training patients, including slices without annotations).

Candidate extraction. Hyperintensity candidates are 8-connected components of the predicted hyperintensity class on each slice; 3D context attributes use 26-connectivity across slices. Distances use the header voxel spacing. Attributes involving the ventricles use the *predicted* ventricle mask, so the pipeline needs no anatomical annotations at test time.

Iterative refinement (attempted, not a reported component). We prompted the LLM with its own program, its calibrated thresholds, its area-weighted training Dice, and up to 16 concrete misclassified candidates (attribute values, the rule that fired, the reference label), and asked for a revision under the same output contract. With thinking disabled, single-attempt revisions were syntactically valid but—checked by diffing against the parent—in two of two completed cases textually identical to the input program modulo whitespace: the model did not use the feedback to change the decision logic. Enabling thinking reproduces the token-budget failure of Section 4.4; the longer feedback-augmented prompt also exhausted GPU memory (up to 47 GB per device on two Quadro RTX 8000 GPUs) after one or two successful revisions regardless of thinking mode, most likely from attention memory scaling with the several-fold longer prompt on an attention kernel that does not consistently take a memory-efficient path for this shape. We did not solve this within our timeline and report it as an open problem for making criteria refinement practical, not as a working part of CRITERIACODE.

C ADDITIONAL RESULTS

Per-fold breakdowns (all five folds individually, not just the mean±SD reported in the main text), per-class HD95 for every system and dataset, and the full manifest of every synthesised program (valid and invalid, with the validation error where applicable) are included with the released code rather than reproduced here.

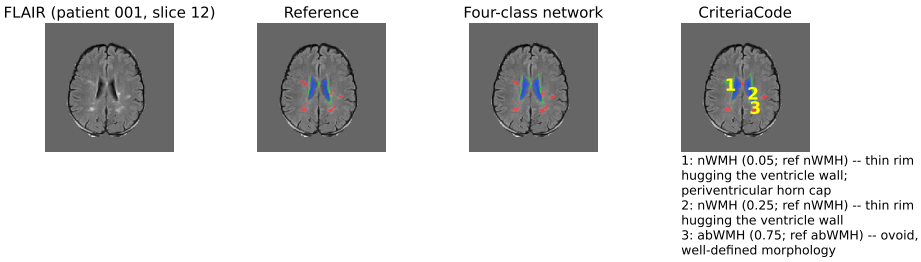


Figure 3: A test patient (real candidates; system = "Program (calibrated)", Table 1), with CRITERIACODE's decision, score and fired rule per candidate: the two rims are correctly separated from the lesion with a checkable reason.

LLM contract satisfaction, full numbers (Section 4.4). Of 8 samples each, 6/8 (Qwen3-14B) and 5/8 (Qwen3-32B) satisfy the output contract; failures are clean validation errors, discarded automatically. Contract satisfaction is not correctness: 14B programs range from 0.684/0.638 down to near-degenerate; 32B is more consistently reasonable (0.350–0.658/0.531–0.771 oracle-candidate Dice).

Criteria-group ablation, full numbers (Section 4.5). For the program, size/shape is by far the most load-bearing group (Dice 0.788/0.792 $\rightarrow$ 0.584/0.499, $\Delta{=}-0.249$, since its rules reference diameter, thickness and elongation directly), location is second ($\Delta{=}-0.078$), orientation/symmetry/contrast contribute little alone ($|\Delta| \leq 0.02$), and context changes nothing. The classifier's worst case is $\Delta{=}-0.022$: with 33 attributes remaining it substitutes correlated evidence for whatever is missing—the flexibility a rule set trades for auditability.

Prompt-content ablation, full numbers (Section 4.5). Qwen3-32B, 8 samples/condition: dropping the written protocol for general clinical knowledge alone yields more valid programs (8/8 vs. 5/8) and higher nWMH Dice (0.661 vs. 0.609) but slightly lower abWMH (0.629 vs. 0.643)—suggestive the protocol's detail encourages the more bug-prone rules of Section 4.4, not conclusive given the small, mixed sample.

Rule-level decision agreement on the held-out test set (Section 4.5). Table 5 gives, for the calibrated hand-written program on the 2,490 test candidates (three-way decision, abWMH/nWMH/background), how often each rule fires and how often the *final* decision (which can combine several fired rules) agrees with the reference label when it does; this is not the isolated correctness

Table 3: The 34 candidate attributes computed for every hyperintensity component. All are deterministic functions of the candidate mask, the ventricle and brain masks and the z-scored FLAIR/T1/T2 slices.

Attribute	Unit	Meaning
Size and shape		
area_mm2	mm^2	in-plane area of the hyperintense candidate on this slice
equiv_diam_mm	mm	diameter of the circle with the same area
major_axis_mm	mm	length of the major axis of the best-fitting ellipse
minor_axis_mm	mm	length of the minor axis of the best-fitting ellipse
elongation	ratio ≥ 1	major_axis / minor_axis; 1 = round, large = thin/elongated
solidity	0-1	area / convex-hull area; low values = irregular or branching shape
mean_thickness_mm	mm	twice the mean distance of candidate pixels to the candidate boundary (thin rims are small)
max_inscribed_diam_mm	mm	diameter of the largest disc that fits inside the candidate
Location		
dist_to_ventricle_mm	mm	3D distance from the candidate to the nearest ventricle voxel; 0 = touching the ventricle
frac_boundary_near_ventricle	0-1	fraction of the candidate boundary within 2 pixels of the ventricle (1 = rim hugging the ventricle wall)
dist_to_anterior_horn_mm	mm	in-plane distance to the anterior (frontal) horn region of the lateral ventricles on this slice
dist_to_posterior_horn_mm	mm	in-plane distance to the posterior (occipital) horn region of the lateral ventricles on this slice
dist_to_brain_edge_mm	mm	in-plane distance to the brain surface (small = juxtacortical / subcortical location)
rel_slice	0-1	relative slice position (0 = most inferior, 1 = most superior slice)
rel_ap	-1..1	centroid position along the posterior(-1) to anterior(+1) axis within the brain bounding box
abs_rel_lr	0-1	absolute centroid distance from the brain midline, 0 = midline, 1 = lateral brain edge
Morphology vs. ventricle		
orientation_vs_ventricle_deg	degrees 0-90	angle between the candidate major axis and the direction pointing away from the ventricle wall; $\sim$0 = elongated perpendicular to the ventricle (Dawson's finger), $\sim$90 = elongated along the ventricle wall (rim/cap)
Symmetry		
sym_overlap_frac	0-1	fraction of the candidate whose mirror image across the brain midline also lies on a hyperintensity (bilateral symmetric changes are typical of normal/age-related WMH)
sym_partner_area_ratio	ratio	area of the mirrored partner hyperintensity divided by this candidate's area; NaN if no partner
Multi-contrast		
flair_z_mean	z-score	mean FLAIR intensity inside the candidate, z-scored within the brain
flair_z_max	z-score	maximum FLAIR intensity inside the candidate
flair_contrast	z-score difference	mean FLAIR inside minus mean FLAIR in a 3-pixel ring of surrounding normal tissue
t2_z_mean	z-score	mean T2-weighted intensity inside the candidate
t2_contrast	z-score difference	T2 inside minus surrounding ring; MS lesions are also hyperintense on T2
t1_z_mean	z-score	mean T1-weighted intensity inside the candidate
t1_contrast	z-score difference	T1 inside minus surrounding ring; negative = T1-hypointense ('black hole', typical of chronic MS lesions)
Context		
overlap_prev_slice	0-1	fraction of the candidate overlapping a hyperintensity on the previous (inferior) slice
overlap_next_slice	0-1	fraction of the candidate overlapping a hyperintensity on the next (superior) slice
volume_3d_mm3	mm^3	volume of the 3D connected hyperintensity containing this candidate
n_slices_3d	count	number of slices spanned by that 3D connected hyperintensity
ventricle_area_slice_mm2	mm^2	ventricle area on this slice (large ventricles = atrophy / older brain)
n_candidates_slice	count	number of hyperintense candidates on this slice
patient_wmh_load_mm2	mm^2	total hyperintense area over all slices of the patient (lesion load)
patient_n_candidates	count	total number of hyperintense candidates in the patient

Table 4: Criteria-group ablation on oracle candidates: out-of-fold Dice (nWMH / abWMH) when an attribute group is removed (missing for the program, dropped for the classifier). Δ is the change in mean Dice over the two classes.

	Criteria program			Learned classifier		
Removed group	Dice nWMH	Dice abWMH	Δ	Dice nWMH	Dice abWMH	Δ
None (all 34 attributes)	0.788	0.792	–	0.847	0.867	–
Location (ventricle/horn/cortex distances, position)	0.708	0.716	-0.078	0.838	0.849	-0.014
Size and shape	0.584	0.499	-0.249	0.827	0.842	-0.022
Morphology vs. ventricle (orientation)	0.799	0.805	+0.012	0.847	0.866	-0.000
Symmetry (bilateral)	0.801	0.815	+0.018	0.848	0.868	+0.001
Multi-contrast (FLAIR/T2/T1 intensity, contrast)	0.795	0.805	+0.010	0.848	0.867	+0.001
Context (slice/3D/patient)	0.788	0.792	+0.000	0.848	0.870	+0.002

Table 5: Decision agreement of the calibrated hand-written program on the held-out test set (2,490 candidates, three-way decision): the fraction of candidates on which the given rule fired *and* the final decision agreed with the reference label, among all candidates on which it fired (n). Not rule-isolated precision (see text).

Rule	n	Decision agreement
Ovoid, well-defined morphology	1713	0.690
Thin rim hugging the ventricle wall	1174	0.725
Periventricular horn cap	1149	0.687
Below 3 mm minimum lesion size	835	0.426
Bilateral symmetric small change	472	0.718
T2 hyperintensity confirmed	48	0.750
T1 hypointense (black hole)	45	0.600
Weak FLAIR contrast (background)	34	0.618
Periventricular, perpendicular (Dawson's finger)	11	0.818
Juxtacortical location	5	0.600
(no rule fired)	100	0.840

of that one rule's own directional claim, since CRITERIACODE's score is additive over every rule that fires. Table 6 gives the FLAIR-only and confound-controlled rows omitted from Table 1, and Table 7 the real-candidate data-efficiency numbers omitted from Figure 2.

LLM prompt (protocol variant). The exact prompt given to Qwen3-14B and Qwen3-32B for the main results (Sections 4.1, 4.4); the "generic" and "data-informed" variants add or remove text as described in Section 3 and were not used in the reported main results.

```
You are an expert neuroradiologist and Python programmer. An automatic detector has found
     every hyperintense region ("candidate") on axial T2-FLAIR brain MRI slices of patients
     with multiple sclerosis (MS), together with the ventricles. For each candidate we
     computed the numeric attributes listed below. Your job is to write an executable,
     interpretable Python program that decides, for one candidate at a time, whether it is:
- "abWMH": an abnormal white matter hyperintensity, i.e. an MS demyelinating lesion,
- "nWMH": a normal / physiological hyperintensity (CSF-contaminated periventricular rim or cap
     , age-related symmetric small-vessel change), or
- "background": not a genuine hyperintensity at all (detector false positive: e.g. negligible
     contrast).

The program must encode the clinical criteria as explicit, readable rules over the attributes.
      Every numeric threshold must be stored in a dictionary named PARAMS (with a short
     comment on its meaning and a clinically plausible default value) so that thresholds can
     later be calibrated on data; the rules themselves must reference PARAMS rather than
     literal numbers. Also expose PARAMS["abnormal_threshold"], the score cut-off for calling
     a candidate "abWMH".

ANNOTATION PROTOCOL (MS3SEG dataset; axial T2-FLAIR, 1.5 T; classes annotated by a trained
     specialist and reviewed by a senior neuroradiologist)

Class definitions:
- background: non-brain tissue and normal brain parenchyma.
- ventricles: CSF-filled ventricular spaces; very low (hypointense) signal on FLAIR.
- normal WMH (physiological hyperintensity): age-related hyperintensities or normal CSF-
     contaminated regions in periventricular areas. Hyperintense but non-pathological, and
     excluded from the MS lesion burden. Identified as (a) thin periventricular rims/caps in
     the anterior and posterior horns of the lateral ventricles, where CSF contamination (
     partial volume with ventricular CSF, "ependymitis granularis") is physiologically normal,
      or (b) bilateral, symmetric deep white matter changes consistent with age-related small-
     vessel disease.
```

Table 6: Rows omitted from Table 1 for space: the decomposition with a FLAIR-only perception network, and the confound-controlled experiment of Section 4.1, in which candidates are re-extracted from the four-class network's own predicted hyperintense region so that every candidate-level method is scored on exactly the pixels that network chose (both input settings). Dice, mean±SD over patients; the network's own native decision on those pixels is repeated in each block for reference.

	Cross-validation (80 patients)		Held-out test (20 patients)	
System	Dice nWMH	Dice abWMH	Dice nWMH	Dice abWMH
Decomposition with a FLAIR-only perception network (cf. the FLAIR+T1+T2 rows of Table 1)				
Four-class net (FLAIR), for reference	0.722±0.06	0.744±0.16	0.744±0.08	0.784±0.07
Oracle (perfect candidate classification)	0.710±0.09	0.761±0.15	0.746±0.10	0.807±0.07
Program (no calibration)	0.600±0.11	0.640±0.17	0.645±0.14	0.670±0.11
Program (calibrated)	0.625±0.11	0.655±0.17	0.641±0.14	0.692±0.10
Decision tree (depth 4, same attributes)	0.635±0.10	0.671±0.17	0.676±0.11	0.697±0.08
LLM-program ensemble (calibrated)	0.606±0.12	0.622±0.17	0.619±0.16	0.661±0.11
Learned classifier (GBDT)	0.648±0.12	0.693±0.17	0.691±0.13	0.747±0.08
Confound-controlled: candidates re-extracted from the four-class net's own region (FLAIR)				
Four-class net (FLAIR), native decision on the same pixels	0.722±0.06	0.744±0.16	0.744±0.08	0.784±0.07
Oracle (perfect candidate classification)	0.715±0.10	0.763±0.14	0.753±0.10	0.807±0.08
Program (calibrated)	0.634±0.13	0.669±0.17	0.679±0.13	0.693±0.10
Learned classifier (GBDT)	0.656±0.12	0.699±0.16	0.699±0.13	0.751±0.08
Confound-controlled: candidates re-extracted from the four-class net's own region (FLAIR+T1+T2)				
Four-class net (FLAIR+T1+T2), native decision on the same pixels	0.713±0.06	0.743±0.15	0.744±0.08	0.793±0.07
Oracle (perfect candidate classification)	0.707±0.10	0.763±0.14	0.747±0.10	0.812±0.08
Program (calibrated)	0.632±0.11	0.666±0.16	0.675±0.13	0.715±0.10
Learned classifier (GBDT)	0.649±0.11	0.697±0.16	0.695±0.13	0.756±0.09

Table 7: Data efficiency on real, perception-derived candidates (agnostic_multi; cf. Figure 2's oracle-candidate curve and Table 1's $n = 64$ row). Out-of-fold Dice (nWMH / abWMH), mean over the 80 cross-validation patients.

System	$n = 4$	$n = 8$	$n = 16$	$n = 32$	$n = 64$
Program, calibrated	0.612/0.650	0.615/0.643	0.610/0.659	0.616/0.652	0.620/0.654
Decision tree (depth 4)	0.612/0.619	0.610/0.643	0.600/0.649	0.625/0.667	0.628/0.672

```
- abnormal WMH (MS lesion): MS-related demyelinating lesions with characteristic imaging and
    spatial features. Requires high T2-FLAIR signal combined with MS-characteristic features:
     periventricular or juxtacortical location, ovoid morphology oriented perpendicular to
    the ventricles (Dawson's fingers), a minimum diameter of about 3 mm, and confirmation of
    hyperintensity on T2-weighted images.

Additional radiological knowledge that may be used (McDonald 2017/2024, MAGNIMS consensus):
- Typical MS lesions are ovoid/round, well-defined, >= 3 mm, located periventricularly (
    abutting the ventricle but extending into the white matter, often perpendicular to the
    ventricular wall along medullary veins), juxtacortically (touching the cortex),
    infratentorially, or in the deep white matter; they are usually asymmetric between
    hemispheres; chronic lesions can be T1-hypointense ("black holes").
- Normal/physiological hyperintensities are thin, smooth, linear rims or caps hugging the
    ventricular wall (especially frontal and occipital horns), following the wall contour
    rather than extending perpendicularly into the white matter, and tend to be bilateral and
     symmetric. Age-related small-vessel changes are punctate or patchy, bilateral, symmetric
    , in deep or periventricular white matter, without the ovoid perpendicular morphology.


Attribute schema (name: unit -- meaning):
- area_mm2: mm^2 -- in-plane area of the hyperintense candidate on this slice
- equiv_diam_mm: mm -- diameter of the circle with the same area
- major_axis_mm: mm -- length of the major axis of the best-fitting ellipse
- minor_axis_mm: mm -- length of the minor axis of the best-fitting ellipse
- elongation: ratio >= 1 -- major_axis / minor_axis; 1 = round, large = thin/elongated
- solidity: 0-1 -- area / convex-hull area; low values = irregular or branching shape
- mean_thickness_mm: mm -- twice the mean distance of candidate pixels to the candidate
    boundary (thin rims are small)
- max_inscribed_diam_mm: mm -- diameter of the largest disc that fits inside the candidate
- dist_to_ventricle_mm: mm -- 3D distance from the candidate to the nearest ventricle voxel; 0
     = touching the ventricle
- frac_boundary_near_ventricle: 0-1 -- fraction of the candidate boundary within 2 pixels of
    the ventricle (1 = rim hugging the ventricle wall)
```

Table 8: Few-shot target-domain adaptation (FLAIR+T1+T2): the program's 20 thresholds recalibrated, or the tree/GBDT refit, on a handful of target-domain scans (2/4/8, five random draws each), then evaluated on every remaining scan in that dataset; draws are not patient-stratified (Limitations). "Regularized" searches only within 1.5× of the source-calibrated value per threshold instead of the from-scratch range, testing whether an unregularized search on little data explains the pattern. Lesion Dice, mean over draws and evaluation patients; "0" is the zero-adaptation row of Table 2 (multi).

Dataset	System	0	2	4	8
Shifts (severe shift)	Program, recalibrated	0.231	0.262	0.245	0.252
	Program, regularized	0.231	0.243	0.258	0.241
	Tree, refit	0.249	0.235	0.246	0.239
	GBDT, refit	0.240	0.262	0.283	0.259
MSLesSeg (moderate shift)	Program, recalibrated	0.566	0.399	0.506	0.497
	Program, regularized	0.566	0.555	0.541	0.551
	Tree, refit	0.557	0.490	0.537	0.548
	GBDT, refit	0.559	0.544	0.538	0.558

Table 9: Controlled shift-severity sweep: MS3SEG's own 20 official test patients, corrupted by an independent, quantified amount (population/scanner/protocol/annotation held fixed; severity 0 is the shared, unmodified control), evaluated zero-shot with every already-trained/calibrated system. Lesion Dice, mean over patients. ‡: program vs. tree paired Wilcoxon signed-rank $p < 0.05$ at that severity; n.s. at every bias-field level.

Axis	Severity	Net	No-reasoning	Program	Tree	GBDT
noise	0.00	0.793	0.493	0.687	0.700	0.740
	0.15	0.755	0.485	0.656‡	0.684	0.716
	0.30	0.650	0.464	0.588‡	0.618	0.650
	0.50	0.477	0.401	0.441‡	0.481	0.495
bias	0.00	0.793	0.493	0.687	0.700	0.740
	0.15	0.798	0.489	0.678	0.687	0.744
	0.30	0.779	0.481	0.667	0.679	0.734
	0.50	0.681	0.456	0.610	0.618	0.644

```
- orientation_vs_ventricle_deg: degrees 0-90 -- angle between the candidate major axis and the
     direction pointing away from the ventricle wall; ~0 = elongated perpendicular to the
    ventricle (Dawson's finger), ~90 = elongated along the ventricle wall (rim/cap)
- dist_to_anterior_horn_mm: mm -- in-plane distance to the anterior (frontal) horn region of
    the lateral ventricles on this slice
- dist_to_posterior_horn_mm: mm -- in-plane distance to the posterior (occipital) horn region
    of the lateral ventricles on this slice
- dist_to_brain_edge_mm: mm -- in-plane distance to the brain surface (small = juxtacortical /
     subcortical location)
- flair_z_mean: z-score -- mean FLAIR intensity inside the candidate, z-scored within the
    brain
- flair_z_max: z-score -- maximum FLAIR intensity inside the candidate
- flair_contrast: z-score difference -- mean FLAIR inside minus mean FLAIR in a 3-pixel ring
    of surrounding normal tissue
- t2_z_mean: z-score -- mean T2-weighted intensity inside the candidate
- t2_contrast: z-score difference -- T2 inside minus surrounding ring; MS lesions are also
    hyperintense on T2
- t1_z_mean: z-score -- mean T1-weighted intensity inside the candidate
- t1_contrast: z-score difference -- T1 inside minus surrounding ring; negative = T1-
    hypointense ('black hole', typical of chronic MS lesions)
- sym_overlap_frac: 0-1 -- fraction of the candidate whose mirror image across the brain
    midline also lies on a hyperintensity (bilateral symmetric changes are typical of normal/
    age-related WMH)
- sym_partner_area_ratio: ratio -- area of the mirrored partner hyperintensity divided by this
     candidate's area; NaN if no partner
- overlap_prev_slice: 0-1 -- fraction of the candidate overlapping a hyperintensity on the
    previous (inferior) slice
- overlap_next_slice: 0-1 -- fraction of the candidate overlapping a hyperintensity on the
    next (superior) slice
- volume_3d_mm3: mm^3 -- volume of the 3D connected hyperintensity containing this candidate
- n_slices_3d: count -- number of slices spanned by that 3D connected hyperintensity
- rel_slice: 0-1 -- relative slice position (0 = most inferior, 1 = most superior slice)
- rel_ap: -1..1 -- centroid position along the posterior(-1) to anterior(+1) axis within the
    brain bounding box
```

Table 10: Attribute drift: standardized Wasserstein distance (candidate-level, real perception-derived candidates; MS3SEG's own distribution as reference, distance divided by MS3SEG's own SD) between MS3SEG and each external dataset, averaged within each of Table 4's six attribute groups. Larger = more shifted.

Attribute group	MSLesSeg	Shifts	Shifts / MSLesSeg
Size and shape	0.218	0.112	0.51×
Location	0.268	0.217	0.81×
Orientation	0.058	0.225	3.88×
Symmetry	0.099	0.061	0.61×
Multi-contrast	0.446	0.782	1.75×
Context	0.467	0.671	1.44×
Overall (34 attrs)	0.324	0.407	1.26×

Table 11: Tree auditability, measured the same way as Table 5: the depth-4 tree of Table 1 (agnostic_multi, seed 0) on the same 2,490 held-out real test candidates, grouped by which of its 16 leaves each candidate reaches. Decision agreement = fraction of candidates in that leaf where the tree's own prediction matches the reference; splits are the tree's, not clinically named. Five largest leaves by candidate count shown; full 16 in the released code.

Leaf split path (root to leaf)	n	Decision agreement
`rim_frac > 0.42, area > 3.63, thickness <= 2.19, diam > 2.95`	614	0.930
`rim_frac <= 0.42, thickness > 1.92, elongation <= 3.82, thickness > 2.09`	517	0.913
`rim_frac <= 0.42, thickness > 1.92, elongation <= 3.82, thickness <= 2.09`	233	0.721
`rim_frac > 0.42, area <= 3.63, area <= 2.02, edge_dist > 16.50`	227	0.665
`rim_frac > 0.42, area > 3.63, thickness <= 2.19, diam <= 2.95`	191	0.749

```
- abs_rel_lr: 0-1 -- absolute centroid distance from the brain midline, 0 = midline, 1 =
     lateral brain edge
- ventricle_area_slice_mm2: mm^2 -- ventricle area on this slice (large ventricles = atrophy /
     older brain)
- n_candidates_slice: count -- number of hyperintense candidates on this slice
- patient_wmh_load_mm2: mm^2 -- total hyperintense area over all slices of the patient (lesion
     load)
- patient_n_candidates: count -- total number of hyperintense candidates in the patient

Output requirements (strict):
1. Return exactly one Python code block. It must define:
   PARAMS: dict[str, float]
   def classify(c: dict, P: dict = None) -> dict
   where c maps attribute names to floats (NaN when unavailable) and P defaults to PARAMS when
     None.
2. classify must return {"decision": one of "abWMH" | "nWMH" | "background", "score_abnormal":
     float in [0, 1], "reasons": list[str]}.
   score_abnormal must be a graded evidence score for "abWMH" (higher = more lesion-like), and
     decision must be "abWMH" exactly when score_abnormal >= P["abnormal_threshold"] and the
     candidate is not rejected as "background".
3. Use only the Python standard library `math`. No other imports, no I/O, no randomness, no
     global state besides PARAMS.
4. Handle NaN attributes gracefully with math.isnan (treat missing evidence as neutral, never
     crash).
5. Use only attributes from the schema; do not invent attributes.
6. Keep PARAMS to at most 20 entries; reasons should name the rules that fired.
Combine location (periventricular vs juxtacortical vs deep), morphology (thin rim hugging the
     ventricle vs ovoid/perpendicular), size (about 3 mm minimum for a lesion), symmetry, and
     T2/T1 confirmation into the score, grounded in whatever clinical criteria were given to
     you above.
```

The hand-written reference program. The program used for every headline in-domain and cross-scanner result in the main text (Tables 1 and 2, Figures 2 and 3), authored by us from the protocol and attribute schema before any LLM output was seen or any real candidate existed, shown here with its uncalibrated defaults; `PARAMS` values are replaced by calibration but the rule logic is not.

```python
import math

# Hand-written baseline rules, authored from the annotation protocol before any LLM program
    was generated.
PARAMS = {
    "min_contrast": 0.3,            # FLAIR contrast below this = not a genuine hyperintensity
```

```
    "rim_boundary_frac": 0.5,       # fraction of boundary on the ventricle wall to count as a
      rim
    "rim_thickness_mm": 2.5,        # rims/caps are thin
    "rim_orientation_deg": 50.0,    # elongated along the wall (large angle) = rim,
     perpendicular (small) = Dawson's finger
    "horn_dist_mm": 3.0,            # distance to frontal/occipital horn to count as a cap
    "cap_thickness_mm": 3.0,
    "sym_frac": 0.3,                # mirrored overlap above this = bilateral symmetric change
    "sym_max_diam_mm": 5.0,
    "min_lesion_diam_mm": 3.0,      # McDonald minimum lesion size
    "periventricular_mm": 3.0,
    "juxtacortical_mm": 3.0,
    "perpendicular_deg": 45.0,
    "ovoid_min_elong": 1.2,
    "ovoid_max_elong": 3.0,
    "min_solidity": 0.8,
    "t1_hypo": -0.3,                # T1 hypointensity relative to surroundings
    "t2_hyper": 0.5,                # T2 confirmation
    "abnormal_threshold": 0.5,
}


def _nan(x):
    return x is None or (isinstance(x, float) and math.isnan(x))


def classify(c, P=None):
    P = PARAMS if P is None else P
    reasons = []
    if not _nan(c.get("flair_contrast")) and c["flair_contrast"] < P["min_contrast"]:
        return {"decision": "background", "score_abnormal": 0.0, "reasons": ["weak␣FLAIR␣
     contrast"]}
    score = 0.5
    diam = c["equiv_diam_mm"]
    thick = c["mean_thickness_mm"]
    dv = c["dist_to_ventricle_mm"]
    ang = c["orientation_vs_ventricle_deg"]
    # physiological patterns
    if c["frac_boundary_near_ventricle"] >= P["rim_boundary_frac"] and thick <= P["
     rim_thickness_mm"] and (_nan(ang) or ang >= P["rim_orientation_deg"]):
        score -= 0.35
        reasons.append("thin␣rim␣hugging␣the␣ventricle␣wall")
    for h in ("dist_to_anterior_horn_mm", "dist_to_posterior_horn_mm"):
        if not _nan(c.get(h)) and c[h] <= P["horn_dist_mm"] and thick <= P["cap_thickness_mm"
     ]:
            score -= 0.2
            reasons.append("periventricular␣horn␣cap")
            break
    if c["sym_overlap_frac"] >= P["sym_frac"] and diam <= P["sym_max_diam_mm"]:
        score -= 0.2
        reasons.append("bilateral␣symmetric␣small␣change")
    if diam < P["min_lesion_diam_mm"]:
        score -= 0.15
        reasons.append("below␣3␣mm␣minimum␣lesion␣size")
    else:
        score += 0.1
    # MS-typical patterns
    if not _nan(dv) and dv <= P["periventricular_mm"] and not _nan(ang) and ang <= P["
     perpendicular_deg"] and thick > P["rim_thickness_mm"]:
        score += 0.3
        reasons.append("periventricular,␣perpendicular␣to␣ventricle␣(Dawson's␣finger)")
    if c["dist_to_brain_edge_mm"] <= P["juxtacortical_mm"]:
        score += 0.2
        reasons.append("juxtacortical␣location")
    if P["ovoid_min_elong"] <= c["elongation"] <= P["ovoid_max_elong"] and c["solidity"] >= P[
     "min_solidity"]:
        score += 0.15
        reasons.append("ovoid,␣well-defined␣morphology")
    if not _nan(c.get("t1_contrast")) and c["t1_contrast"] <= P["t1_hypo"]:
        score += 0.15
        reasons.append("T1␣hypointense␣(black␣hole)")
    if not _nan(c.get("t2_contrast")) and c["t2_contrast"] >= P["t2_hyper"]:
        score += 0.1
        reasons.append("T2␣hyperintensity␣confirmed")
    score = min(1.0, max(0.0, score))
    decision = "abWMH" if score >= P["abnormal_threshold"] else "nWMH"
    return {"decision": decision, "score_abnormal": score, "reasons": reasons}
```

Example LLM-generated program. One of the reasonable (non-degenerate) Qwen3-32B programs from Section 4.4, included verbatim and unedited.

```
import math

# Thresholds for classification (clinically plausible defaults)
PARAMS = {
    "min_flair_z_mean": -2.0,              # Minimum FLAIR z-score to be considered a genuine
     hyperintensity
    "min_flair_contrast": 0.5,             # Minimum FLAIR contrast (candidate vs surroundings
     )
    "min_equiv_diam_mm": 3.0,              # Minimum equivalent diameter for a candidate to be
      considered a lesion
    "min_t2_z_mean": -1.0,                 # Minimum T2 z-score for confirmation as lesion
    "t1_black_hole_threshold": -1.0,       # T1 contrast threshold for chronic MS lesions ("
     black holes")
    "thin_rim_thickness": 1.0,             # Thickness in mm for a thin rim (normal
     periventricular rim/cap)
    "hugging_ventricle_ratio": 0.7,        # Fraction of boundary near ventricle for a rim/cap
    "perpendicular_orientation_deg": 30,   # Maximum angle from perpendicular to ventricle for
      "Dawson's finger"
    "symmetry_overlap_threshold": 0.5,     # Symmetry overlap fraction for age-related
     symmetric changes
    "symmetry_area_ratio_threshold": 0.5,  # Symmetric partner area ratio for age-related
     changes
    "abnormal_threshold": 0.7,             # Threshold for calling a candidate "abWMH"
    "max_flair_z_mean_background": 1.0,    # Max FLAIR z-score to be considered background (
     not hyperintense)
    "min_flair_z_mean_background": -10.0,  # Min FLAIR z-score to be considered background (
     not hyperintense)
    "min_flair_contrast_background": -1.0, # Min FLAIR contrast to be considered background
    "min_t2_contrast_background": -1.0,    # Min T2 contrast to be considered background
    "background_thickness_threshold": 0.5, # Max mean thickness for background (too thin to be
      a lesion)
    "background_max_diameter": 1.5,        # Max diameter for background (too small to be a
     lesion)
    "background_flair_z_max_threshold": 1.0, # Max FLAIR z-score max for background
    "juxtacortical_distance_threshold": 3.0, # Distance to brain edge for juxtacortical
     location
}

def classify(c: dict, P: dict = None) -> dict:
    if P is None:
        P = PARAMS

    reasons = []
    score_abnormal = 0.0

    # 1. Check if the candidate is not hyperintense at all (background)
    if (not math.isnan(c["flair_z_mean"]) and
        c["flair_z_mean"] < P["min_flair_z_mean"] and
        (math.isnan(c["flair_contrast"]) or c["flair_contrast"] < P["min_flair_contrast"])):
        reasons.append("low_flair_signal")
        return {"decision": "background", "score_abnormal": 0.0, "reasons": reasons}

    if (not math.isnan(c["flair_z_mean"]) and
        c["flair_z_mean"] > P["max_flair_z_mean_background"] and
        (not math.isnan(c["flair_contrast"]) and c["flair_contrast"] < P["
     min_flair_contrast_background"])):
        reasons.append("high_flair_but_low_contrast")
        return {"decision": "background", "score_abnormal": 0.0, "reasons": reasons}

    if (not math.isnan(c["flair_z_max"]) and
        c["flair_z_max"] < P["min_flair_z_mean_background"]):
        reasons.append("low_flair_max")
        return {"decision": "background", "score_abnormal": 0.0, "reasons": reasons}

    if (not math.isnan(c["equiv_diam_mm"]) and
        c["equiv_diam_mm"] < P["background_max_diameter"] and
        (not math.isnan(c["mean_thickness_mm"]) and c["mean_thickness_mm"] < P["
     background_thickness_threshold"])):
        reasons.append("too_small_and_thin")
        return {"decision": "background", "score_abnormal": 0.0, "reasons": reasons}

    # 2. Check for periventricular "normal" rims/caps (nWMH)
    if (not math.isnan(c["frac_boundary_near_ventricle"]) and
        c["frac_boundary_near_ventricle"] > P["hugging_ventricle_ratio"] and
        (not math.isnan(c["mean_thickness_mm"]) and c["mean_thickness_mm"] < P["
     thin_rim_thickness"])):
        reasons.append("thin_periventricular_rim")
        return {"decision": "nWMH", "score_abnormal": 0.0, "reasons": reasons}
```

```
    if (not math.isnan(c["dist_to_ventricle_mm"]) and c["dist_to_ventricle_mm"] == 0 and
        not math.isnan(c["orientation_vs_ventricle_deg"]) and
        c["orientation_vs_ventricle_deg"] > P["perpendicular_orientation_deg"]):
        reasons.append("hugging_ventricle_along_wall")
        return {"decision": "nWMH", "score_abnormal": 0.0, "reasons": reasons}

    if (not math.isnan(c["sym_overlap_frac"]) and c["sym_overlap_frac"] > P["
     symmetry_overlap_threshold"] and
        not math.isnan(c["sym_partner_area_ratio"]) and c["sym_partner_area_ratio"] > P["
     symmetry_area_ratio_threshold"]):
        reasons.append("symmetric_bilateral_change")
        return {"decision": "nWMH", "score_abnormal": 0.0, "reasons": reasons}

    # 3. Add evidence for abnormality (abWMH)
    # a. Size
    if (not math.isnan(c["equiv_diam_mm"]) and c["equiv_diam_mm"] >= P["min_equiv_diam_mm"]):
        score_abnormal += 0.2
        reasons.append("sufficient_size")

    # b. Perpendicular to ventricle (Dawson's finger)
    if (not math.isnan(c["dist_to_ventricle_mm"]) and c["dist_to_ventricle_mm"] == 0 and
        not math.isnan(c["orientation_vs_ventricle_deg"]) and
        c["orientation_vs_ventricle_deg"] <= P["perpendicular_orientation_deg"]):
        score_abnormal += 0.2
        reasons.append("perpendicular_to_ventricle")

    # c. FLAIR contrast and intensity
    if (not math.isnan(c["flair_z_mean"]) and c["flair_z_mean"] > P["min_flair_z_mean"] and
        not math.isnan(c["flair_contrast"]) and c["flair_contrast"] >= P["min_flair_contrast"
     ]):
        score_abnormal += 0.2
        reasons.append("high_flair_signal")

    # d. T2 contrast and intensity (MS lesions are also T2 hyperintense)
    if (not math.isnan(c["t2_z_mean"]) and c["t2_z_mean"] > P["min_t2_z_mean"] and
        not math.isnan(c["t2_contrast"]) and c["t2_contrast"] > 0):
        score_abnormal += 0.15
        reasons.append("high_t2_signal")

    # e. T1 hypointensity (chronic MS lesion "black hole")
    if (not math.isnan(c["t1_z_mean"]) and c["t1_z_mean"] < P["t1_black_hole_threshold"] and
        not math.isnan(c["t1_contrast"]) and c["t1_contrast"] < P["t1_black_hole_threshold"]):
        score_abnormal += 0.15
        reasons.append("t1_hypointensity_black_hole")

    # f. Shape: ovoid (not elongated)
    if (not math.isnan(c["elongation"]) and c["elongation"] < 3.0):
        score_abnormal += 0.1
        reasons.append("ovoid_shape")

    # g. Juxtacortical location
    if (not math.isnan(c["dist_to_brain_edge_mm"]) and c["dist_to_brain_edge_mm"] < P["
     juxtacortical_distance_threshold"]):
        score_abnormal += 0.1
        reasons.append("juxtacortical_location")

    # Final decision
    if score_abnormal >= P["abnormal_threshold"]:
        return {"decision": "abWMH", "score_abnormal": score_abnormal, "reasons": reasons}
    else:
        return {"decision": "nWMH", "score_abnormal": score_abnormal, "reasons": reasons}
```