ADAPTIVE SEQUENCE MODELING FOR PROGRESSIVE EMERGENCY MEDICAL SERVICE PROTOCOL PREDICTION UNDER STREAMING INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Emergency Medical Technicians (EMTs) must select treatment protocols under severe time and cognitive constraints while patient information is still being collected. Existing Emergency Medical Service (EMS) prediction systems typically require complete records, leaving early response unsupported. We present a Temporal EMS Protocol Prediction System that progressively ranks protocols as evidence accumulates from dispatch through provider impressions. Progressive Clinical Feature Learning (PCFL) uses independently trained predictors at successive, capture-time-audited stages, restricting predictions to available information. Stage-wise Ensembles for Temporal Pooling (STEP) adaptively combines stage-wise and modality-specific predictions without additional patient information. The Dispatch–Protocol Confidence Score (DPCS) provides stage-dependent reliability estimates from empirical dispatch–protocol relationships. Staged-Evidence Prompting (SEP) evaluates large language models (LLMs) under matched evidence constraints, enabling record-level comparisons with structured predictors. Evaluation on the National Emergency Medical Services Information System (NEMSIS) dataset shows improving protocol-ranking accuracy as clinical evidence accumulates, with structured predictors consistently outperforming zero-shot LLMs under matched conditions. STEP further improves ranking performance, while DPCS provides an interpretable reliability signal with observed accuracy. These findings support modeling temporal EMS documentation structure to provide progressively informed decision support throughout patient care.

1 INTRODUCTION

Emergency Medical Services (EMS) decisions are made under severe time pressure, incomplete information, and high clinical stakes. Field providers must select an appropriate protocol from a large protocol set while the patient's condition is still being assessed and the prehospital record is still being constructed. Dispatch initially provides a coarse description of the call. Scene observations, clinical findings, vital signs, and provider impressions then progressively clarify the encounter. Consequently, protocol selection is not a single decision made from a complete record, but a process that unfolds as evidence becomes available.

Most clinical prediction systems are evaluated using a fixed representation assembled after the relevant information has been collected. Although this setting measures performance from a completed record, it cannot determine when useful protocol discrimination first becomes possible or how predictions change as evidence accumulates. This distinction is particularly important in EMS, where earlier predictions may be operationally valuable even when the available evidence remains incomplete. Existing EMS assistants have explored protocol-oriented concept extraction, multimodal sensing, and interactive support (Jin et al., 2023; Preum et al., 2020; Weerasinghe et al., 2024; Jin et al., 2025a;b), but progressive protocol prediction under a common, stage-specific evidence constraint remains insufficiently studied. In parallel, large language models (LLMs) have demonstrated broad capabilities on clinical tasks (Chen et al., 2024; Labrak et al., 2024), yet their behavior on fine-grained EMS protocol ranking as structured evidence is progressively revealed remains unclear.

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.

We address this problem through progressive EMS protocol prediction, in which the historically recorded protocol is ranked repeatedly as new feature blocks become available. We evaluate this formulation over six cumulative stages: dispatch, complaint location, complaint organ system, scene assessment, vital signs, and provider impressions. The prediction task covers 82 protocol classes derived from NEMSIS 2024. The scene-assessment stage includes both documented symptoms and audited scene-context features, consistent with the current feature-availability definitions. These stages form a workflow-informed feature-revelation protocol rather than a reconstruction of the exact timestamps at which individual observations were acquired. The resulting stage-wise evaluation reveals how protocol identifiability changes across the encounter instead of reducing performance to a single measurement after the record is complete.

This progressive setting introduces several connected machine-learning challenges. First, each stage must contain only features that could legitimately be available at that point in the workflow, requiring explicit feature-availability definitions and safeguards against future information leakage. Second, heterogeneous predictors must receive matched records and matched evidence for their performance to be compared meaningfully. Third, predictions produced from earlier and modality-specific views of the encounter may provide useful estimation structure, but their relevance can change as additional evidence becomes available. Finally, a displayed shortlist should communicate empirical reliability in a form that complements model probabilities without being interpreted as a guarantee of correctness.

The challenge therefore extends beyond selecting a predictive architecture. Constructing a progressive evaluation requires a cumulative and auditable feature representation, consistent records across stages, and data-integrity controls that prevent later information from entering earlier predictions. Furthermore, the model will need to distinguish between a highly imbalanced set of protocols at a larger set to allow for more fine-grained protocol prediction which may aid novice EMT or complex protocol situations which generally get mis-labeled as a "general" protocol.

In summary, our paper establishes the following contributions:

- Progressive Clinical Feature Learning (PCFL) formulates EMS protocol ranking through audited, nested feature blocks that restrict predictions to stage-available evidence, with matched records and hit definitions across predictors. To our knowledge, machine learning literature has not established such a system in pre-hospital care.
- Stage-wise Ensembles for Temporal Pooling (STEP) optimizes information gain in a high tail imbalanced multi-stage framework to extract prior-stage information with a Temporal Feature Aggregation (TFA) component.
- Dispatch–Protocol Confidence Score (DPCS) annotates STEP shortlists with stage-dependent reliability tiers derived from protocol-recognition rates and dispatch-conditioned accuracy, without fitting to test labels or reranking predictions, supporting temporal prediction interpretation and reliance.
- SEP extends PCFL's information contract to LLMs by serializing only stage-permitted evidence, enabling comparisons with structured predictors on identical records under matched observability. We evaluate locally served Qwen3-32B-AWQ and MedGemma-27B-text in zero-shot and few-shot settings.

Our paper further conducts an extensive empirical study for this framework through PCFL at a fundamental level. PCFL is predictor-agnostic and we evaluate it across multiple neural architectures including neural networks, LSTMs, transformers, decision trees as well as cascaded PCFL variants. Furthermore, we explore the effects of our STEP ensembled pooling approach under significant hyperparameter modifications along with a paired McNemar test for validating information gain across each comparison. Lastly, we fully explore and compare our method with local LLM inferences through Qwen3 and MedGemma, frontier medical domain LLMs, and conduct an in-depth comparison of the protocol class capture context for each method across all stages with multi-shot context. Beyond EMS, our formulation motivates research in other settings where features become available in a clinically ordered sequence, such as intensive-care decision support, mass-casualty triage, and general first responder scenarios.

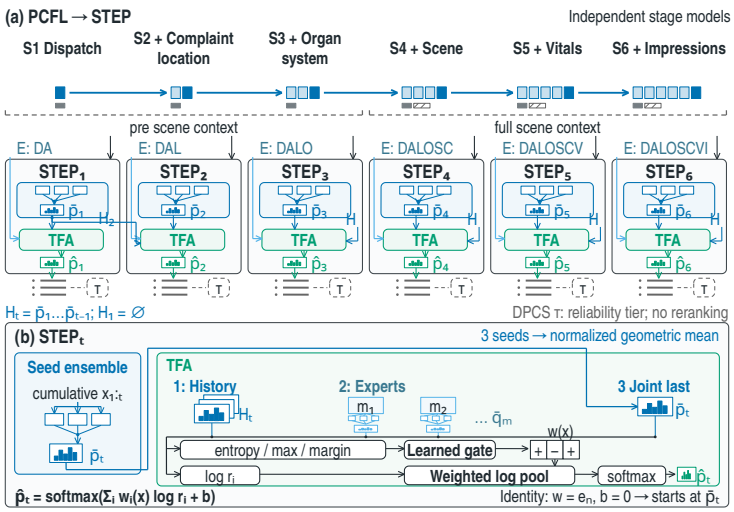

Figure 1: **Progressive EMS protocol prediction with STEP.** (a) Temporal EMS System Architecture. (b) STEP framework (seed ensembling, early stage posterior, single-modality expert)

2 Related Work

EMS decision support. EMSAssist pairs on-device speech recognition with protocol-oriented assistance and EMSConText builds an EMS concept lexicon (Jin et al., 2023; Preum et al., 2020); multimodal assistants integrate scene signals (Weerasinghe et al., 2024), and smart-glasses and mobile-video systems extend sensing to wearable and remote platforms (Jin et al., 2025a;b). That work concerns what an assistant can capture; here the question is what can be inferred from it.

Early prediction, stacking and opinion pools. Early time-series classification trades accuracy against earliness (Schäfer & Leser, 2020), anytime classifiers emit a usable prediction as computation proceeds (Jazbec et al., 2023), and active feature acquisition learns which variable to query next (Li & Oliva, 2021); here the acquisition order is the clinical workflow's, so the quantity of interest is the incremental value of each block when it arrives. Recurrent and attention models exploit repeated measurements in a fixed representation (Harutyunyan et al., 2019; Choi et al., 2016), whereas an encounter's representation expands across block types, and tuned MLPs are competitive on high-dimensional sparse tabular inputs (Grinsztajn et al., 2022; Gorishniy et al., 2021; Kadra et al., 2021; Borisov et al., 2024). STEP assembles established parts. Its stage and expert predictors are geometric-mean deep ensembles (Lakshminarayanan et al., 2017), and its pooling subcomponent TFA is stacked generalisation (Wolpert, 1992; Ting & Witten, 1999): a weighted logarithmic opinion pool, equivalently a product of experts (Genest & Zidek, 1986; Hinton, 2002; Kittler et al., 1998), with input-dependent weights as in mixtures of experts (Jacobs et al., 1991), fitted on out-of-fold posteriors so that the second-stage learner never sees overconfident in-sample predictions (Chernozhukov et al., 2018); what is specific here is their arrangement along the stage axis under a fixed information contract. Temporal recommenders re-rank as a history accumulates (Fan et al., 2021; Cho et al., 2020) and confidence-aware ranking separates relevance from reliability (Knyazev & Oosterhuis, 2023); because neural probabilities are often miscalibrated (Guo et al., 2017), the displayed confidence is an empirical reliability estimate rather than the model's own posterior.

LLMs on clinical tasks. Large language models show zero- and few-shot competence across clinical benchmarks (Chen et al., 2024; Labrak et al., 2024), and EMS-specific work explores prompting and retrieval augmentation (Ge et al., 2026). Ranking a closed set of 82 fine-grained protocols from stage-limited coded evidence differs from open-ended clinical question answering: the label set is fixed, the evidence is structured, and the binding constraint is which fields exist yet; SEP (§3.4) turns that constraint into an interface any instruction-tuned model can be run through.

3 Progressive EMS Protocol Prediction

Our temporal EMS architecture is shown in Figure 1. Throughout, $t \in \{1, \ldots, 6\}$ indexes the stage, e an encounter, $\mathcal{P}$ the protocol set ($|\mathcal{P}| = 82$), y_e the recorded protocol, I_e the dispatch code, $\bar{p}_t$ the

seed-ensembled and $\hat{p}_t$ the STEP (pooled) stage-t posterior, and $\mathcal{K}_e^{(t)}$ the displayed shortlist; *gate* refers only to the pool's learned weight network.

3.1 Progressive Clinical Feature Learning (PCFL)

Task and stages. Let $B_{e,1}, \ldots, B_{e,6}$ be the six ordered blocks of a record: dispatch, complaint location, complaint organ system, the scene assessment, vital signs and provider impressions. At stage t,

$$x_e^{(t)} = \text{concat}\left(B_{e,1}, \ldots, B_{e,t}\right) \in \mathbb{R}^{D_t}, \qquad \pi_e^{(t)} = f_t\big(x_e^{(t)}\big) \in \Delta^{|\mathcal{P}|}, \tag{1}$$

and the displayed artifact is the shortlist $\mathcal{K}_e^{(t)}$ of the $K = 3$ protocols with the largest posterior; a hit at depth K is $h_K(e,t) = \mathbb{I}[y_e \in \mathcal{K}_e^{(t)}]$, from which every metric and test in the paper is built. The dimension grows as $D_t = 56, 66, 78, 1,321, 1,398, 2,073$ (App. A). The nesting $X^{(1)} \subset \cdots \subset X^{(6)}$ is verified by set comparison of the column lists, Our features were selected through guidance from an EMS domain expert. We acknowledge that sequence ordering can vary from different stages and we leave exploration for permutations as future study.

3.2 Stage-wise Ensembles for Temporal Pooling (STEP)

At each of six stages, STEP combines independently trained networks into a seed ensemble, then applies confidence-dependent temporal pooling (TFA) with earlier-stage and single-modality posteriors.

Stage predictor and seed ensemble. Each stage uses an MLP with hidden widths $[256, 128]$, ReLU, dropout 0.2 and 82 logits, trained with softmax cross-entropy (Adam, 10^{-3}, batch 256, 200 epochs, last-epoch weights). Predictions from seeds $\{0, 1, 2\}$ form a normalised geometric mean,

$$\bar{p}_t(y \mid x) = \frac{\prod_{s=0}^{2} p_t^{(s)}(y \mid x)^{1/3}}{\sum_{y'} \prod_{s=0}^{2} p_t^{(s)}(y' \mid x)^{1/3}}, \tag{2}$$

equivalent to averaging and renormalising log-softmax outputs. Single-modality experts use the same recipe on their raw blocks. Three seeds are fixed in advance to stabilise pooling inputs; single-seed stage-level test top-1 varies by up to 0.78 percentage points (App. C).

Temporal pooling (TFA). TFA combines two input axes: *stage history*, comprising unpooled posteriors $\bar{p}_1, \ldots, \bar{p}_{t-1}$, and *modality experts*, comprising ensembled MLPs $q_{m,t}$ for each available block: dispatch, location, organ system, symptoms, vitals, impressions and context. The context expert uses age before the scene stage and full context thereafter (Table 4). Pools are not chained. The current joint posterior $\bar{p}_t$ is the last input, yielding $n = 2t + 1$ inputs: three at stage 1 and thirteen at stage 6. For ordered inputs $r_1, \ldots, r_n$,

$$\hat{p}_t(y \mid x) = \text{softmax}_y\left(\sum_{s=1}^{n} w_s(x)\, \log r_s(y \mid x) + b_y\right), \tag{3}$$

$$w(x) = w_0 + W_2 \tanh\big(W_1 c(x) + b_1\big) + b_2. \tag{4}$$

The confidence trajectory $c(x) \in \mathbb{R}^{3n}$ collects each input's entropy, maximum probability and top-1/top-2 margin. The gate has hidden width 32, class-tied weights and per-class bias $b \in \mathbb{R}^{82}$. Weights are unconstrained and unnormalised, not attention probabilities. Each stage pools the record separately.

Fitting and selection. *Identity initialisation*: $w_0 = e_n$, $W_2 = 0$, $b_2 = 0$ and $b = 0$ yield $\hat{p}_t = \bar{p}_t$. *Cross-fitting*: each pool uses only stages $1, \ldots, t$ and is fitted on five-fold out-of-fold posteriors from fit rows (fold seed $100 \cdot \text{seed} + \text{fold}$), avoiding overconfident in-sample inputs. Validation and test posteriors use base models trained only on fit rows (App. C). *Selection*: validation NLL selects the epoch, including epoch 0, so pools that fail to improve NLL retain the ensemble (Figure 16). Test labels never enter fitting, selection or pooling decisions.

Scope of pooling. All inputs respect stage-t field availability. At stages 1–5, the joint input contains every earlier-stage and expert feature, so pooling offers estimation benefits without additional information. At stage 6, joint inputs merge symptom and impression indicators sharing an ICD-10-CM code; preserving their provenance during retraining leaves ensemble performance unchanged

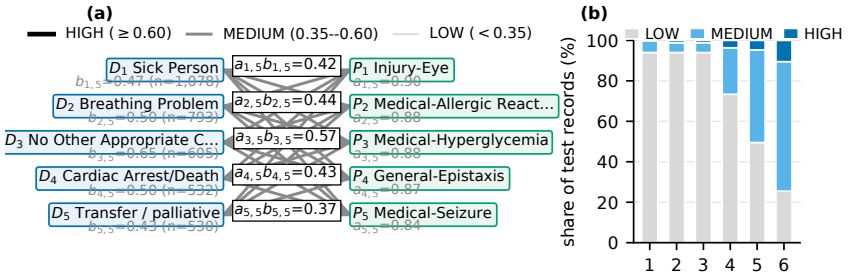

Figure 2: DPCS on the STEP shortlist. (a) The dispatch–protocol reliability graph at stage 5. (b) The share of test records in each tier at every stage (counts in App. D).

(App. C). With fixed class-tied weights, pooling is additive in expert class log-odds, making the joint predictor necessary to represent cross-modality interactions under that restriction (App. C). Empirically, expert-only pooling underperforms the ensemble at every stage from 2 onward.

3.3 Dispatch–Protocol Confidence Score (DPCS)

DPCS annotates shortlists using stage-specific protocol recall and dispatch-conditioned accuracy. For N_j inner-validation encounters with protocol P_j and M_i with dispatch code D_i,

$$a_{j,t} = \frac{1}{N_j} \sum_{r:\, y_r = j} \mathbb{I}[\hat{y}_{r,t} = j], \qquad b_{i,t} = \frac{1}{M_i} \sum_{r:\, I_r = i} \mathbb{I}[\hat{y}_{r,t} = y_r], \tag{5}$$

where $\hat{y}_{r,t}$ is STEP's top-1 prediction from pooled out-of-fold base posteriors. Both rates use only the 7,378 inner-validation rows. The shortlist score and tier are

$$\mathrm{DPCS}_e^{(t)} = b_{I_e,t} \max_{j \in \mathcal{K}_e^{(t)}} a_{j,t}, \qquad \tau_e^{(t)} = \begin{cases} \textsc{low} & \mathrm{DPCS}_e^{(t)} < 0.35 \\ \textsc{medium} & 0.35 \le \mathrm{DPCS}_e^{(t)} < 0.60 \\ \textsc{high} & \mathrm{DPCS}_e^{(t)} \ge 0.60. \end{cases} \tag{6}$$

Figure 2 shows (a) the stage-5 graph with edge weights $a_{j,5} b_{i,5}$ and (b) tier proportions across six stages. DPCS is an uncalibrated reliability proxy that preserves rankings and accuracy. It requires only a shortlist and inner-validation hit rates, making it predictor-agnostic.

3.4 Staged-Evidence Prompting (SEP) of existing LLMs

Interface. SEP shares the supervised arm's stage definitions and per-field capture-time cutoffs through a common module. Tests verify that scene-time fields appear only from the symptom block onward. Unmeasured vitals render as `[not measured]`; medications, procedures and record identifiers are excluded.

Prompt assembly. Each prompt contains a byte-identical prefix (task, output contract, field legend and 82 protocols as ordered `id<TAB>name` pairs), k demonstrations and the stage-t query (App. B). A seeded traversal of fit rows selects the first record from each previously unselected protocol, yielding k distinct-protocol demonstrations rendered at the query's stage with gold answers. Fixed ordering enables prefix caching; list-order sensitivity is ablated in App. B.

Output contract and shortlist. The model returns one JSON object ranking ten protocol identifiers. For language model g and stage-t demonstrations $\mathcal{D}_k^{(t)}$,

$$\rho_e^{(t)} = \mathrm{parse}\left(g\big(\mathrm{prompt}_t\left(x_e^{(t)}; \mathcal{D}_k^{(t)}\right)\big)\right), \qquad \mathcal{K}_e^{(t)} = \rho_e^{(t)}[1{:}K], \tag{7}$$

where parse removes leaked reasoning or markdown fences, attempts JSON parsing, then regular-expression extraction. Failure returns an empty list, scored as a miss at every depth. Both arms use the same hit indicator $h_K(e,t)$ and test rows, enabling paired comparisons.

Instantiation and selection. Qwen3-32B-AWQ (Yang et al., 2025) runs locally through vLLM (Kwon et al., 2023) on two A30 GPUs (App. B). Prompt, format and $k \in \{0, 3, 5\}$ were selected on `val820`, a hashed inner-validation subset using the historical acceptance test (App. E). Three-shot prompting was accepted over zero-shot; five-shot was rejected because it improved top-1 but

stage	NN top-1 / top-3	ens. top-1 / top-3	STEP top-1 / top-3	Δ_1 [95% CI] (p_1)	Δ_3 [95% CI] (p_3)
1 dispatch	20.36 / 38.09	20.42 / 38.02	20.72 / 38.34	$+0.30$ [0.00, 0.61] (0.06)	$+0.32$ [-0.07, 0.72] (0.13)
2 +location	24.54 / 43.38	24.54 / 43.79	24.63 / 43.91	$+0.09$ [-0.18, 0.35] (0.59)	$+0.12$ [-0.29, 0.55] (0.6)
3 +organ system	27.31 / 47.79	27.56 / 48.16	27.85 / 48.39	$+0.29$ [-0.04, 0.61] (0.098)	$+0.23$ [-0.20, 0.65] (0.3)
4 +symptoms, scene	45.84 / 67.32	48.16 / 69.60	48.24 / 70.52	$+0.09$ [-0.51, 0.67] (0.8)	$+0.91$ [0.39, 1.43] (6.2×10^{-4})
5 +vitals	50.34 / 72.13	52.94 / 74.53	53.71 / 75.56	$+0.77$ [0.24, 1.31] (0.0051)	$+1.02$ [0.54, 1.51] (3.7×10^{-5})
6 +impressions	56.03 / 75.98	58.34 / 78.56	59.31 / 79.82	$+0.96$ [0.38, 1.51] (7.7×10^{-4})	$+1.27$ [0.78, 1.77] (6.8×10^{-7})

Table 1: Balanced test split ($n = 8{,}198$), clean boundary: NN, ens. (3-seed geometric mean) and STEP, top-1 / top-3 (%); Δ: STEP $-$ ens., paired, bootstrap 95% CI and exact McNemar p (the corrected family of 12 tests, Bonferroni $0.05/12 = 0.0042$).

reduced top-3. Further variants used only `val820` (App. B); the test set was read once per final configuration. SEP requires no model training. MedGemma-27B-text (Sellergren et al., 2025) uses Qwen3's selected prompt at seed 0 with $k \in \{0, 3\}$, without separate selection (App. B).

4 Experimental Setup

Data. The NEMSIS 2024 public-release research dataset (NEMSIS Technical Assistance Center, 2025; Gregor et al., 2026) provides 81,978 encounters over 82 protocols, with each contributing 1,000 encounters: the cohort is balanced by construction, so the stage curve measures how identifiable a protocol is rather than how often it occurs, and no absolute accuracy here is a deployment estimate. An outer 90/10 stratified split (seed 42) gives 73,780 training and 8,198 test encounters; an inner 10% of the training split (seed 1234; 66,402 fit and 7,378 inner-validation rows) carries every selection decision. Medications and procedures are documented after the protocol decision and are excluded, as is the record identifier. Cohort and feature construction are in App. A.

Comparators and metrics. The primary comparison is STEP against a matched single-seed network (NN): seed 0 of the same stage MLP, with identical features, records, splits, architecture, 200-epoch training budget and last-epoch weights. Seed 0 is the project's established single-seed control and was not chosen by its result; the other seeds and the three-seed ensemble without pooling, which separates STEP's two steps, are in App. C. Four neural alternatives to the per-stage MLP (an LSTM and a Transformer over the stage sequence, and cascaded variants of the MLP and the Transformer that pass the previous stage's hidden representation forward) were run on the project's explicit three-way partition of the same cohort without the context block and are compared against their own MLP control on that partition (App. A); the two SEP configurations complete the set. Top-1 and top-3 accuracy are co-primary, because what a clinician sees is a shortlist.

Stage-application policy and tests. Every comparison is an exact, two-sided, paired McNemar test (McNemar, 1947; Dietterich, 1998) on per-record hit vectors, computed with the hit defined as h_1 and as h_3 over the same records. The 12 test-split comparisons that decompose STEP (three contrasts, six stages, two metrics) are Bonferroni-corrected at $0.05/12$; SEP's configuration was selected on validation rows with the historical acceptance test.

5 Results

5.1 How do stages and pooling affect ranking accuracy?

Table 1 and Figure 3(a) show accuracy increasing at every stage, from 20.72% to 59.31% top-1 and 38.34% to 79.82% top-3. The scene stage contributes 20.4 top-1 %, exceeding the other four transitions combined. Seed ensembling accounts for main gains over the matched NN at stages 4–6 (+2.3 to +2.6 top-1 %). Beyond ensembling, pooling adds $+0.77$ top-1 and $+1.02$ top-3 % at the vital-sign stage ($p = 0.0051$, 3.7×10^{-5}), and $+0.96$ and $+1.27$ at the impression stage ($p = 7.7\times10^{-4}$, 6.8×10^{-7}). At the scene stage, it improves top-3 by $+0.91$ ($p = 6.2\times10^{-4}$), with top-1 unchanged; earlier gains are negligible and nonsignificant. Of 12 paired tests, 4 survive the family-wise threshold $0.05/12$: top-3 at stages 4–6 and top-1 at stage 6. Pooling also reduces negative log-likelihood at every stage (stage 6: 1.696 to 1.498), with recalibration alone accounting for roughly half this reduction (1.592). Full results and implementation details appear in App. C.

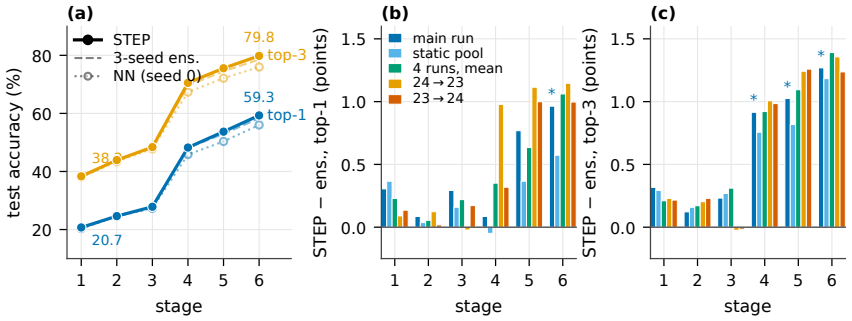

Figure 3: Balanced test split ($n = 8{,}198$; values in Tables 1 and 20). (a) Top-1 and top-3 by stage(b, c) STEP − ens. in % under five sources of evidence (stars: comparisons clearing $0.05/12$).

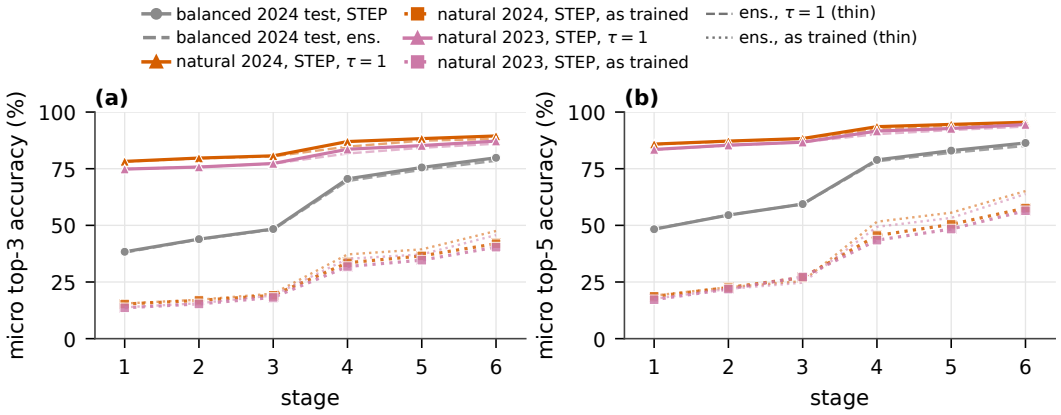

Figure 4: Micro (a) top-3 and (b) top-5 by stage on natural 2024 (90,759) and 2023 (100,000) cohorts: as trained (dotted), prior-corrected at τ=1 (solid), STEP (markers), and ensemble (thin).

5.2 Is the increment robust?

Figure 3(b, c) summarizes three robustness checks (Table 20, App. C): across checks and both metrics, increments remain near zero at stages 1–3 and around one point at stages 5–6. *Replications.* Four full-pipeline replications with disjoint seed triples yield mean stage-6 increments of $+1.06$ top-1 (range $[+0.96, +1.17]$) and $+1.39$ top-3. *Static against gated.* A log-linear pool with fixed per-input weights and per-class biases (Wołczyk et al., 2021; Heskes, 1997) recovers most of the increment on the same inputs ($+0.37$ / $+0.82$ at stage 5; $+0.57$ / $+1.18$ at stage 6). Gating adds $+0.40$ and $+0.39$ top-1 %, respectively; non-negative (Breiman, 1996) or ridge-shrunk static weights change results by only ± 0.06. Fixed weights thus explain most of the gain. *Cross-year transfer.* Training on one year's balanced fit rows and evaluating on the other year preserves the full increment in both directions: stage-6 top-1 / top-3 gains are $+1.15$ / $+1.36$ for 2024→2023 under matched cohort rules and $+1.00$ / $+1.24$ for 2023→2024. *Answer:* pooling on these inputs produces consistent gains across seed triples and years, with a smaller contribution from gating.

5.3 What changes under natural and multi-year distributions?

The balanced cohort measures protocol identifiability: with about 100 test records per protocol, top-k accuracy approximates mean per-protocol recall. Natural cohorts instead measure prevalence-weighted performance: one protocol accounts for 54% of the 2024 cohort (90,759 records, excluded from training and selection) and 53% of the 2023 cohort (100,000 records), making micro accuracy sensitive to majority-class performance and prior mismatch. Figure 4 and Table 26 (App. F) evaluate the models from Table 1 as trained and with full prior correction (Saerens et al., 2002), $p_\tau(y \mid x) \propto p(y \mid x)(\pi_{\text{target}}(y)/\pi_{\text{train}}(y))^\tau$, at τ=1. The training prior is balanced; target priors are estimated from each random cohort half and applied to the other, with both halves scored. This is an oracle-prior evaluation relative to deployment, following Table 27.

As trained. Both models fall below balanced-test performance: stage-6 top-3 is 47.52% / 41.92% on 2024 and 45.71% / 40.38% on 2023 (ensemble / STEP). STEP reduces micro top-3 by -5.60 and -5.34 percentage points, respectively, as its balanced-fit class bias shifts mass away from the majority protocol. Protocol-level gains persist: 2024 stage-6 macro top-1 rises from 58.18% to 60.05%.

Prior-corrected. At τ=1, STEP's stage-6 top-3 reaches 89.45% on 2024 (gain: +47.53 percentage points) and 87.18% on 2023 (gain: +46.80 percentage points); top-5 reaches 95.51% and 94.50%. STEP outperforms the ensemble on every metric, with stage-6 top-3 / top-5 gains of +1.32 / +0.85 on 2024 and +1.16 / +0.91 on 2023. Corrected curves exceed balanced-test accuracy at every stage. On balanced cross-year hold-outs, prior correction is nearly the identity, and STEP retains its full advantage in both directions without correction.

Qualifications. Label-free EM prior estimation (Saerens et al., 2002) preserves the ordering (Table 27). Refitting with zero class bias also turns the as-trained micro loss into a gain (stage-6 top-1: 25.22% versus 24.42%), but remains exploratory because it was selected after inspecting external results. *Answer:* STEP's protocol-level gains transfer across years, while its micro-accuracy gains on natural cohorts depend on accounting for the target prior.

5.4 What does the increment consist of?

Under one selection policy and a validation sweep for each, no LSTM, Transformer or cascade beats the per-stage MLP ensemble at any stage; all trail it by 0.1 to 1.7 top-1 % at stages 5–6 (Table 6). *A stronger base predictor*: a gradient-boosted model tuned on inner validation beats the single-seed MLP by +1.4 to +2.2 top-1 % at stages 4–6 and matches the ensemble; STEP leads it by $+1.0$ to $+1.2$ ($p \leq 0.007$; Table 21), and the same pool on boosted stage and expert posteriors adds $+0.90$ / $+0.74$ and $+0.87$ / $+0.68$ at stages 5–6 over the single boosted model ($p \leq 0.0031$).

What the earlier posteriors carry (App. F). At each transition 2.7% to 5.8% of records are ranked correctly by the earlier network and incorrectly by the later one, and for 71% to 82% of them the earlier answer stays in the later top-3, which is why the top-3 increment is the more robust one; a hard rule that reverts to the earlier prediction cannot separate these from legitimate corrections (net change between -2.9 and $+0.5$ % over reversion fractions of 1–20%), the soft pool can. A training-time congruence term toward the earlier stage (Yan et al., 2021) raises the single network but not the seed ensemble, a block-masked network trained over all six reveal stages matches the ensemble with about 33% fewer harmful transitions at the two richest stages, and in the conjunction test the jointly trained model recovers 29–33% of the records whose true protocol is in the top-3 of at least two block experts but which an experts-only pool misses, against 7–12% for two more expressive decision-level combiners. With the synthetic study the evidence is consistent with a finite-sample estimation effect rather than added information, without establishing that mechanism uniquely.

Generality (App. F). Tuning the synthetic study's stage-2 joint model over weight decay, dropout and early stopping before pooling removes at most a fifth of a material gain (pure-noise block: $+11.4$, $+8.2$ and $+3.5$ top-1 % at n = 2,000, 8,000 and 32,000 against $+14.4$, $+8.4$ and $+3.6$ untuned; Table 31). On UCI Covertype with an artificial, documented reveal order (Table 32) the pool is inert while every block is low-dimensional (increments within ± 0.7, all $p > 0.2$), and once the last block is padded with 500 sparse, weakly class-dependent columns, the regime the synthetic study identifies, it adds $+3.9$ / $+1.1$ top-1 / top-3 at n = 5,000 ($p = 3.3 \times 10^{-6}$) and $+3.9$ / $+0.8$ at n = 20,000 ($p = 2.0 \times 10^{-22}$; static pool $+4.1$) and halves the harmful-transition rate (10.9% to 4.7% at n = 5,000). *Answer:* the increment is not recalibration, not a fixed product, not an artefact of a weak base predictor or of the MLP, and appears in the predicted regime on an unrelated dataset; the mechanism is consistent with finite-sample estimation but not established uniquely.

5.5 How do prompted LLMs compare under the same stage contract?

Figure 5 shows a confusion matrix comparison between the methods. The figure shows Stage-6 top-1 predictions collapsed to the ten coarse protocol families on the balanced test split (printed when $\geq$ 0.5%; square-root colour scale) for (a) STEP, (b) Qwen3-32B-AWQ k=3 seed 0 and (c) MedGemma-27B-text k=3 seed 0. Both prompted models keep most of their mass inside the Medical and General families and rarely predict the Airway, Cardiac-arrest and OB/GYN families, whereas STEP's errors stay largely within the true family. Figure 8 pairs both prompted arms with STEP record for record. Locally hosted LLMs reproduce the stage-wise performance trend but in-

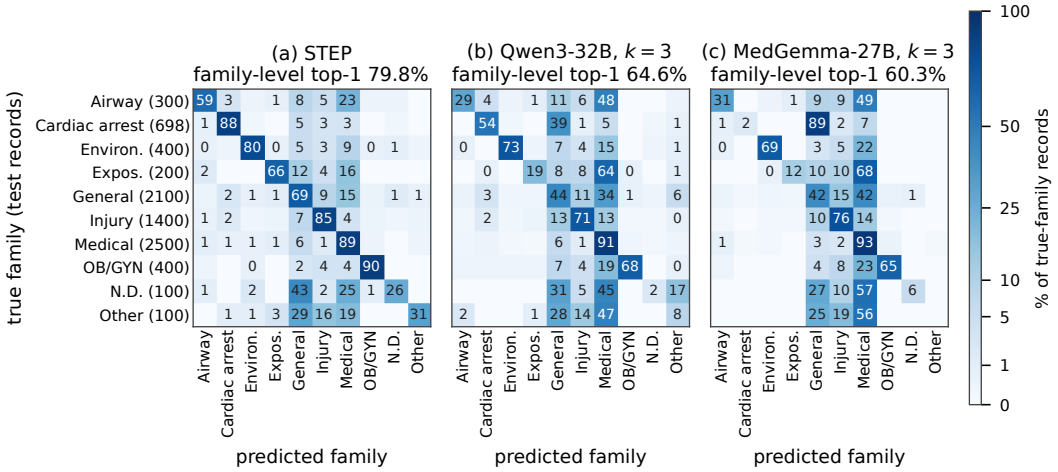

Figure 5: Confusion Matrix Comparison: STEP vs. Qwen vs. MedGemma

creasingly trail STEP as evidence accumulates, while STEP offers substantially faster inference. We did not evaluate commercial API models such as GPT, Claude or Gemini: sending prehospital records to a third-party API raises privacy and data-governance questions that the local arms avoid. *Answer:* under matched evidence both locally served prompted models trail STEP at every stage and the gap widens with the evidence; furthermore, STEP is able to surpass LLM baselines on all protocols and achieve higher fine-grained discrimination. Complete per protocol confusion matrices as well as per protocol f1 scores can be found in App. B.

5.6 Does DPCS carry reliability information?

From the scene stage on, observed shortlist accuracy rises monotonically from LOW to HIGH (stage-6 top-1: 42.82%, 62.21% and 80.61%), so the displayed tier carries information about reliability; but the predictor's own confidence orders accuracy better on the same records, with a stage-6 AUROC for detecting a wrong top-1 of 0.809 for the maximum posterior against 0.650 for DPCS (App. D). *Answer:* DPCS is an interpretable, stage-dependent annotation whose tiers order observed accuracy; it is not a reliability score to prefer over the model's confidence.

6 Conclusion

This work formulates pre-hospital protocol ranking as progressive prediction under accumulating clinical evidence. PCFL defines workflow-informed evidence stages, SEP enables matched comparisons with language models, STEP combines temporal and modality-specific predictions, and DPCS adds empirical reliability tiers. The supervised pipeline achieves 60.09% top-1 and 80.28% top-3 accuracy at the final stage, consistently outperforming the evaluated zero-shot and few-shot Qwen3-32B-AWQ configurations under matched evidence constraints. The widening performance gap as evidence accumulates motivates research into temporally aware language models and agents. Future work should evaluate prospective utility with practicing EMTs and integrate real-time multimodal evidence to support progressively informed protocol review while preserving first responders' clinical authority.

Ethics Statement

This work concerns sensitive prehospital records and a safety-critical decision context. The system predicts historical protocol labels for research and assistive ranking; it is not an autonomous treatment system. No patient-level record, identifier or example appears in this paper or its figures; all qualitative material is class-level aggregate. A hosted-LLM arm was deliberately not run pending a data-use determination, and the local arm never places an identifier, a medication or a procedure in a prompt. Prospective use would require data governance, institutional oversight, fairness

and distribution-shift evaluation, validation with EMS professionals, and an interface that preserves clinician authority.

REPRODUCIBILITY STATEMENT

Our code is available for reproducing results at `https://anonymous.4open.science/r/step-ems-protocol-F025/`. Splits derive from seeds 42 (outer, 73,780 / 8,198) and 1234 (inner, 66,402 / 7,378) and the loader asserts row alignment across stages; the LLM subsets are derived at load time and SHA-256 hashed (App. B). App. A gives the cohort construction, capture-time audit, defect measurements, the exact stage-model hyperparameters and splits; App. B the prompt template, decoding and serving configuration; App. C the pool hyperparameters, experts, the all-stage STEP decomposition and the pool-reproducibility experiments; App. E the historical acceptance test, the multiplicity policy and the per-wave summary of every tested comparison, whose full machine-readable ledger ships with the code. Every table and figure is generated from the saved logits and result files by scripts in the accompanying repository, which record the track, training protocol, ensemble, split and n behind each number.

AI USE STATEMENT

Generative AI assisted literature discovery, manuscript restructuring, language editing, and the scripts that regenerate tables and figures from saved artifacts. It did not generate experimental observations or alter results. The authors verified the assisted material and take responsibility for the final content.

REFERENCES

Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6):7499–7519, 2024. doi: 10.1109/TNNLS.2022.3229161.

Leo Breiman. Stacked regressions. *Machine Learning*, 24(1):49–64, 1996.

Xiaolan Chen, Jiayang Xiang, Shanfu Lu, Yexin Liu, Mingguang He, and Danli Shi. Evaluating large language models in medical applications: A survey. *arXiv preprint arXiv:2405.07468*, 2024. doi: 10.48550/arXiv.2405.07468. URL `https://arxiv.org/abs/2405.07468`.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. doi: 10.1111/ectj.12097.

Sung Min Cho, Eunhyeok Park, and Sungjoo Yoo. MEANTIME: Mixture of attention mechanisms with multi-temporal embeddings for sequential recommendation. In *Proceedings of the 14th ACM Conference on Recommender Systems (RecSys)*, pp. 515–520, 2020. doi: 10.1145/3383313.3412216.

Edward Choi, Mohammad Taha Bahadori, Joshua A. Kulas, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL `https://proceedings.neurips.cc/paper/2016/hash/231141b34c82aa95e48810a9d1b33a79-Abstract.html`.

Thomas G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923, 1998. doi: 10.1162/089976698300017197.

Ziwei Fan, Zhiwei Liu, Jiawei Zhang, Yun Xiong, Lei Zheng, and Philip S. Yu. Continuous-time sequential recommendation with temporal graph collaborative transformer. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 433–442, 2021. doi: 10.1145/3459637.3482242.

Xueren Ge, Sahil Murtaza, Anthony Cortez, and Homa Alemzadeh. Expert-guided prompting and retrieval-augmented generation for emergency medical service question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 30798–30806, 2026. doi: 10.1609/aaai.v40i36.40337.

Christian Genest and James V. Zidek. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1):114–135, 1986. doi: 10.1214/ss/1177013825.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18932–18943, 2021.

Casey F. Gregor, Amanda M. Pope, Kevin Nicholes, Mengtao Dai, Rylee E. Beckstead, Bengie Alexandre, and N. Clay Mann. Description of the 2024 NEMSIS public-release research dataset. *Prehospital Emergency Care*, pp. 1–8, 2026. doi: 10.1080/10903127.2026.2687827.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 35, pp. 507–520, 2022.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1):96, 2019. doi: 10.1038/s41597-019-0103-9.

Tom Heskes. Selecting weighting factors in logarithmic opinion pools. In *Advances in Neural Information Processing Systems*, volume 10, 1997. URL https://papers.nips.cc/paper/1413-selecting-weighting-factors-in-logarithmic-opinion-pools.

Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002. doi: 10.1162/089976602760128018.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991. doi: 10.1162/neco.1991.3.1.79.

Metod Jazbec, James Urquhart Allingham, Dan Zhang, and Eric Nalisnick. Towards anytime classification in early-exit architectures by enforcing conditional monotonicity. In *Advances in Neural Information Processing Systems*, volume 36, pp. 56138–56168, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/af2d9fb5bcee19ef2dfa70d843520c97-Abstract-Conference.html.

Liuyi Jin, Tian Liu, Amran Haroon, Radu Stoleru, Michael Middleton, Ziwei Zhu, and Theodora Chaspari. EMSAssist: An end-to-end mobile voice assistant at the edge for emergency medical services. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services (MobiSys)*, pp. 275–288, 2023. doi: 10.1145/3581791.3596853.

Liuyi Jin, Pasan Gunawardena, Amran Haroon, Runzhi Wang, Sangwoo Lee, Radu Stoleru, Michael Middleton, Zepeng Huo, Jeeeun Kim, and Jason Moats. A smart-glasses for emergency medical services via multimodal multitask learning. arXiv:2511.13078, 2025a.

Liuyi Jin, Amran Haroon, Radu Stoleru, Pasan Gunawardena, Michael Middleton, and Jeeeun Kim. Real-time mobile video analytics for pre-arrival emergency medical services. arXiv:2511.14119, 2025b.

Arlind Kadra, Marius Lindauer, Frank Hutter, and Josif Grabocka. Well-tuned simple nets excel on tabular datasets. In *Advances in Neural Information Processing Systems*, volume 34, pp. 23928–23941, 2021.

Yigitcan Kaya, Sanghyun Hong, and Tudor Dumitras. Shallow-deep networks: Understanding and mitigating network overthinking. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *PMLR*, 2019. URL https://proceedings.mlr.press/v97/kaya19a.html.

Josef Kittler, Mohamad Hatef, Robert P. W. Duin, and Jiri Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998. doi: 10.1109/34.667881.

Norman Knyazev and Harrie Oosterhuis. A lightweight method for modeling confidence in recommendations with learned beta distributions. In *Proceedings of the 17th ACM Conference on Recommender Systems (RecSys)*, pp. 306–317, 2023. doi: 10.1145/3604915.3608788.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, pp. 611–626, 2023. doi: 10.1145/3600006.3613165.

Yanis Labrak, Mickael Rouvier, and Richard Dufour. A zero-shot and few-shot study of instruction-finetuned large language models applied to clinical and biomedical tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 2049–2066, Torino, Italia, 2024. URL https://aclanthology.org/2024.lrec-main.185.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6402–6413, 2017.

Yang Li and Junier Oliva. Active feature acquisition with generative surrogate models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6450–6459. PMLR, 2021. URL https://proceedings.mlr.press/v139/li21p.html.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947. doi: 10.1007/BF02295996.

NEMSIS Technical Assistance Center. 2024 NEMSIS public-release research dataset, NEMSIS v3.5.0. National Emergency Medical Services Information System, https://nemsis.org/using-ems-data/request-research-data/, 2025. User manual: https://nemsis.org/wp-content/uploads/2025/09/2024-NEMSIS-RDS-350-User-Manual-v3.pdf.

Sarah Masud Preum, Sile Shu, Homa Alemzadeh, and John A. Stankovic. EMSContExt: EMS protocol-driven concept extraction for cognitive assistance in emergency response. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 13350–13355, 2020. doi: 10.1609/aaai.v34i08.7048.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1):21–41, 2002.

Patrick Schäfer and Ulf Leser. TEASER: Early and accurate time series classification. *Data Mining and Knowledge Discovery*, 34(5):1336–1362, 2020. doi: 10.1007/s10618-020-00690-z.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin,

Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. MedGemma technical report. arXiv:2507.05201, 2025.

Kai Ming Ting and Ian H. Witten. Issues in stacked generalization. *Journal of Artificial Intelligence Research*, 10:271–289, 1999. doi: 10.1613/jair.594.

Keshara Weerasinghe, Saahith Janapati, Xueren Ge, Sion Kim, Sneha Iyer, John A. Stankovic, and Homa Alemzadeh. Real-time multimodal cognitive assistant for emergency medical services. In *2024 IEEE/ACM Ninth International Conference on Internet-of-Things Design and Implementation (IoTDI)*, pp. 85–96, 2024. doi: 10.1109/IoTDI61053.2024.00012.

Maciej Wołczyk, Bartosz Wójcik, Klaudia Bałazy, Igor Podolak, Jacek Tabor, Marek Śmieja, and Tomasz Trzciński. Zero time waste: Recycling predictions in early exit neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2106.05409.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992. doi: 10.1016/S0893-6080(05)80023-1.

Sijie Yan, Yuanjun Xiong, Kaustav Kundu, Shuo Yang, Siqi Deng, Meng Wang, Wei Xia, and Stefano Soatto. Positive-congruent training: Towards regression-free model updates. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. URL https://arxiv.org/abs/2011.09161.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. arXiv:2505.09388, 2025.

A PCFL: IMPLEMENTATION, COHORT, FEATURES AND SPLITS

stage model (`direct_nn`)	value
hidden widths / activation / dropout	$[256, 128]$ / ReLU / 0.2 after each hidden layer
normalisation / feature scaling	none / none (17 continuous columns enter raw: five vitals as mean and standard deviation, GCS mean and minimum, pain mean and maximum, $ETCO_2$ mean and standard deviation, age)
output / loss	82 logits / unweighted softmax cross-entropy
optimiser / learning rate / schedule / weight decay	Adam / 10^{-3} / none / 0
batch size / epochs / checkpoint	256 / 200 / last epoch (best-validation epoch tested and rejected)
weight initialisation	PyTorch `nn.Linear` default (Kaiming-uniform, $a = \sqrt{5}$)
seeds	$\{0, 1, 2\}$ per stage; splits 42 / 1234; cross-fitting fold seed $100 \cdot$ seed $+$ fold
hardware	64 CPU cores, 3× NVIDIA A30 (24 GB); stage models and experts run concurrently on CPU/GPU, the LLM arm on two GPUs

Table 2: Stage-model hyperparameters and hardware, read from the code.

Table 2 lists the stage-model hyperparameters and the hardware. The reported feature set is called config D in the generated tables.

Cohort. NEMSIS 2024 public release; `Pub_PCRevents.txt` (dispatch, complaint location and organ system, scene/context and age), `FACTPCRVITAL.txt` (timestamped vitals), `PCRVITALECGGROUP.txt` (ECG rhythm), with medications and procedures read only to be excluded. About 1,000 encounters were sampled per protocol for the 82 protocols with enough volume (978–1,000 retained per class). Invalid and "not recorded" sentinels are stripped before code features are built. Procedure coverage is 69.27% of the cohort (683 distinct codes), and 94.9% of procedures with a recorded answer to "performed prior to this unit's care" were performed by the documenting unit, i.e. after the protocol decision.

Features. Every feature is an indicator, a small count, or a raw value carrying its own missing flag; nothing is scaled, normalised or imputed. Vitals are re-joined from the raw tables so that every value column gets its own missing indicator and an unmeasured vital never collapses onto a measured zero; glucose is retained, and ECG rhythm (multi-hot), GCS mean and minimum, AVPU, pain and $ETCO_2$ are added on the same terms. Age enters as a raw value, a missing flag and eight fixed clinical bands at 1, 5, 13, 18, 40, 65 and 80 years. Table 3 gives the block breakdown per stage; the information in the context block is carried by incident location type (169 codes) and age, the other three scene-time fields being near-constant. The merged `clinical_*` block sums four source indicators: the symptom vocabulary has 1,063 codes, the impression vocabulary 1,246, and 571 codes are in both, so the union is the 1,738-column stage-6 block, in which 43.7% of records carry a value of 2 (usually one ICD-10-CM code documented both as a symptom and as an impression) and three cells reach 3. This is not a label leak, since both fields are recorded before the protocol decision, but it does discard which field a shared code came from; the provenance-preserving alternative (the stage-5 matrix with the impression block kept separate, 2,676 columns) is trained and compared in App. C.

The capture-time audit in the data. Figure 6 contrasts the gain of the context block under the audited per-field timing with the uniform placement of all five fields at every stage that the audit replaced: uniform placement credited stages 1–3 with 7.8, 7.4 and 7.0 top-1 % that the fields cannot legitimately supply, and the two placements coincide from the scene stage on. Table 5 gives the accepted gain of the audited block at every stage; it peaks at the scene stage ($+5.53$ top-1), where the four scene-time fields legitimately arrive, and the right panel of Figure 6 shows the same pivot as the per-transition gain of STEP against the input dimension.

Splits and protocols. Outer split: 90/10 stratified on protocol, seed 42 (73,780 / 8,198). Inner split: 10% of the training split stratified on protocol, seed 1234 (66,402 / 7,378). The six stage datasets are built from the same cohort in the same row order, so the split is identical across stage directories; the loader raises unless the label and identifier vectors match exactly. Two training

stage	block revealed	D_t	new columns	composition
1	Dispatch	56	56	disp. 46; age 10
2	+Complaint location	66	10	disp. 46; age 10; loc. 10;
3	+Organ system	78	12	disp. 46; age 10; loc. 10; organ 12
4	+Symptoms, scene context	1,321	1,243	disp. 46; age 10; loc. 10; organ 12 clinic. 1,063; scene 190
5	+Vitals	1,398	77	disp. 46; age 10; loc. 10; org. 12 clinic. 1,063; scene 190; vit. 77
6	+Impressions	2,073	675	disp. 46; age 10; loc. 10; org. 12 clinic. 1,738; scene 190; vit. 77

Table 3: Cumulative feature stages of current PCFL framework. D_t is counted from the built datasets; every transition adds columns and drops none (the stage-$(t-1)$ columns are a subset of the stage-t columns with identical values at stages 1–5; at stage 6 the merged clinical vocabulary changes the values of the shared symptom columns, App. C). The context block is 10 columns (age) before symptoms arrive and 190 after, the capture-time audit of §4 made visible in the data; the clinical block is one merged symptom/impression vocabulary (1,063 codes at stages 4–5; 1,738 at stage 6).

field	NEMSIS element	capture time	evidence from the element's own definition
`ctx_first_ems`	eScene.01	scene	"first EMS unit … *on the scene*": an arrival fact
`ctx_num_patients`	eScene.06	scene	"how many patients were *at the scene*"; NONE = "crew were unable to locate anyone"
`ctx_mci`	eScene.07	scene	"*overwhelmed* … by the number and severity of victims at an emergency scene"
`ctx_location_type`	eScene.09	scene	a two-tier pick list that *EMS personnel* choose from, not an address derivation
`ctx_age_years`	ePatient.15/16	dispatch	a demographic with no scene-observation dependency; a CAD element in its own right (eCad.66)

Table 4: Per-field capture timing from the NEMSIS element definitions. Scene-time fields may not appear before stage 4; age is the one dispatch-time field, and the one whose placement rests on a judgement rather than on the element's own wording.

protocols are used throughout, and no number mixes them. *Fit-rows* models are trained on the 66,402 fit rows, so their predictions on the 7,378 inner-validation rows are out-of-sample; these are the quantities every acceptance test reads. *Full-train* models are trained on all 73,780 training rows, which is what would be deployed, and their test predictions are the reported quantities. Selecting under one protocol and reporting under the other is an approximation; the measured difference between them is about 0.3 to 0.4 %, and at stage 6 it is $+0.00$ top-1 and $+0.07$ top-3.

Per-stage detail. The single seeds, the three-seed ensemble and STEP at every stage with top-5 and top-10, the per-seed values and the seed spread (the source of the 0.70-point figure of §3.2) are in App. C (Table 16). Table 6 gives the four alternative architectures against their own MLP control on the project's held-out partition. None of them improves on the per-stage MLP on test at any stage: the one candidate accepted on validation, the LSTM at stage 4, is 0.04 top-1 % behind on its partition's test rows, and by stage 5 all four are 2.2 to 4.7 % behind ($p < 10^{-7}$).

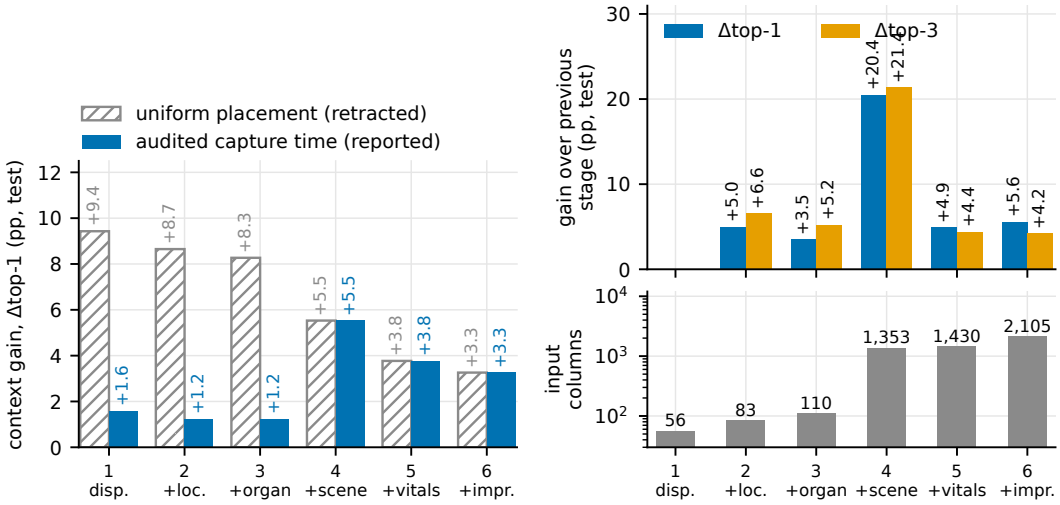

Figure 6: Left: the gain from the scene/context block under the audited capture timing against the uniform placement the audit replaced. Right: per-transition gain of STEP (top) and the input dimension (bottom; log axis).

stage	no ctx. top-1	timed top-1	Δtop-1	Δtop-3	val p_1	val p_3	test p_1	test p_3
1	18.83	20.42	+1.59	+2.09	5.3×10^{-9}	1.1×10^{-11}	4.0×10^{-9}	2.3×10^{-12}
2	24.03	25.27	+1.24	+1.84	8.3×10^{-5}	2.1×10^{-9}	4.2×10^{-5}	1.6×10^{-8}
3	27.48	28.71	+1.23	+2.35	2.8×10^{-4}	1.2×10^{-10}	4.6×10^{-5}	1.5×10^{-13}
4	43.19	48.72	+5.53	+5.50	4.2×10^{-37}	2.6×10^{-28}	9.5×10^{-37}	1.3×10^{-40}
5	49.60	53.37	+3.77	+3.65	2.6×10^{-16}	3.2×10^{-20}	4.0×10^{-21}	3.4×10^{-23}
6	55.60	58.86	+3.26	+2.67	1.2×10^{-19}	4.9×10^{-19}	4.6×10^{-17}	5.7×10^{-15}

Table 5: The scene/context + age block at its audited capture time: the same features without context versus with context placed when each field legitimately exists. Accepted at every stage on both metrics. The gain is a stage-4 effect (where the four scene-time fields legitimately arrive), not a dispatch-time one. Test deltas and test McNemar p are reported next to the validation acceptance test.

st.	architecture (selected config)	params	seed top-1 (mean)	ens. top-1	top-3	top-5	top-10	NLL	Δ_1 / Δ_3 vs MLP (p)
1	mlp (lr $3{\cdot}10^{-4}$, wide)	182k	20.44	20.57	38.39	48.32	63.06	3.213	–
	lstm (lr $3{\cdot}10^{-4}$, wide)	585k	20.38	20.53	38.22	48.08	62.75	3.216	-0.04 (0.91) / -0.17 (0.42)
	transformer (lr $3{\cdot}10^{-4}$, wide)	1113k	20.48	20.40	38.38	48.19	62.71	3.225	-0.17 (0.45) / -0.01 (1)
	cascaded_mlp (lr $3{\cdot}10^{-4}$, wide)	3683k	20.55	20.58	38.45	48.17	62.91	3.215	$+0.01$ (1) / $+0.06$ (0.79)
	cascaded_transformer (default)	1253k	20.20	20.60	37.90	47.90	62.61	3.242	$+0.04$ (0.92) / -0.49 (0.034)
2	mlp (lr $3{\cdot}10^{-4}$, wide)	187k	24.64	24.66	44.18	54.59	68.46	2.998	–
	lstm (lr $3{\cdot}10^{-4}$, wide)	588k	24.44	24.74	43.99	54.62	68.33	3.004	$+0.07$ (0.73) / -0.20 (0.41)
	transformer (lr $3{\cdot}10^{-4}$, wide)	1117k	24.36	24.54	43.94	54.51	68.25	3.013	-0.12 (0.53) / -0.24 (0.32)
	cascaded_mlp (lr $3{\cdot}10^{-4}$, wide)	3683k	24.64	24.81	44.40	54.71	68.43	2.997	$+0.15$ (0.36) / $+0.22$ (0.26)
	cascaded_transformer (default)	1253k	24.19	24.54	43.62	54.00	68.11	3.030	-0.12 (0.61) / -0.56 (0.027)
3	mlp (lr $3{\cdot}10^{-4}$, wide)	193k	27.82	28.03	48.95	59.60	72.93	2.821	–
	lstm (lr $3{\cdot}10^{-4}$, wide)	592k	27.51	27.53	48.87	59.60	72.98	2.828	-0.50 (0.018) / -0.09 (0.76)
	transformer (lr $3{\cdot}10^{-4}$, wide)	1121k	27.48	27.88	48.66	59.36	73.16	2.840	-0.15 (0.48) / -0.29 (0.22)
	cascaded_mlp (lr $3{\cdot}10^{-4}$, wide)	3683k	27.68	27.99	48.78	59.59	72.93	2.827	-0.04 (0.89) / -0.17 (0.45)
	cascaded_transformer (default)	1253k	26.77	27.26	48.23	59.22	72.82	2.864	-0.77 (0.0014) / -0.72 (0.0044)
4	mlp (lr $3{\cdot}10^{-4}$, wide)	829k	47.49	48.17	70.05	78.56	87.62	1.972	–
	lstm (lr $3{\cdot}10^{-4}$, wide)	911k	46.97	47.94	69.94	78.35	87.34	2.009	-0.23 (0.44) / -0.11 (0.72)
	transformer (lr $3{\cdot}10^{-4}$, wide)	1441k	45.95	47.28	69.18	78.13	87.23	2.042	-0.89 (0.0031) / -0.88 (0.0036)
	cascaded_mlp (lr $3{\cdot}10^{-4}$, wide)	3683k	47.46	48.06	69.52	78.26	87.40	1.992	-0.11 (0.69) / -0.54 (0.026)
	cascaded_transformer (default)	1253k	46.19	47.98	69.36	78.04	86.80	2.060	-0.20 (0.54) / -0.70 (0.022)
5	mlp (lr $3{\cdot}10^{-4}$, wide)	869k	51.48	52.48	74.69	82.58	89.99	1.813	–
	lstm (lr $3{\cdot}10^{-4}$, wide)	932k	49.82	50.95	73.36	81.24	89.14	1.889	-1.52 (7.6e-05) / -1.33 (6.9e-05)
	transformer (lr $3{\cdot}10^{-4}$, wide)	1462k	49.41	50.74	73.33	81.29	89.45	1.918	-1.73 (9.6e-06) / -1.35 (6.4e-05)
	cascaded_mlp (lr $3{\cdot}10^{-4}$, wide)	3683k	51.62	52.33	73.70	81.82	89.66	1.833	-0.15 (0.64) / -0.99 (0.00024)
	cascaded_transformer (default)	1253k	49.91	52.07	73.79	81.57	89.17	1.901	-0.40 (0.27) / -0.90 (0.0056)
6	mlp (lr $3{\cdot}10^{-4}$, wide)	1214k	57.32	58.45	78.60	85.59	91.99	1.591	–
	lstm (lr $3{\cdot}10^{-4}$, wide)	1105k	55.70	56.92	77.86	84.79	91.35	1.695	-1.54 (4.7e-05) / -0.74 (0.021)
	transformer (lr $3{\cdot}10^{-4}$, wide)	1636k	55.20	57.00	78.02	85.00	91.41	1.715	-1.45 (0.00023) / -0.59 (0.072)
	cascaded_mlp (lr $3{\cdot}10^{-4}$, wide)	3683k	56.66	57.26	78.08	85.25	91.73	1.610	-1.20 (1.4e-05) / -0.52 (0.03)
	cascaded_transformer (default)	1253k	55.02	56.75	77.46	84.35	91.03	1.700	-1.71 (1.2e-06) / -1.15 (0.00022)

Table 6: Matched architecture comparison on the feature contract, canonical partition (fit rows 66,402 / inner validation 7,378 / test 8,198), identical features at every stage (age from stage 1, scene/context block from stage 4), seeds $\{0, 1, 2\}$, one policy for every architecture: train on the fit rows, keep the best-validation-NLL epoch (early stop after 25 epochs without improvement), report the 3-seed geometric-mean ensemble on test; the single-seed mean is given for comparison and is not the ensemble. Each architecture is shown at the configuration a validation sweep over learning rate (10^{-3}, $3{\cdot}10^{-4}$) and width (128 / [256,128] against 256 / [512,256]) selected on seed-0 validation top-1 at stages 4–6 (default: lr 10^{-3}, narrow). Δ: ensemble minus the selected MLP ensemble, exact paired McNemar p on the 8,198 test records (descriptive; 48 comparisons). Cascades train all six stages jointly (sum of per-stage cross-entropies). Parameters are for one seed at that stage.

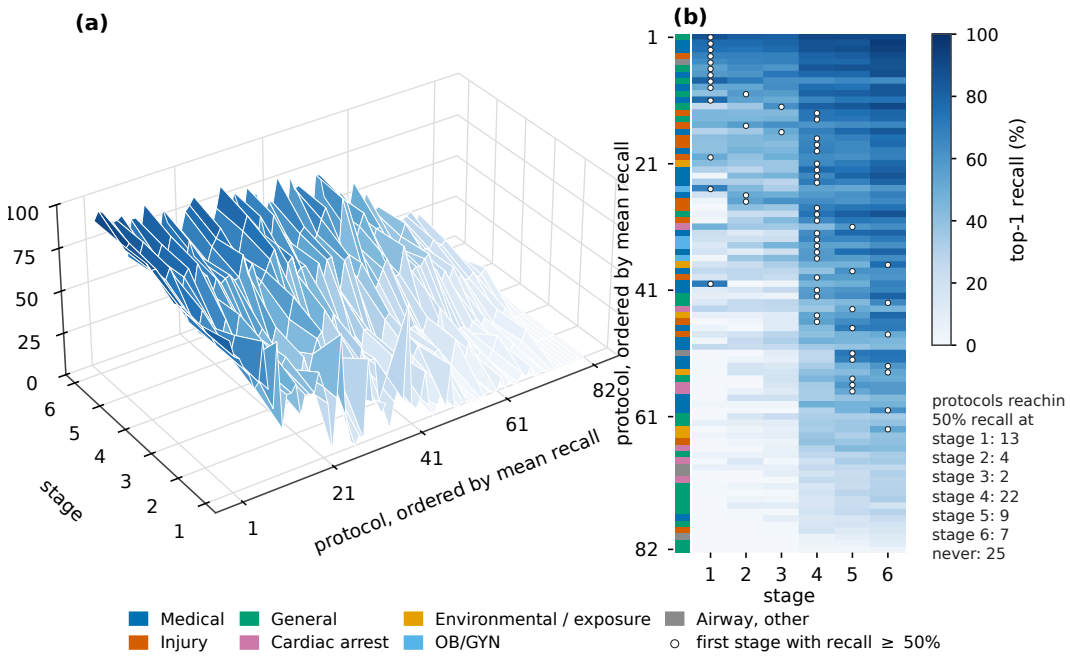

Figure 7: When does protocol discrimination become possible? (a) STEP top-1 recall of every protocol at every stage (balanced test split, 100 records per protocol, clean boundary), protocols ordered by their mean recall over the six stages. (b) The same matrix as a heatmap, with the coarse family on the left and a dot at the first stage at which a protocol's recall reaches 50%; the counts of protocols reaching 50% at each stage, and never, are printed beside the panel. No number should be read off the surface; the heatmap and the printed counts carry the values.

B SEP: SERVING, PROMPT, SUBSETS AND PER-STAGE RESULTS

Serving and decoding. Qwen3-32B-AWQ (Yang et al., 2025) is served locally by vLLM (Kwon et al., 2023) with decoding pinned for reproducibility (Table 7). The 2,050-record checkpoint subset reproduces as an exact prefix of the full test run under greedy decoding, which is the determinism claim of §5.5; the per-record wall-clock quoted there is the run's elapsed time over the 8,198 test records at request concurrency 32.

model	Qwen3-32B-AWQ (4-bit AWQ, `awq_marlin`)
serving	vLLM, max model length 8,192, prefix caching, eager execution, GPU memory fraction 0.88
parallelism	tensor parallel 1, pipeline parallel 2 (tensor parallelism hung in the first forward pass on this host's PCIe-only GPU pair)
decoding	temperature 0 (greedy), top-p 1.0, max 80 output tokens, seed 0, `chat_template_kwargs={enable_thinking: false}`
retries	transport failures only; a parse failure is a miss (a retry is a no-op under greedy decoding)

Table 7: Serving and decoding configuration of the LLM arm.

Subsets and metrics. Three evaluation subsets are derived from the split at load time, never read from a file, and identified by a SHA-256 over their sorted record identifiers plus an order hash, checked at load: `val820` (ten inner-validation records per protocol; every selection decision), `test2050` (25 test records per protocol; a checkpoint) and `test8198` (all test rows; the reported numbers). Two accuracies are defined: valid-only over parsed responses, and *overall*, which counts an unparseable response as a miss. Overall is the primary metric because it is what a deployment sees; the measured invalid rate is 0.00% at every stage, so the two coincide. Table 8 gives the arm at every stage with the seed spread of the $k = 3$ configuration and the within-arm verdicts. In the list-order ablation on `val820`, a second fixed candidate ordering changed zero-shot top-1 by $+0.7$, -1.5, -0.9, $+3.1$, $+2.1$ and $+1.0$ % at stages 1–6 and cleared the acceptance threshold at stage 4 only; thinking mode ran at 13.8 to 66.2 times the wall-clock of the adopted configuration (mean 34.8), which would have turned the 4.6-hour test run into about 160 hours.

Against STEP. Figure 13 and Table 9 pair every prompted run with STEP on the same 8,198 test records. The prompted model reproduces the shape of the stage curve, including the jump at the scene stage, but the distance grows with the evidence: with three demonstrations it is 5.71 top-1 % behind STEP at dispatch, 27.41 after vital signs and 26.72 after impressions, and all 48 paired tests (four prompted runs, six stages, two metrics) favour STEP at $p < 0.05/48$. Within the prompted arm, the seed-0 $k = 3$ run is accepted over zero-shot at every stage but the second (Table 8). A stage-6 record is 2,105 mostly binary columns, which the supervised model consumes directly and the prompted model reads as a serialised field list, and it falls furthest behind where that list is longest, consistent with a limit of the evidence format rather than of model scale (no scale sweep was run). Earlier versions of this comparison used the three-seed ensemble without pooling as the supervised comparator, because the pooled model's per-record outputs had not been exported; the last two columns of Table 9 keep those gaps under their correct label. Against STEP every gap widens slightly, by the pool's own gain.

MedGemma-27B-text: protocol and caveats. The second prompted arm is MedGemma-27B-text (Sellergren et al., 2025), a medically tuned instruction model, served locally by vLLM in bfloat16 (pipeline parallel over three A30 GPUs, maximum model length 4,096, greedy decoding, seed 0, at most 160 output tokens because its tokenizer splits digits individually, no thinking mode exists for it) under the identical staged-evidence prompts, subsets and metric code as Qwen3, with the record identifier, medications and procedures never rendered. It was run at Qwen3's selected configuration (the same prefix, candidate order, $k \in \{0, 3\}$ and demonstration seed 0), fixed before any MedGemma test row was read, so it answers "same prompt, different model" and not "MedGemma at its own best". Every MedGemma row is matched to the same PcrKey in the Qwen3 run and in the supervised arm; an unmatched row is an error, not a warning. Table 10 gives the paired results at every stage; Table 11 the three models at depths 1, 3, 5 and 10; Figure 14 the same surface

stage	ZS top-1	ZS top-3	k=3 top-1 (mean ± sd)	k=3 top-3	invalid%	ZS→ k=3 test (seed 0)
1	14.99	23.33	14.98 ± 0.72	22.79	0.00	selected (top-1)
2	15.55	23.58	16.00 ± 0.39	24.75	0.00	rejected
3	15.56	24.52	16.35 ± 0.63	26.14	0.00	selected (top-3)
4	22.69	33.34	25.74 ± 0.54	35.99	0.00	selected (top-1, top-3)
5	24.07	33.97	27.10 ± 0.68	36.67	0.00	selected (top-1, top-3)
6	29.79	39.88	33.37 ± 0.74	42.68	0.00	selected (top-1, top-3)

Table 8: The prompted LLM arm (Qwen3-32B-AWQ, greedy, thinking disabled) on the same 8,198 test records and the same six stages. Zero-shot and k=3 demonstrations (mean ± std over three demonstration seeds); the invalid rate is the share of unparseable responses, counted as misses in the overall accuracy shown. Right: the verdict of the within-arm zero-shot→ k=3 acceptance test on test rows, paired on the seed-0 k=3 run (test was read once per final configuration; all selection was on `val820`).

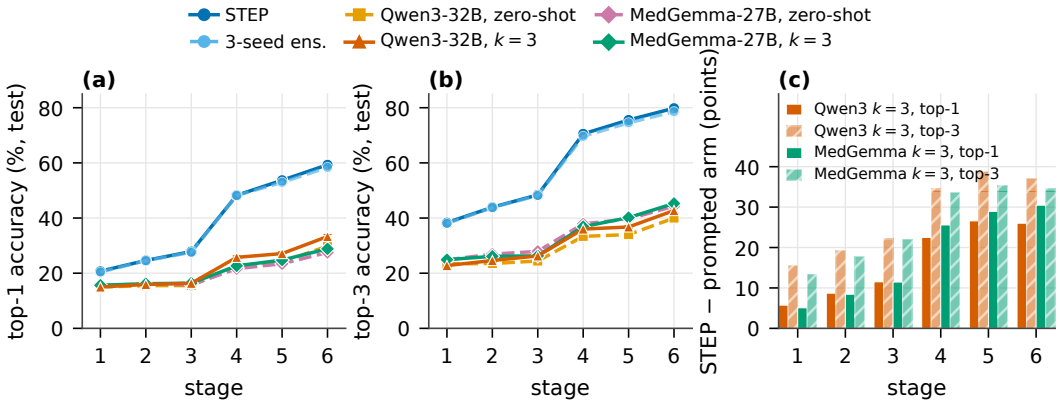

Figure 8: SEP on the same 8,198 test records: (a) top-1 and (b) top-3 by stage for STEP, the ensemble, Qwen3-32B-AWQ and MedGemma-27B-text (c) STEP minus each arm at k=3

over stage and depth; Figure 5 and Table 28 the family-level confusions and macro precision, recall and F1; Figures 9 and 10–12 and Table 14 the per-protocol view at stage 6. Five caveats qualify the arm. (i) *Precision differs between the arms*: Qwen3 ran as a 4-bit AWQ checkpoint, MedGemma in bf16, because no usable quantised checkpoint of MedGemma-27B-text existed; bf16 is the more favourable setting, so it cannot explain a MedGemma deficit, but it is a real asymmetry. (ii) *Sliding-window asymmetry*: MedGemma's architecture uses a 1,024-token local attention window on five of every six layers, and the fixed candidate list alone is 1,290 tokens under its tokenizer, so on those layers the model attends to the query record without the candidate list in its local window, whereas Qwen3 used full attention throughout; this is reported, not compensated for, because changing the prompt would break the frozen-prompt rule. (iii) *Seed 0 only*: Qwen3's k=3 result is a mean over three demonstration seeds with a stage-6 top-1 spread of 0.75 %, MedGemma has one seed, so the paired MedGemma-vs-Qwen3 tests compare seed 0 with seed 0 and the cross-arm differences are of the order of Qwen3's own seed spread. (iv) *The prompt configuration was selected on Qwen3*, not on MedGemma, so the arm may understate what MedGemma would reach with its own selection. (v) *Reproducibility is not exact across batch composition*: on a 2,050-record checkpoint subset, 17.0% of rows received a different ranked list when scored standalone and inside the full 8,198-row pass (batched floating-point reduction order), with single-digit row counts changing a top-1 or top-3 hit, whereas the Qwen3 arm reproduced exactly; the invalid-response rate is at most 0.21% (Table 10). The reported arms comprise 295,128 prompted calls (196,752 for Qwen3 over four conditions per stage, 98,376 for MedGemma over two) and 531 million prompt tokens in all, which is the scale behind the cost remark of §5.5.

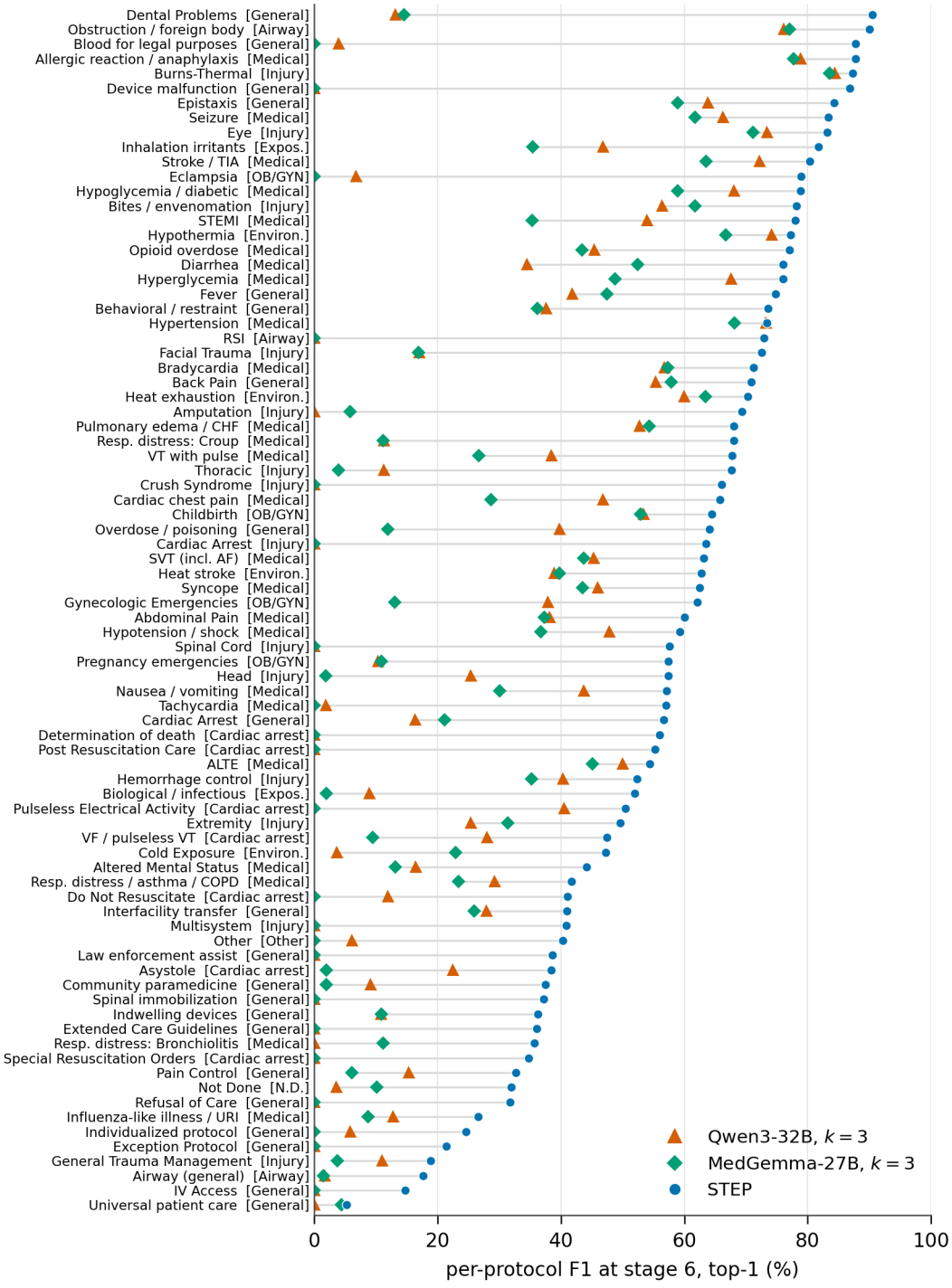

Figure 9: F1 comparison of STEP vs. Qwen3 and MedGemma LLM models per protocol

continued

Table 14: Per-protocol stage-6 results on the balanced test split (100 records per protocol), in %: recall (R), precision (P) and F1 of the top-1 prediction for STEP, and top-1 recall and F1 of the two prompted arms (k=3, seed 0). The index is the row order of the confusion heatmaps (Figs. 10–12); protocols are grouped by coarse family.

#	protocol (family)	STEP			Qwen3 k=3		MedGemma k=3	
		R	P	F1	R	F1	R	F1
1	Airway	12	33	18	1	2	1	1
2	Airway-Obstruction/Foreign Body	86	95	90	69	75	74	77
3	Airway-Rapid Sequence Induction (RSI-Paralytic)	74	72	73	0	0	0	0
4	Cardiac Arrest-Asystole	34	44	38	50	23	1	2
5	Cardiac Arrest-Determination of Death /Withholding Resuscitative Efforts	63	50	56	0	0	0	0
6	Cardiac Arrest-Do Not Resuscitate	47	36	41	7	12	0	0
7	Cardiac Arrest-Post Resuscitation Care	58	53	55	0	0	0	0
8	Cardiac Arrest-Pulseless Electrical Activity	58	45	50	32	37	0	0
9	Cardiac Arrest-Special Resuscitation Orders	29	44	35	0	0	0	0
10	Cardiac Arrest-Ventricular Fibrillation/Pulseless Ventricular Tachycardia	42	55	47	22	28	5	9
11	Environmental-Cold Exposure	39	60	47	2	4	15	23
12	Environmental-Heat Exposure/Exhaustion	78	64	70	71	60	72	63
13	Environmental-Heat Stroke/Hyperthermia	53	77	63	29	39	30	40
14	Environmental-Hypothermia	78	76	77	76	74	59	67
15	Exposure-Airway/Inhalation Irritants	76	88	82	33	47	23	35
16	Exposure-Biological/Infectious	53	51	52	5	9	1	2
17	General-Back Pain	84	61	71	81	56	64	58
18	General-Behavioral/Patient Restraint	78	70	74	67	38	43	36
19	General-Cardiac Arrest	51	64	57	39	16	93	21
20	General-Community Paramedicine /Mobile Integrated Healthcare	33	43	38	6	9	1	2
21	General-Dental Problems	90	91	90	7	13	8	15
22	General-Epistaxis	86	83	84	82	64	71	59
23	General-Exception Protocol	18	26	21	0	0	0	0
24	General-Extended Care Guidelines	29	48	36	0	0	0	0
25	General-Fever	80	70	75	32	42	37	47
26	General-Individualized Patient Protocol	17	45	25	5	6	0	0
27	General-Indwelling Medical Devices/Equipment	29	48	36	6	11	6	11
28	General-Interfacility Transfers	63	30	41	43	28	58	26
29	General-IV Access	10	28	15	0	0	0	0
30	General-Law Enforcement - Assist with Law Enforcement Activity	39	38	39	0	0	0	0
31	General-Law Enforcement - Blood for Legal Purposes	90	86	88	2	4	0	0
32	General-Medical Device Malfunction	86	88	87	0	0	0	0
33	General-Overdose/Poisoning/Toxic Ingestion	65	63	64	58	40	7	12
34	General-Pain Control	32	33	33	17	15	4	6
35	General-Refusal of Care	30	34	32	0	0	0	0
36	General-Spinal Immobilization/Clearance	34	41	37	0	0	0	0
37	General-Universal Patient Care/Initial Patient Contact	3	21	5	0	0	6	4
38	Injury-Amputation	70	69	69	0	0	3	6
39	Injury-Bites and Envenomations-Land	77	79	78	44	56	50	62
40	Injury-Bleeding/Hemorrhage Control	51	54	52	51	40	44	35
41	Injury-Burns-Thermal	89	86	87	80	83	76	84
42	Injury-Cardiac Arrest	60	67	63	0	0	0	0
43	Injury-Crush Syndrome	75	59	66	0	0	0	0
44	Injury-Extremity	65	40	50	20	25	35	31
45	Injury-Eye	79	88	83	62	73	58	71
46	Injury-Facial Trauma	74	71	73	75	17	90	17
47	Injury-General Trauma Management	15	25	19	10	11	4	4
48	Injury-Head	66	51	57	18	25	1	2
49	Injury-Multisystem	40	42	41	0	0	0	0
50	Injury-Spinal Cord	68	50	58	0	0	0	0
51	Injury-Thoracic	71	65	68	6	11	2	4
52	Medical-Abdominal Pain	69	53	60	72	38	70	37
53	Medical-Allergic Reaction/Anaphylaxis	93	83	88	95	79	94	78
54	Medical-Altered Mental Status	47	42	44	62	16	52	13
55	Medical-Apparent Life Threatening Event (ALTE)	43	74	54	34	50	30	45
56	Medical-Bradycardia	73	70	71	44	57	49	57
57	Medical-Cardiac Chest Pain	74	59	66	90	47	91	29
58	Medical-Diarrhea	81	72	76	21	34	38	52
59	Medical-Hyperglycemia	84	69	76	80	68	38	49
60	Medical-Hypertension	80	68	73	75	73	61	68
61	Medical-Hypoglycemia/Diabetic Emergency	80	78	79	68	68	76	59
62	Medical-Hypotension/Shock (Non-Trauma)	62	57	59	44	48	29	37
63	Medical-Influenza-Like Illness/Upper Respiratory Infection	23	32	27	15	13	6	9
64	Medical-Nausea/Vomiting	66	50	57	55	44	29	30
65	Medical-Opioid Poisoning/Overdose	84	71	77	39	45	80	43
66	Medical-Pulmonary Edema/CHF	68	68	68	63	52	47	54

continued

#	protocol (family)	STEP			Qwen3 k=3		MedGemma k=3	
		R	P	F1	R	F1	R	F1
67	Medical-Respiratory Distress-Bronchiolitis	45	30	36	0	0	8	11
68	Medical-Respiratory Distress-Croup	69	67	68	6	11	6	11
69	Medical-Respiratory Distress/Asthma/COPD/Reactive Airway	43	41	42	80	29	87	23
70	Medical-Seizure	95	74	83	94	66	96	62
71	Medical-ST-Elevation Myocardial Infarction (STEMI)	85	72	78	45	54	24	35
72	Medical-Stroke/TIA	88	74	80	92	72	80	63
73	Medical-Supraventricular Tachycardia (Including Atrial Fibrillation)	71	57	63	72	45	43	44
74	Medical-Syncope	80	51	62	67	46	64	44
75	Medical-Tachycardia	59	55	57	1	2	0	0
76	Medical-Ventricular Tachycardia (With Pulse)	61	76	68	25	38	16	27
77	OB/GYN-Childbirth/Labor/Delivery	59	71	64	60	53	87	53
78	OB/GYN-Eclampsia	71	89	79	5	7	0	0
79	OB/GYN-Gynecologic Emergencies	64	60	62	38	38	8	13
80	OB/GYN-Pregnancy Related Emergencies	62	53	57	6	10	7	11
81	Not Done	26	41	32	2	4	6	10
82	Other	31	57	40	8	6	0	0

Exploration on `val820`. Three variants were explored on `val820` only and never read on test: a second candidate ordering, which moved zero-shot top-1 by between -1.5 and $+3.1$ % across the stages and cleared the threshold at one (details above); an alternative prefix, which was not carried into the test protocol; and thinking mode, which improved top-3 at five stages but at about 35 times the wall-clock and was rejected on cost. The candidate order is fixed because shuffling it per query would destroy the server's prefix cache.

Serving cost. Latency is the batch-1 per-record forward time on one CPU core (20 warm-up and 200 timed repeats, weights of the stage-6 STEP shapes): 39 MLPs (stages 1–5 and the joint stage-6 model, and seven experts, three seeds each) and the stage-6 pool, 7.71 M parameters in all, take 4.8 ms per record, and the three-seed stage-6 ensemble alone 0.53 ms. The prompted model's 235 ms is the elapsed wall-clock of the 8,198-record test run divided by its size, a throughput figure at request concurrency 32 that flatters the prompted model. Both keep the record on local hardware. The comparison argues for a division of labour rather than a replacement: the supervised ranker stays the stage predictor on structured evidence, and a language model would be the component to reach for where evidence is not yet coded, which NEMSIS, lacking a narrative field, cannot test; no language-model output feeds STEP or DPCS here.

The prompt. Figure 15 gives the template. The fixed prefix is byte-identical for a whole run and placed first, so the server's prefix cache covers it; the k demonstrations and the query record follow. Only the template is shown: a rendered prompt contains a patient record.

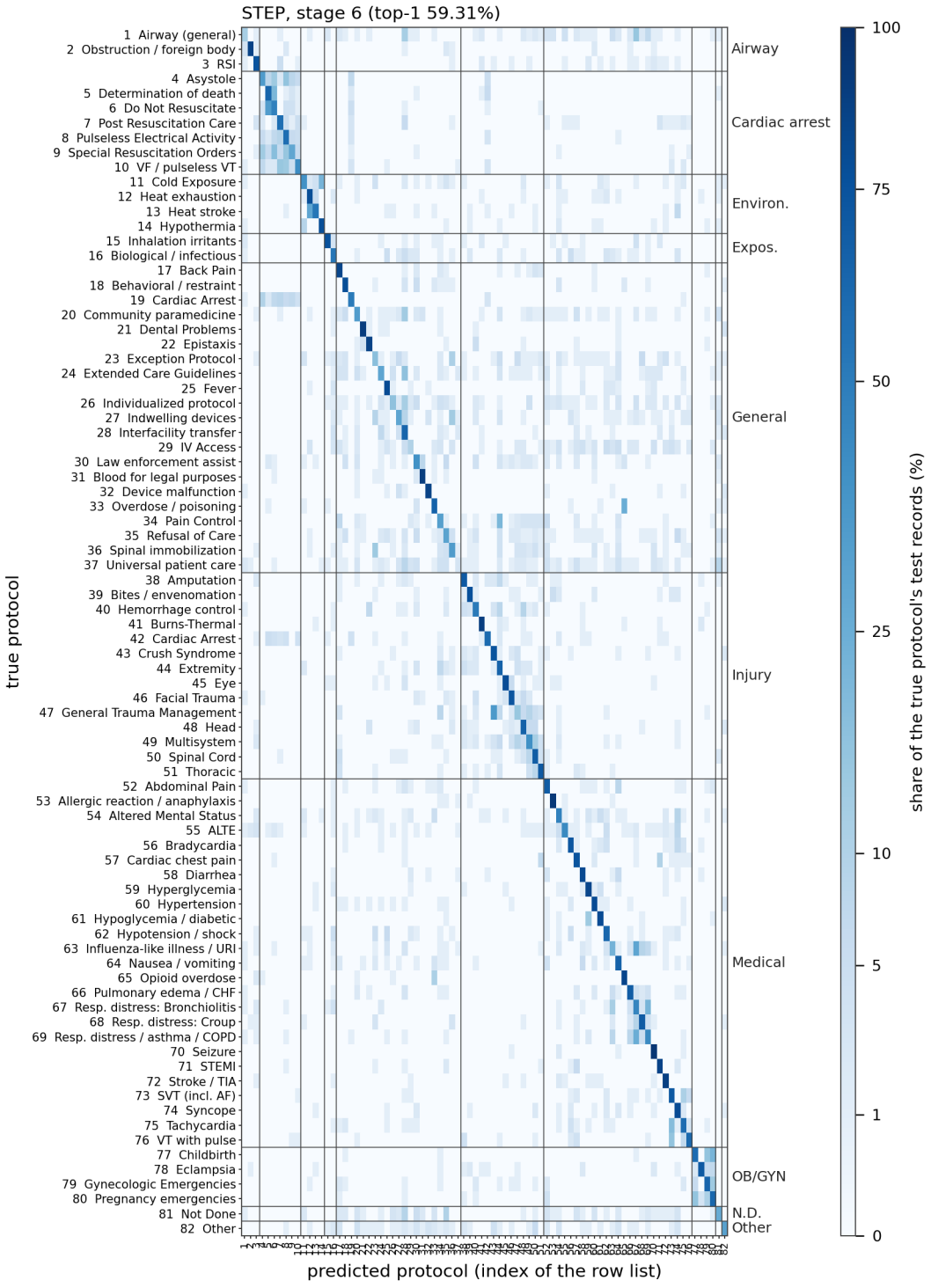

Figure 10: Complete confusion matrix of STEP vs. Qwen3 and MedGemma LLM models per protocol

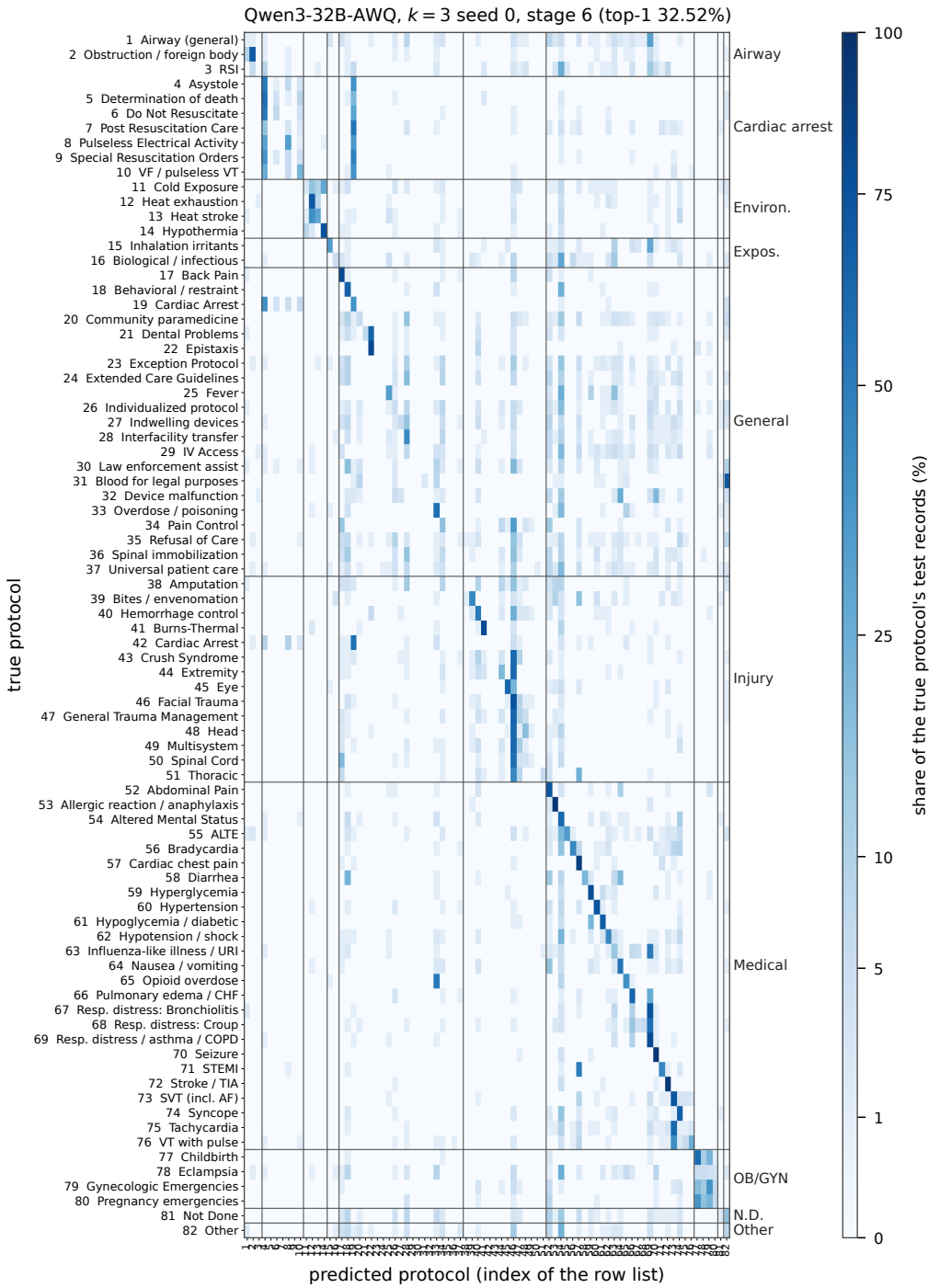

Figure 11: As Figure 10 for Qwen3-32B-AWQ (k=3, seed 0). Takeaway: mass collapses onto a few general and medical columns.

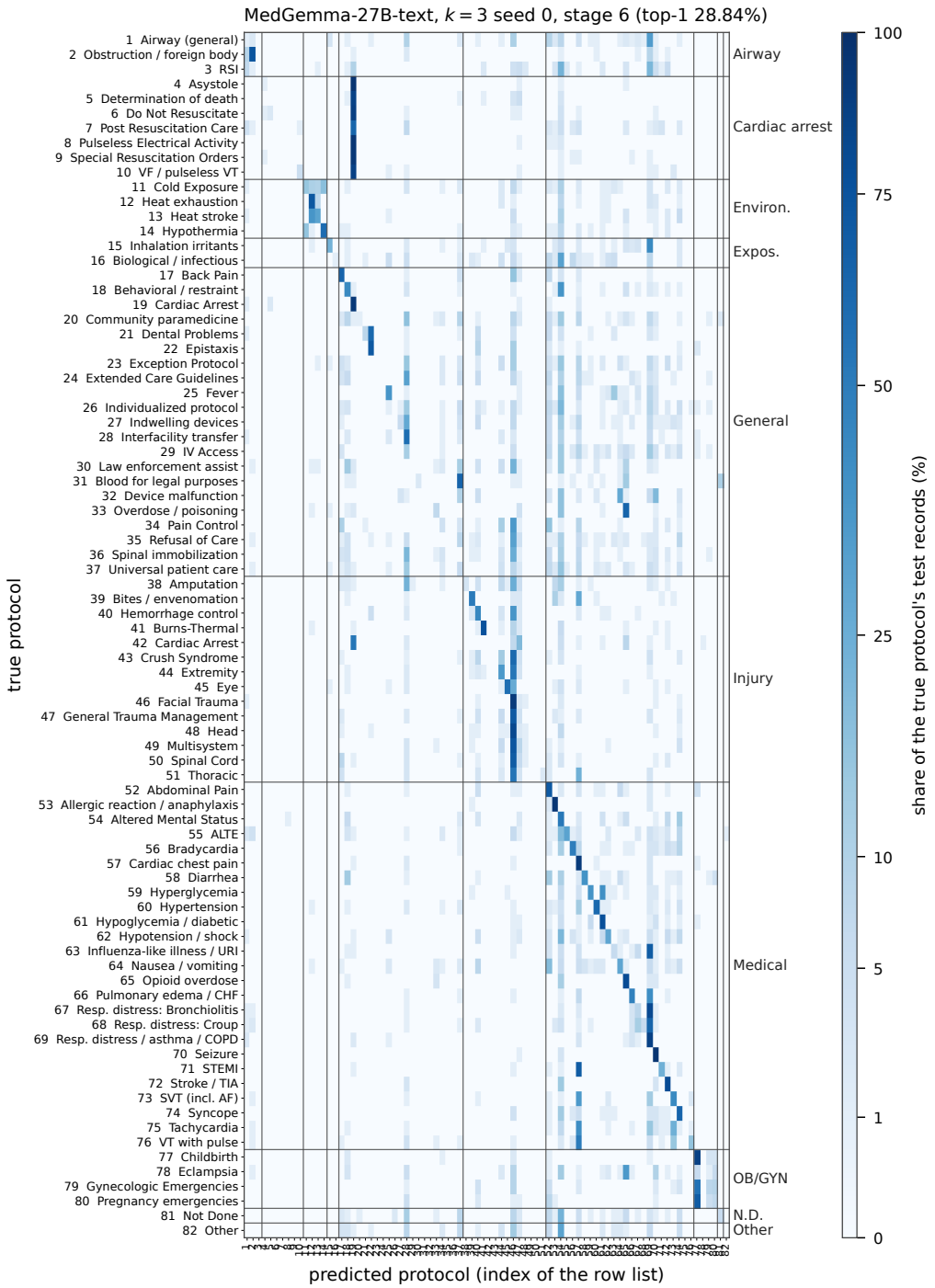

Figure 12: As Figure 10 for MedGemma-27B-text (k=3, seed 0). Takeaway: the same collapse as Qwen3, with a stronger pull towards the medical columns.

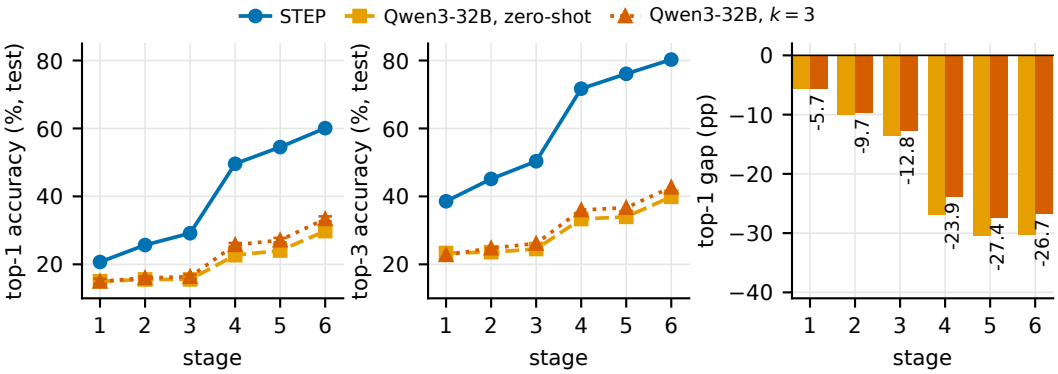

Figure 13: SEP against STEP across the six stages on the same 8,198 test records: top-1 (left), top-3 (centre) and the top-1 gap of the two prompted configurations, prompted minus STEP (right; labels for k=3, the mean over three demonstration seeds).

	zero-shot − STEP		k=3 − STEP (mean)			top-1 gap to the ensemble	
stage	top-1	top-3	top-1	top-3	largest p	zero-shot	k=3
1	−5.70	−15.22	−5.71	−15.77	1.7×10^{-34}	−5.59	−5.60
2	−10.12	−21.55	−9.68	−20.39	1.4×10^{-83}	−9.78	−9.34
3	−13.59	−25.82	−12.80	−24.20	7.9×10^{-125}	−13.70	−12.91
4	−26.91	−38.36	−23.86	−35.72	$<10^{-300}$	−26.62	−23.56
5	−30.45	−42.11	−27.41	−39.41	$<10^{-300}$	−29.17	−26.13
6	−30.30	−40.40	−26.72	−37.59	$<10^{-300}$	−29.07	−25.49

Table 9: The prompted arm against STEP on the same 8,198 test records, in percentage points. Each of the four prompted runs per stage (zero-shot; k=3 with demonstration seeds 0–2) is paired record for record with STEP by exact McNemar on top-1 and top-3; the largest of the eight p-values per stage is shown, and all 48 lie below the Bonferroni level $0.05/48$. The last two columns give the historical comparator, the three-seed ensemble without pooling.

stage	config	MedGemma top-1 / top-3		− Qwen3 seed 0 (p)		− STEP (p)		inv.%	ZS−
		top-1	top-3	Δ_1	Δ_3	Δ_1	Δ_3		
1	zero-shot	15.30	24.85	+0.30 (0.18)	$+1.46\,(4.7\times10^{-5})^*$	$-5.43\,(1.8\times10^{-40})^*$	$-13.49\,(7.7\times10^{-143})^*$	0.00	
	k=3	15.64	24.90	−0.01 (1)	$+1.22\,(7.5\times10^{-6})^*$	$-5.09\,(6.6\times10^{-34})^*$	$-13.44\,(1.6\times10^{-140})^*$	0.00	r
2	zero-shot	16.08	26.92	+0.60 (0.022)	$+3.40\,(1.9\times10^{-19})^*$	$-8.55\,(7.1\times10^{-77})^*$	$-16.99\,(3.8\times10^{-200})^*$	0.00	
	k=3	16.19	26.12	+0.37 (0.099)	$+2.01\,(2.8\times10^{-13})^*$	$-8.44\,(1.5\times10^{-72})^*$	$-17.80\,(1.8\times10^{-216})^*$	0.00	r
3	zero-shot	15.76	27.91	+0.24 (0.4)	$+3.50\,(4.8\times10^{-20})^*$	$-12.09\,(1.2\times10^{-125})^*$	$-20.48\,(5.6\times10^{-247})^*$	0.00	
	k=3	16.39	26.30	$+0.81\,(0.001)^*$	−0.07 (0.83)	$-11.45\,(9.1\times10^{-116})^*$	$-22.09\,(2.0\times10^{-289})^*$	0.00	r
4	zero-shot	21.63	37.89	−1.09 (0.002)	$+4.54\,(3.9\times10^{-28})^*$	$-26.62\,(<10^{-300})^*$	$-32.63\,(<10^{-300})^*$	0.00	
	k=3	22.66	36.87	$-2.56\,(1.9\times10^{-16})^*$	+1.09 (0.004)	$-25.58\,(<10^{-300})^*$	$-33.64\,(<10^{-300})^*$	0.00	r
5	zero-shot	23.43	39.66	−0.67 (0.094)	$+5.66\,(2.0\times10^{-41})^*$	$-30.28\,(<10^{-300})^*$	$-35.90\,(<10^{-300})^*$	0.00	
	k=3	24.76	40.14	$-1.61\,(9.3\times10^{-6})^*$	$+3.05\,(2.6\times10^{-15})^*$	$-28.95\,(<10^{-300})^*$	$-35.41\,(<10^{-300})^*$	0.21	selected
6	zero-shot	27.52	44.30	$-2.29\,(2.1\times10^{-8})^*$	$+4.31\,(3.5\times10^{-24})^*$	$-31.79\,(<10^{-300})^*$	$-35.52\,(<10^{-300})^*$	0.00	
	k=3	28.84	45.28	$-3.68\,(2.2\times10^{-21})^*$	$+1.76\,(1.1\times10^{-5})^*$	$-30.47\,(<10^{-300})^*$	$-34.55\,(<10^{-300})^*$	0.13	selected (top-1

Table 10: MedGemma-27B-text (bf16, seed 0, zero-shot and k=3, Qwen3's selected prompt configuration) on the same 8,198 test records of the balanced cohort, paired record for record with Qwen3-32B-AWQ seed 0 rerun and with STEP (Table 1). Δ: MedGemma minus the comparator in points, exact two-sided McNemar p; * marks the comparisons that clear the Bonferroni level of family C ($0.05/48$: the 24 MedGemma-vs-STEP and 24 MedGemma-vs-Qwen3 tests). inv.%: unparseable responses, counted as misses. Right: the within-arm zero-shot→ k=3 dual gate on test (accept if top-1 or top-3 improves at $p < 0.025$ with no harm at $p < 0.05$), applied post hoc for symmetry with Table 8; no MedGemma configuration was selected on validation.

stage	model	top-1	top-3	top-5	top-10
1 dispatch	STEP (Table 1)	**20.72**	**38.34**	**48.29**	**62.98**
	Qwen3-32B, zero-shot	14.99	23.38	28.78	36.28
	Qwen3-32B, k=3, 3 seeds	14.98 $\pm$0.73	22.77 $\pm$1.02	28.28 $\pm$1.54	35.01 $\pm$1.19
	MedGemma-27B, zero-shot	15.30	24.85	30.92	40.74
	MedGemma-27B, k=3, seed 0	15.64	24.90	31.25	40.52
2 +location	STEP (Table 1)	**24.63**	**43.91**	**54.53**	**68.04**
	Qwen3-32B, zero-shot	15.48	23.52	28.92	34.76
	Qwen3-32B, k=3, 3 seeds	15.93 $\pm$0.40	24.60 $\pm$0.69	29.87 $\pm$1.00	35.97 $\pm$0.50
	MedGemma-27B, zero-shot	16.08	26.92	33.24	42.64
	MedGemma-27B, k=3, seed 0	16.19	26.12	31.28	39.74
3 +organ system	STEP (Table 1)	**27.85**	**48.39**	**59.40**	**72.90**
	Qwen3-32B, zero-shot	15.52	24.41	30.25	37.53
	Qwen3-32B, k=3, 3 seeds	16.34 $\pm$0.66	26.26 $\pm$0.48	31.91 $\pm$0.46	38.72 $\pm$0.90
	MedGemma-27B, zero-shot	15.76	27.91	34.39	43.86
	MedGemma-27B, k=3, seed 0	16.39	26.30	31.79	40.72
4 +symptoms, scene	STEP (Table 1)	**48.24**	**70.52**	**78.89**	**87.70**
	Qwen3-32B, zero-shot	22.71	33.35	37.90	43.71
	Qwen3-32B, k=3, 3 seeds	25.74 $\pm$0.54	36.00 $\pm$0.48	40.25 $\pm$0.53	44.81 $\pm$0.59
	MedGemma-27B, zero-shot	21.63	37.89	44.36	52.61
	MedGemma-27B, k=3, seed 0	22.66	36.87	42.45	50.66
5 +vitals	STEP (Table 1)	**53.71**	**75.56**	**82.98**	**90.36**
	Qwen3-32B, zero-shot	24.10	34.00	37.89	43.71
	Qwen3-32B, k=3, 3 seeds	27.12 $\pm$0.69	36.76 $\pm$0.55	40.66 $\pm$0.40	45.23 $\pm$1.00
	MedGemma-27B, zero-shot	23.43	39.66	45.54	54.28
	MedGemma-27B, k=3, seed 0	24.76	40.14	45.29	53.24
6 +impressions	STEP (Table 1)	**59.31**	**79.82**	**86.36**	**92.45**
	Qwen3-32B, zero-shot	29.81	40.00	44.11	49.09
	Qwen3-32B, k=3, 3 seeds	33.29 $\pm$0.75	42.72 $\pm$0.70	45.49 $\pm$1.07	49.10 $\pm$0.58
	MedGemma-27B, zero-shot	27.52	44.30	50.63	59.21
	MedGemma-27B, k=3, seed 0	28.84	45.28	50.10	56.77

Table 11: STEP against the two prompted arms on the same 8,198 test records of the balanced 82-protocol cohort at depths 1, 3, 5 and 10 (test accuracy, %; best per stage and depth in bold). STEP is the clean-boundary model of Table 1. Both language models were served locally with greedy decoding under the staged-evidence prompt of App. B: Qwen3-32B-AWQ (4-bit, thinking disabled; run `t3`; k=3 is the mean $\pm$ sd over demonstration seeds 0–2) and MedGemma-27B-text (bf16, seed 0 only; App. B). Top-5 and top-10 never gated any decision.

		STEP			stage MLP (seed 0)			Qwen3-32B, k=3			MedGemma-27B, k=3		
k	stage	P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	1	19.1	20.7	16.0	19.0	20.4	16.3	12.4	15.0	10.5	11.4	15.6	10.4
	2	23.2	24.6	20.9	22.7	24.5	21.3	12.5	15.9	11.1	12.0	16.2	10.9
	3	26.9	27.8	24.5	25.3	27.3	23.8	13.2	16.3	11.1	13.2	16.4	11.3
	4	47.7	48.2	46.6	45.2	45.8	44.7	32.3	25.7	21.0	29.2	22.7	18.0
	5	53.5	53.7	52.2	50.0	50.3	49.0	32.2	27.1	23.0	29.5	24.8	20.1
	6	58.7	59.3	58.0	56.1	56.0	55.2	36.2	33.3	28.9	35.9	28.8	24.4
3	1	15.2	38.3	19.7	15.2	38.1	19.8	13.1	22.8	12.3	15.6	24.9	13.8
	2	15.9	43.9	22.1	15.8	43.4	22.1	12.5	24.6	12.6	15.4	26.1	14.6
	3	17.0	48.4	24.3	16.8	47.8	24.2	12.8	26.3	12.9	16.6	26.3	14.8
	4	25.4	70.5	36.8	25.2	67.3	36.0	26.6	36.0	21.3	27.3	36.9	21.6
	5	27.3	75.6	39.6	26.6	72.1	38.2	26.9	36.7	23.1	26.5	40.1	23.3
	6	30.1	79.8	42.9	30.2	76.0	42.1	30.0	42.7	27.7	30.8	45.3	26.9

Table 12: Macro-averaged precision (P), recall (R) and F1 over the 82 protocols on the 8,198 test records of the balanced cohort, in %, with "predicted positive" = "appears in the top k" (at k=1 this is ordinary multi-class macro P/R/F1; at k=3 every record contributes three predicted positives, so precision falls and recall equals top-3 accuracy). Because the cohort is balanced (100 test records per protocol), macro and support-weighted averages coincide to the first decimal. STEP and the stage MLP are clean-boundary models of Table 1; the prompted arms ran on the same family (Qwen3: mean over demonstration seeds 0–2; MedGemma: seed 0).

st.	method	micro top-1	micro top-3	macro $F1_1$	macro $F1_3$	EM top-1	EM top-3	EM macro $F1_1$
4	ensemble	19.49	37.18	15.62	12.69	37.85	75.50	21.01
	static pool	18.22	32.50	15.28	12.25	46.80	82.27	18.91
	STEP	18.31	33.49	15.07	12.18	48.27	83.69	18.17
	static pool, no class bias	19.48	37.11	15.62	12.66	58.83	82.83	14.34
	STEP, no class bias	19.76	37.28	15.63	12.60	59.50	83.81	13.63
	masked ens.	18.87	35.98	15.27	12.56	43.50	79.95	21.49
5	ensemble	20.92	39.40	17.30	13.79	53.06	84.84	24.55
	static pool	20.11	35.86	17.12	13.47	57.54	85.95	23.57
	STEP	20.23	36.52	17.09	13.48	54.24	86.17	23.90
	static pool, no class bias	21.63	40.86	17.66	13.96	60.78	85.39	18.01
	STEP, no class bias	21.65	40.97	17.76	14.00	62.06	86.23	17.85
	masked ens.	20.61	38.67	17.30	13.61	49.76	83.22	24.60
6	ensemble	24.42	47.52	20.04	15.60	55.38	85.45	28.48
	static pool	23.20	41.11	19.69	15.03	61.06	88.32	27.90
	STEP	23.24	41.92	19.58	15.01	60.41	88.41	26.98
	static pool, no class bias	25.46	48.06	20.60	15.76	61.69	85.26	21.44
	STEP, no class bias	25.22	47.65	20.38	15.68	62.78	87.71	17.99
	masked ens.	23.66	43.46	19.46	14.91	50.05	84.50	27.84

Table 13: External cohort (90,759 naturally distributed records), stages 4–6: micro accuracy, macro F1 (top-1 and top-3 conventions), and the same after a label-free EM estimate of the deployment prior from the unlabelled external posteriors (Saerens et al., 2002). "No class bias": the pool refitted with its per-class bias frozen at zero.

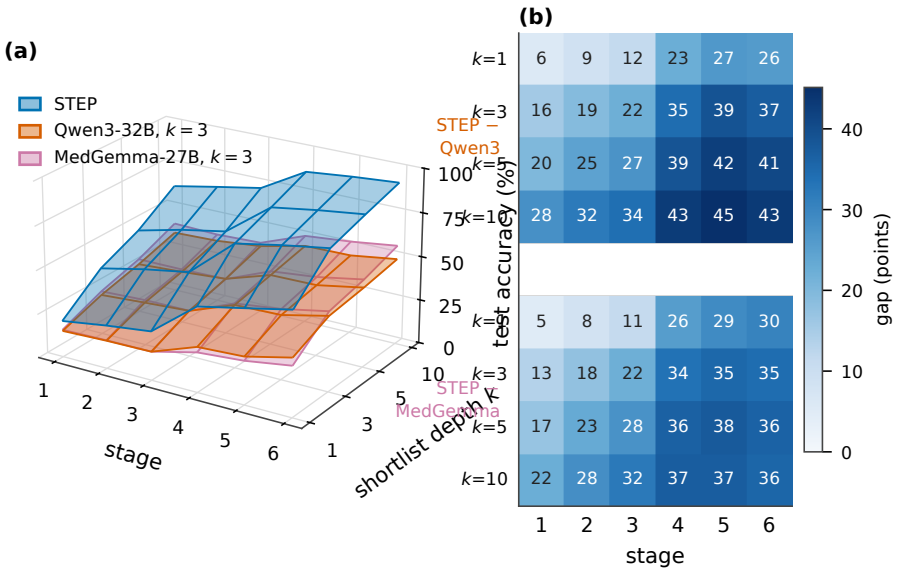

Figure 14: Accuracy over stage and shortlist depth on the same 8,198 balanced test records. (a) One surface per method: STEP, Qwen3-32B-AWQ (k=3, mean of three demonstration seeds) and MedGemma-27B-text (k=3, seed 0), both under the staged-evidence prompt. (b) The vertical distance between the surfaces: STEP minus each prompted arm at every (stage, k), in %; the values are those of Table 11, and no number should be read off the surface. Takeaway: STEP sits above both prompted arms at every stage and depth, and the gap widens with the evidence and with the depth.

```
You are an experienced EMS protocol classifier for a US NEMSIS 2024 patient care record.

A dispatcher or paramedic will act on a SHORTLIST, not on a single answer, so rank the
candidates: put the protocol you consider most likely first, then the next most likely,
and so on.

Information about a call arrives progressively. You are shown only what is documented by
the current point in the workflow. Fields that were not documented are shown explicitly
as [not recorded] or [not measured]; treat that as genuine absence of a measurement, and
do NOT assume a normal value in its place.

Answer with ONE compact JSON object and nothing else:
{"ranked":[id1,id2,id3,id4,id5,id6,id7,id8,id9,id10]}
- exactly 10 protocol IDs, most likely first
- IDs only, as integers, drawn from the candidate list below
- no duplicates, no names, no explanation, no markdown fence

HOW TO READ A RECORD
- Dispatch complaint: the call-taker's coded reason for the call.
- Complaint location / organ system: where the patient's complaint is, and which system.
- Symptoms and impressions: ICD-10-CM terms documented by the crew.
- Vitals are summarised over the call as mean+-sd; a single reading shows one number.
- GCS is 3-15 (15 = normal). AVPU is Alert / Verbal / Painful / Unresponsive.
- Pain is a 0-10 score. ETCO2 is end-tidal CO2 in mmHg. SpO2 is a percentage.

CANDIDATE PROTOCOLS (82 classes, id<TAB>name):
<82 lines, one per protocol, in one fixed ordering>
                              --- end of the cached prefix ---
<k demonstrations: a record rendered at the query's stage, then its gold ranked answer>
<the query record, rendered at the current stage>
```

Figure 15: The prompt template, reproduced from the builder. Everything above the marker is the fixed, cacheable prefix; the marker itself is not part of the prompt. Angle brackets stand for content that varies: the candidate block is 82 lines of `id<TAB>name`, and the demonstrations and query are records rendered by the same stage-aware code that produces the supervised arm's features. No rendered prompt is shown, because it would contain a patient record.

C STEP: Decomposition, Pooling Axes, Gate and Selection Stability

Implementation. The stage models and the experts follow Table 2. The pool's gate is $3n \to 32$ (tanh) $\to n$ with a zero-initialised output layer, base weights e_n and a zero per-class bias; fitting uses AdamW (learning rate 10^{-3}, weight decay 0, batch 512), at most 100 epochs with patience 15 on validation negative log-likelihood (improvement $> 10^{-5}$), and selects the best validation-NLL epoch with epoch 0 (the identity) a candidate. The gate's initialisation is seeded (seed 0) immediately before the pool is constructed. Inputs at stage t, in order: $\bar{p}_1, \ldots, \bar{p}_{t-1}$; the experts dispatch and age-only context at stages 1–3, plus location from stage 2 and organ system from stage 3; dispatch, location, organ system, symptoms and full context at stage 4, plus vitals from stage 5 and impressions at stage 6; then $\bar{p}_t$. Table 18 lists the expert datasets: eight exist for seven expert slots because the context expert has two mutually exclusive capture-time variants.

Cross-fitting and selection boundaries. The out-of-fold posteriors of every base model come from a five-fold split of all 73,780 training rows, so each inner-validation row is predicted by fold models that never trained on it, and the validation deltas of this appendix compare out-of-sample posteriors. The pool is fitted on the 66,402 fit rows and its epoch is selected on the 7,378 inner-validation rows. The fold models that produce the fit-row inputs did train on inner-validation rows, so validation labels reach the pool's inputs indirectly and the validation NLL is not an untouched estimate for the stacked system; the test rows are untouched by every step, and test inputs come from base models trained on the whole training split. The single-seed NN's validation posteriors are the seed-0 out-of-fold posteriors, so NN, ensemble and STEP are compared on the same footing on both splits.

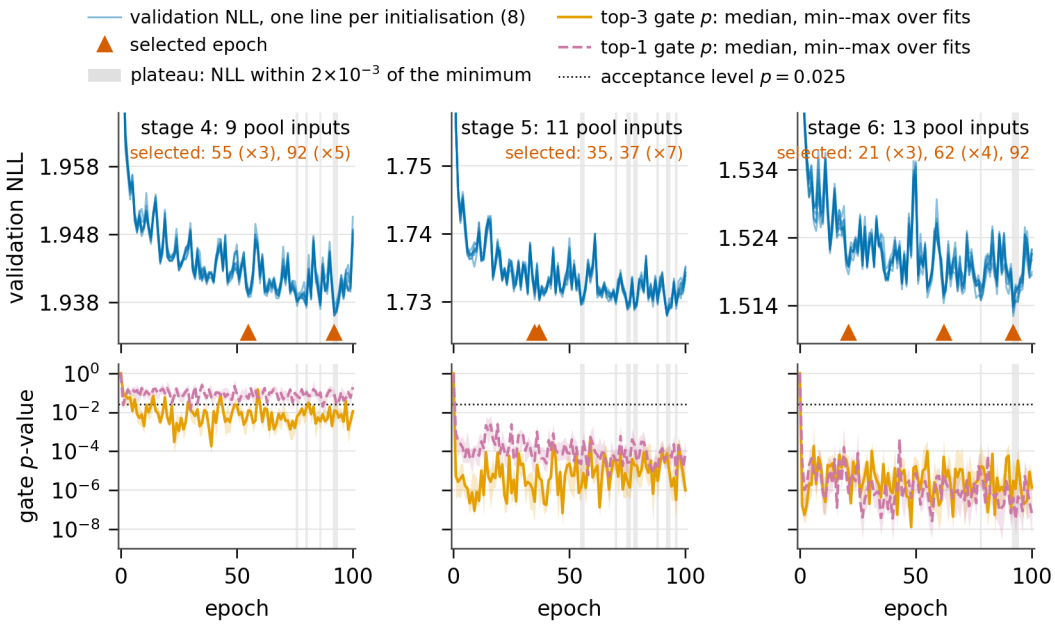

Figure 16: Epoch selection boundaries with validation NLL and gate p-value

Reproduction. The all-stage pools were refitted on CPU with deterministic algorithms from the saved logits. At every stage the refit selects the historically recorded epoch and reproduces the recorded test top-1/3/5/10 exactly; at stages 4–6 the test logits equal the previously saved pool logits (maximum absolute difference 0), so the stage-5 and stage-6 STEP numbers are those previously reported for the pooled pipeline.

Stage-wise results in detail. STEP ranks the recorded protocol first for 20.69% of test encounters from dispatch alone and for 60.09% once impressions are documented, and places it in the top three

		ensemble − NN				STEP − ensemble (TFA)				STEP − NN			
st.	ep.	Δ_1	p	Δ_3	p	Δ_1	p	Δ_3	p	Δ_1	p	Δ_3	p
test split, $n = 8{,}198$ (reported)													
1	30	+0.39	0.094	−0.05	0.82	+0.11	0.55	+0.24	0.17	+0.50	0.030	+0.20	0.36
2	78	+0.37	0.065	+0.24	0.29	+0.34	0.072	+0.20	0.34	+0.71	0.003	+0.44	0.093
3	37	+0.54	0.012	+0.71	0.002	−0.11	0.60	+0.22	0.31	+0.43	0.074	+0.93	5.0e-4
4	92	+2.33	1.2e-13	+2.81	8.5e-21	+0.29	0.30	+1.18	3.3e-6	+2.62	8.9e-14	+3.99	1.0e-32
5	35	+2.16	3.5e-11	+3.10	1.1e-24	+1.28	2.4e-6	+1.50	1.3e-10	+3.44	4.1e-22	+4.60	2.6e-46
6	92	+2.06	3.1e-10	+2.31	8.7e-16	+1.23	2.7e-5	+1.18	1.8e-6	+3.29	2.0e-19	+3.49	1.7e-27
inner-validation rows, $n = 7{,}378$ (out-of-fold base posteriors; the rows that select the pool's epoch)													
1	30	+0.24	0.27	+0.43	0.058	−0.01	1.00	+0.30	0.16	+0.23	0.36	+0.73	0.006
2	78	+0.01	1.00	+0.34	0.21	+0.27	0.20	+0.26	0.24	+0.28	0.29	+0.60	0.042
3	37	+0.88	6.8e-4	+1.42	4.5e-7	−0.14	0.52	−0.03	0.95	+0.75	0.007	+1.40	4.0e-6
4	92	+2.36	5.5e-12	+2.86	3.3e-16	+0.49	0.11	+0.75	0.010	+2.85	6.9e-14	+3.61	1.7e-21
5	35	+2.81	9.4e-15	+2.37	4.7e-13	+1.15	4.0e-5	+1.34	2.5e-7	+3.96	2.7e-25	+3.71	2.1e-26
6	92	+2.91	6.3e-17	+2.32	5.7e-14	+1.56	1.6e-7	+1.21	1.0e-5	+4.47	1.8e-30	+3.52	4.1e-25

Table 15: Decomposition of STEP's gain over the single-seed NN into seed ensembling and pooling, per stage, in percentage points (Δ_1 top-1, Δ_3 top-3) with exact paired McNemar p. ep.: the pool epoch selected by validation NLL (0 would return the ensemble unchanged; it was never selected). On test, 36 comparisons (3 contrasts × 6 stages × 2 metrics) form the family; the Bonferroni level is 0.0014. Under the historical acceptance test on the validation rows (App. E) the pool was rejected at stage 1, rejected at stage 2, rejected at stage 3, accepted per candidate but not admitted at stage 4, admitted at stage 5, admitted at stage 6; that test no longer decides where the pool is applied.

	NN (seed 0)				3-seed ensemble				STEP			
stage	top-1	top-3	top-5	top-10	top-1	top-3	top-5	top-10	top-1	top-3	top-5	top-10
1	20.19	38.36	48.15	62.71	20.58	38.31	48.27	62.65	20.69	38.56	48.04	63.26
2	24.97	44.69	55.44	69.58	25.34	44.94	55.86	70.10	25.68	45.13	55.82	70.15
3	28.73	49.41	60.12	73.33	29.26	50.12	61.04	73.82	29.15	50.34	61.28	74.14
4	46.97	67.71	76.12	85.78	49.30	70.52	78.57	87.63	49.60	71.70	79.24	88.19
5	51.07	71.48	79.81	87.98	53.23	74.58	81.98	90.05	54.51	76.08	83.28	90.56
6	56.79	76.79	83.62	90.51	58.86	79.09	85.58	91.83	60.09	80.28	86.62	92.67

stage	test top-1, seeds 0/1/2	spread	test top-3, seeds 0/1/2	spread	val top-1, seeds 0/1/2	spread
1	20.19 / 20.44 / 20.47	0.28	38.36 / 38.23 / 38.18	0.18	19.67 / 19.45 / 20.13	0.68
2	24.97 / 24.98 / 25.32	0.35	44.69 / 44.47 / 44.95	0.48	24.94 / 24.64 / 24.94	0.30
3	28.73 / 28.65 / 28.95	0.29	49.41 / 49.89 / 49.94	0.52	27.81 / 28.07 / 28.15	0.34
4	46.97 / 47.27 / 46.65	0.62	67.71 / 68.25 / 68.39	0.68	45.77 / 46.04 / 45.57	0.47
5	51.07 / 51.10 / 51.77	0.70	71.48 / 71.94 / 72.74	1.26	50.30 / 50.12 / 51.30	1.18
6	56.79 / 57.09 / 56.43	0.66	76.79 / 76.81 / 76.88	0.10	55.07 / 55.49 / 55.31	0.42

Table 16: Top: test accuracy at depths 1–10 of the single-seed NN, the three-seed ensemble and STEP. Bottom: every ensemble member (seeds 0–2) of each stage model, on test (full-train models) and on the inner-validation rows (out-of-fold), with the spread (max−min) in percentage points. Seed 0 is the NN baseline; it was fixed as the project's single-seed control before these comparisons, not chosen by its result.

for 38.56% and 80.28% (Table 1). Identifiability is not spread evenly over the record: the scene stage alone is worth 20.44 top-1 %, more than the other four transitions together, because the symptom codes and the four scene-time context fields arrive there at once; the remaining transitions add 3.48 to 5.57 % each (Fig. 6, right). Against the single-seed NN, STEP adds 2.62 to 3.44 top-1 and 3.49 to 4.60 top-3 % from the scene stage on ($p < 10^{-13}$ for all six comparisons). Before it the gains are at most 0.71 top-1 and 0.93 top-3 %, and only the stage-3 top-3 gain survives the correction for the 12 comparisons; the stage-1 top-3 gain of 0.20 % ($p = 0.36$) is a null effect. Over the ensemble, the pool adds 1.23 to 1.28 top-1 and 1.18 to 1.50 top-3 % at stages 5–6 and a top-3 gain of 1.18 % at stage 4 (top-1 $+0.29$, $p = 0.30$); before the scene stage its effect ranges from -0.11 to $+0.34$ % on test, and none of those six comparisons is significant even before correction. No pool loses a full point to the ensemble on either split or metric; the largest decrease is 0.14 validation top-1 % at stage 3. Validation never selected the identity epoch, so every pool departs from the ensemble; where the historical acceptance test did not admit it (stages 1–4), the departure is inert or, at stage 4, a top-3 gain.

pool inputs	Δtop-1	p	Δtop-3	p	Δtop-1	p	Δtop-3	p
stage 1: 3-seed ensemble 20.58 / 38.31 (test)								
modality axis (STEP; no history)	-0.01	1.00	$+0.30$	0.16	$+0.11$	0.55	$+0.24$	0.17
stage 2: 3-seed ensemble 25.34 / 44.94 (test)								
stage-history axis only	$+0.15$	0.35	$+0.05$	0.77	$+0.07$	0.67	$+0.04$	0.86
modality-expert axis only	$+0.27$	0.19	$+0.14$	0.55	$+0.24$	0.20	$+0.24$	0.23
both axes (STEP)	$+0.27$	0.20	$+0.26$	0.24	$+0.34$	0.072	$+0.20$	0.34
stage 3: 3-seed ensemble 29.26 / 50.12 (test)								
stage-history axis only	$+0.28$	0.046	$+0.05$	0.79	-0.09	0.58	$+0.02$	0.93
modality-expert axis only	-0.05	0.83	$+0.08$	0.75	$+0.05$	0.84	$+0.16$	0.46
both axes (STEP)	-0.14	0.52	-0.03	0.95	-0.11	0.60	$+0.22$	0.31
stage 4: 3-seed ensemble 49.30 / 70.52 (test)								
stage-history axis only	$+0.22$	0.38	$+0.33$	0.17	$+0.15$	0.50	$+0.24$	0.24
modality-expert axis only	$+0.20$	0.50	$+0.57$	0.042	$+0.22$	0.42	$+0.81$	0.001
both axes (STEP)	$+0.49$	0.11	$+0.75$	0.010	$+0.29$	0.30	$+1.18$	3.3×10^{-6}
stage 5: 3-seed ensemble 53.23 / 74.58 (test)								
stage-history axis only	$+0.42$	0.099	$+0.84$	2.4×10^{-4}	$+0.72$	0.004	$+0.79$	8.3×10^{-5}
modality-expert axis only	$+0.54$	0.044	$+1.03$	2.4×10^{-5}	$+0.26$	0.32	$+0.94$	1.6×10^{-5}
both axes (STEP)	$+1.15$	4.0×10^{-5}	$+1.34$	2.5×10^{-7}	$+1.28$	2.4×10^{-6}	$+1.50$	1.3×10^{-10}
stage 6: 3-seed ensemble 58.86 / 79.09 (test)								
stage-history axis only	$+0.62$	0.020	$+0.92$	8.2×10^{-5}	$+0.54$	0.031	$+0.78$	3.8×10^{-4}
modality-expert axis only	$+0.61$	0.027	$+0.73$	0.002	$+0.40$	0.15	$+0.63$	0.005
both axes (STEP)	$+1.56$	1.6×10^{-7}	$+1.21$	1.0×10^{-5}	$+1.23$	2.7×10^{-5}	$+1.18$	1.8×10^{-6}
validation								*test*

Table 17: What the pool adds over the three-seed ensemble at every stage, by axis, including the stages where it adds nothing or loses. Every pool uses the same recipe (App. C); the combined pool is the one STEP uses. Left: inner-validation rows (n = 7,378); right: test rows (n = 8,198); exact paired McNemar p, descriptive.

expert	columns	seed 0	seed 1	seed 2
dispatch	46	18.39	18.22	18.42
complaint location	10	5.79	5.73	5.71
complaint organ system	12	7.08	7.09	7.08
symptoms	1063	30.88	30.89	31.04
vitals	77	16.93	17.09	17.12
impressions	1246	40.12	40.63	39.97
context full	190	10.80	10.98	10.86
context age	10	3.62	3.46	3.77

Table 18: The eight block experts on the corrected family (one MLP per block, frozen recipe, trained on the fit rows), inner-validation top-1 per seed.

Decomposition. Table 15 splits STEP's gain over the NN into ensembling and pooling on both splits, and Table 16 gives every depth and every seed. On test the ensemble adds 2.06 to 2.33 top-1 % at stages 4–6 and 0.37 to 0.54 before; at stage 1 it does not move top-3 (-0.05). The pool adds 1.23 to 1.28 top-1 % at stages 5–6 and is within ±0.35 top-1 % of the ensemble at stages 1–4; at depth 5 it is 0.23 % below the ensemble at stage 1, a reported-only metric. Most of what the complete workflow recovers is therefore attributable to seed ensembling; we read the ensembling gain as variance reduction, consistent with the seed spread of Table 16 (0.28–0.70 test top-1 %), without having isolated that mechanism.

The two pooling axes. Table 17 fits the history-only and the modality-only pool with the same recipe at every stage where both are defined. At stages 5–6 each axis alone gains 0.26 to 0.72 test top-1 % and their combination beats each, as if the history said when evidence arrived and the experts which source it came from. At stages 2–3 no pool, single-axis or combined, moves top-1 or top-3 by more than 0.35 % on either split; at stage 4 the gains are mainly on top-3 (up to 1.18 test % for the combined pool) and top-1 moves by at most 0.49. Pooling the experts *without* the joint posterior was tested historically on validation with the same recipe and loses to the ensemble on both metrics at every stage from 2 on (top-1 -0.64, -1.84, -2.75, -3.56, -3.27 at stages 2–6), which is why the joint posterior is always an input.

Selection stability. Table 19 reports two historical experiments on the stage 4–6 pools. Across eight fits per stage on identical inputs and the seed-0 initialisation, spanning CUDA with and without deterministic algorithms and CPU, the validation NLL agrees to 10^{-7} and one epoch is selected per stage: given its seeds the fitting is deterministic. Across eight gate initialisations the selected epoch moves at every stage; at stages 5–6 every initialisation is accepted on both metrics with test top-1 inside 0.2 %, whereas at stage 4 all eight are accepted on top-3 alone with p_3 between 0.0065 and 0.024 and p_1 never below 0.041, and one of four unseeded refits was rejected. That initialisation sensitivity is why the historical pipeline did not admit the stage-4 pool. The revised pipeline uses the seed-0 initialisation at every stage, as before; across the eight initialisations the stage-4 pool's test top-1 lies between 49.59 and 49.84, so the reported stage-4 accuracy is not sensitive to it even though the historical verdict was.

st.	n	selected epoch(s)	accepts	val p_3 range	val p_1 range	test top-1	NLL spread
A. eight code paths, seed-0 gate initialisation (cuda non-deterministic ×3, cuda deterministic ×2, cpu ×3)							
4	9	92	8/8	0.010 (identical)	0.109 (identical)	49.60–49.60	1.2e-07
5	11	35	8/8	2.5×10^{-7} (identical)	4.0×10^{-5} (identical)	54.51–54.51	1.2e-07
6	13	92	8/8	1.0×10^{-5} (identical)	1.6×10^{-7} (identical)	60.09–60.09	2.4e-07
B. eight gate initialisation seeds (0–7), training seed 0, cuda							
4	9	55 (3×), 92 (5×)	8/8	0.007–0.024	0.041–0.129	49.59–49.84	3.5e-03
5	11	35 (1×), 37 (7×)	8/8	9.8×10^{-8}–1.0×10^{-5}	2.5×10^{-5}–1.7×10^{-4}	54.35–54.55	7.0e-04
6	13	21 (3×), 62 (4×), 92 (1×)	8/8	6.3×10^{-6}–3.9×10^{-4}	1.6×10^{-7}–1.2×10^{-5}	59.92–60.09	7.0e-03

Table 19: Selection stability of the combined pool. A: eight fits per stage on identical inputs and the seed-0 gate initialisation across floating-point code paths; the fitting is deterministic given its seeds (validation NLL agrees to 10^{-7}, one selected epoch, the recorded gate reproduced exactly). B: eight gate initialisations; the selected epoch moves at every stage, but at stages 5–6 every initialisation accepts on both metrics far below the threshold while at stage 4 all accepts rest on top-3 alone with p_3 within a factor of four of 0.025 (and one of four unseeded refits, not shown, rejected at $p = 0.059$). Full 100-epoch curves; selection rule (best validation NLL, improvement $> 10^{-5}$, patience 15, epoch 0 a candidate) simulated offline.

The pool as a logarithmic opinion pool. Equation 3 gives $\hat{p}_t(y \mid x) \propto \exp(b_y) \prod_s r_s(y \mid x)^{w_s(x)}$, a weighted logarithmic opinion pool with per-record weights. At $w = e_n$, $b = 0$ it equals $r_n = \bar{p}_t$ exactly, which is the identity initialisation; with equal fixed weights $1/S$ over S seeds it is the geometric-mean ensemble of Equation 2.

Conjunctions under fixed and record-dependent weights. Let u and v be two modalities (a dispatch code and a vital sign) and suppose the class posterior depends on their conjunction: $P(y \mid u, v)$ is high only when both take particular values and near the prior otherwise. With fixed class-tied weights a product of per-modality experts has the form $\log P_{\text{pool}}(y \mid u, v) = w_u \log q(y \mid u) + w_v \log q(y \mid v) + b_y + \text{const}$, which is *additive* in the two log-posteriors: for any u, u' and v, $\log P(y \mid u, v) - \log P(y \mid u', v)$ does not depend on v. A conjunction has exactly such a dependence, so a fixed-weight pool can only approximate it by sharpening whichever marginal expert carries the signal. With the gate, $w = w(c(u, v))$ depends on both experts' confidence summaries, so the pooled log-posterior is no longer additive and the argument above does not apply: the gate can make the weight on $q(\cdot \mid u)$ depend on how confident $q(\cdot \mid v)$ is, which is a restricted interaction. It cannot condition on which value v takes beyond what the entropy, maximum and margin of $q(\cdot \mid v)$ reveal, and its weights are shared by all 82 classes, so we expect it to recover little of a class-specific conjunction; we have not characterised its expressivity formally. The empirical evidence is the one that matters for the design: without the joint posterior the pool loses to the ensemble from stage 2 on, and with it the pool gains at stages 4–6 (Table 17).

A stronger tabular baseline: gradient-boosted trees. Does the increment depend on an under-competitive base predictor? Table 21 trains XGBoost (GPU histogram method) at every stage on the identical feature matrices, records and splits of the dataset with a small grid (depth 4, 6, 8 at learning rate 0.1 and depth 6 at 0.05; subsample 0.8, column subsample 0.5, 64 bins, seed 0), early stopping and selection on inner-validation log-loss, and the selected fit-rows model evaluated once on test. The boosted model beats the single-seed MLP by 1.4 to 2.2 top-1 % from the scene stage on

stage	main run: STEP − ens. (p)		static pool	four runs: mean [min, max]		cross-year: STEP − ens.	
	Δ_1	Δ_3	Δ_1 / Δ_3	Δ_1	Δ_3	24→23 Δ_1 / Δ_3	23→24 Δ_1 / Δ_3
1 dispatch	+0.30 (0.06)	+0.32 (0.13)	+0.37 / +0.29	+0.23 [+0.15, +0.30]	+0.21 [+0.11, +0.32]	+0.09 / +0.23	+0.14 / +0.22
2 location	+0.09 (0.59)	+0.12 (0.6)	+0.04 / +0.16	+0.05 [-0.09, +0.12]	+0.17 [+0.11, +0.27]	+0.12 / +0.20	+0.02 / +0.23
3 organ sys.	+0.29 (0.098)	+0.23 (0.3)	+0.16 / +0.27	+0.22 [+0.06, +0.35]	+0.31 [+0.15, +0.45]	−0.02 / −0.02	+0.17 / −0.02
4 scene	+0.09 (0.8)	+0.91 (6.2×10^{-4})	−0.05 / +0.76	+0.35 [+0.09, +0.71]	+0.92 [+0.72, +1.12]	+0.98 / +1.01	+0.32 / +0.98
5 vitals	+0.77 (0.0051)	+1.02 (3.7×10^{-5})	+0.37 / +0.82	+0.63 [+0.38, +0.83]	+1.09 [+0.81, +1.43]	+1.11 / +1.24	+1.00 / +1.26
6 impressions	+0.96 (7.7×10^{-4})	+1.27 (6.8×10^{-7})	+0.57 / +1.18	+1.06 [+0.96, +1.17]	+1.39 [+1.21, +1.68]	+1.15 / +1.36	+1.00 / +1.24

Table 20: Is the increment robust? Group 1: the main run (STEP minus the 3-seed ensemble, exact McNemar p; the corrected family). Group 2: the static log-linear pool (fixed weights, no gate) on the same inputs. Group 3: the whole pipeline repeated with four disjoint seed triples. Group 4: models trained on one year's balanced fit rows, applied once to the other year's balanced cohort (80,284 and 81,978 records; every stage-4–6 cross-year increment has $p < 10^{-3}$. Groups 2–4 are descriptive.

stage	config	GBM top-1 / top-3	GBM − NN	GBM − ens.	STEP − GBM Δ_1 [CI] (p)	Δ_3 (p)
1	d4, η=0.1, 169	20.16 / 38.29	−0.20 / +0.20	−0.26 (0.35) / +0.27 (0.32)	+0.56 [0.02, 1.10] (0.043)	+0.05 (0.89)
2	d4, η=0.1, 200	24.52 / 43.71	−0.02 / +0.33	−0.02 (0.96) / −0.09 (0.79)	+0.11 [-0.34, 0.57] (0.68)	+0.21 (0.47)
3	d4, η=0.1, 224	27.32 / 48.29	+0.01 / +0.50	−0.23 (0.42) / +0.13 (0.67)	+0.52 [-0.01, 1.07] (0.054)	+0.10 (0.76)
4	d6, η=0.05, 576	47.23 / 68.87	+1.39 / +1.55	−0.93 (0.014) / −0.73 (0.042)	+1.01 [0.29, 1.74] (0.007)	+1.65 (2.3e-06)
5	d6, η=0.05, 563	52.53 / 74.16	+2.18 / +2.04	−0.41 (0.32) / −0.37 (0.33)	+1.18 [0.38, 1.93] (0.003)	+1.39 (8.9e-05)
6	d6, η=0.05, 528	58.09 / 79.24	+2.06 / +3.26	−0.26 (0.54) / +0.68 (0.054)	+1.22 [0.46, 1.95] (0.0017)	+0.59 (0.083)

Table 21: A tuned gradient-boosted baseline (XGBoost, GPU histogram method) on same dataset family, identical feature matrices, records and splits: grid over depth 4/6/8 at learning rate 0.1 and depth 6 at 0.05, early stopping and selection on inner-validation log-loss, fit-rows model evaluated once on test. Paired against the single-seed MLP, the 3-seed MLP ensemble and STEP (all fit-rows, clean boundary); descriptive, 18 comparisons.

($p < 10^{-3}$) and matches the three-seed MLP ensemble (-0.93, -0.41, -0.26 top-1 at stages 4–6, only the first significant; top-3 -0.73, -0.37, $+0.68$), so the single network is an under-competitive comparator but the ensemble is not. STEP leads the boosted model by $+1.01$ / $+1.65$, $+1.18$ / $+1.39$ and $+1.22$ / $+0.59$ top-1 / top-3 at stages 4–6 ($p = 0.007$, 0.003, 0.0017 on top-1) and is level with it at stages 1–3.

stage	LOW (<0.35)			MEDIUM (0.35–0.60)			HIGH (≥ 0.60)		
	n	top-1	top-3	n	top-1	top-3	n	top-1	top-3
1	7,707	17.93	36.32	482	64.73	74.27	9	(22.22)§	(44.44)§
2	7,707	23.04	43.10	393	61.83	73.54	98	(87.76)§	(90.82)§
3	7,706	26.65	48.43	392	63.27	77.55	100	88.00	91.00
4	6,015	43.44	67.41	1,883	64.15	82.16	300	81.67	92.00
5	4,059	46.59	70.51	3,751	60.30	80.57	388	81.44	90.98
6	2,086	45.25	70.81	5,250	62.70	82.44	862	80.05	90.02

Table 22: DPCS tier distribution and observed reliability of the displayed STEP top-3 list, by stage. $a_{j,t}$ and $b_{i,t}$ are estimated from STEP's predictions on the 7,378 inner-validation rows (unsmoothed empirical rates); DPCS is then computed for every test record (K=3) and bucketed at the fixed, unfitted cut points 0.35/0.60. Counts are shown with every accuracy: buckets with $n < 100$ are parenthesised and marked § and are never cited in the text. At stages 1–3 94% of records are LOW and at most 100 are HIGH; from stage 4 every tier holds at least 300 records.

D DPCS: All Six Stages and Class-Level Trajectories

Table 22 gives the tier distribution and the observed reliability of the displayed STEP top-3 list at every stage with counts; because the rates are estimated from the pipeline's own validation predictions, applying the pool at stages 1–4 changed their tier tables, while stages 5–6 are unchanged (the pools there are bit-identical to the previously reported ones); its share columns are Figure 2(b). The rates $a_{j,t}$ and $b_{i,t}$ are unsmoothed: the median inner-validation support per dispatch code is 83.5 records, and 2.2% of test records carry a dispatch code with fewer than 30 inner-validation records. Figure 18 shows two class-level rank trajectories under STEP: for each protocol, the median and interquartile range of the true protocol's rank across that class's test records at each stage, annotated with the modal DPCS tier and its share. Protocol A converges late (its median rank enters the shortlist only at the scene stage) and its modal tier rises from LOW to MEDIUM one stage later; Protocol B is ranked first from dispatch on and its tier rises at the scene stage, when the dispatch-level reliability $b_{I_e,t}$ and the recognisability $a_{j,t}$ both improve. No individual record is drawn.

Tier reliability by stage. Figure 17 gives observed top-1 and top-3 by tier at the three stages where all three tiers are populated. Before the scene stage 94% of records are LOW and the HIGH tier holds at most 100 test records, which is what fixed cut % should produce when dispatch evidence alone cannot separate 82 protocols; MEDIUM already orders above LOW there. From the scene stage on every tier holds at least 300 records and the tiers order monotonically on both metrics: at stage 6 top-1 accuracy is 45.25% in LOW, 62.70% in MEDIUM and 80.05% in HIGH, and the mix shifts towards HIGH as the record fills (Fig. 2b).

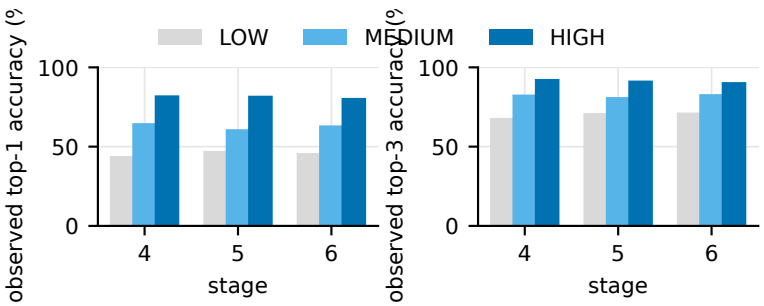

Figure 17: Observed top-1 (left) and top-3 (right) accuracy of the displayed STEP shortlist within each DPCS tier at the three stages where all three tiers are populated (tier shares: Fig. 2b). Every bucket drawn holds at least 300 test records; the rates the score multiplies are estimated on the inner-validation rows.

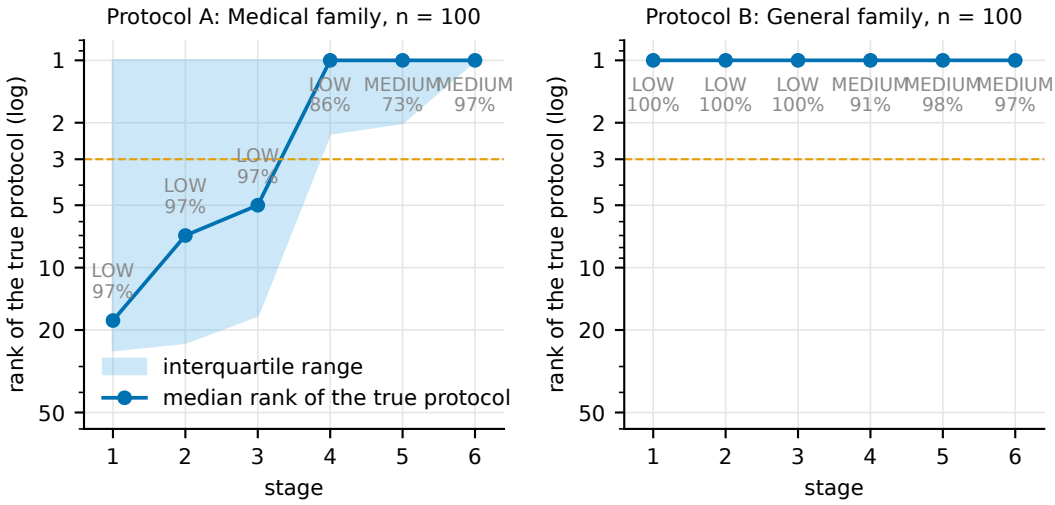

Figure 18: Class-level rank trajectories with DPCS tiers for a late-converging protocol (A, Medical family) and an early-converging one (B, General family); the dashed line is the top-3 boundary and the annotation is the modal tier with its share. Aggregates over each class's test records; no individual record is drawn.

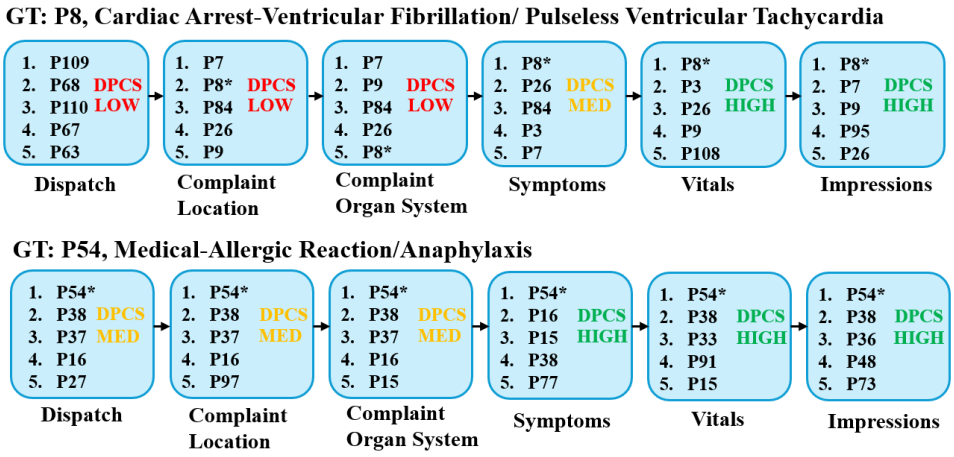

Figure 19: System UI example for the DPCS workflow during deployment

E STATISTICAL PROTOCOL AND MULTIPLICITY

The historical acceptance test. Before this revision a component entered the pipeline only when a paired validation test accepted it: a candidate was accepted when either metric improved at $p < 0.025$ and neither was harmed at $p < 0.05$ (the accept level is the Bonferroni split of 0.05 across the two metrics; the harm level stays at 0.05 because making harm harder to detect would be the wrong trade for a safety-critical display). A candidate declared in the script before the test ran faced that bar alone; a post-hoc finding also had to clear a family-wise correction over its wave. The rule still documents how the feature set, the pool's form and SEP's configuration were chosen, but it no longer decides the stages at which STEP pools. For control and candidate hit vectors $h^c, h^d \in \{0,1\}^n$ on the same records, let $b = \#\{h^c = 1, h^d = 0\}$ and $c = \#\{h^c = 0, h^d = 1\}$. The p-value is the two-sided exact binomial probability of c successes in $b + c$ trials at rate $\frac{1}{2}$ (1 if $b + c = 0$). It is computed twice, with the hit defined by h_1 and by h_3, on the same records; the top-3 hit set contains the top-1 hit set, so the two tests are strongly positively dependent and the Bonferroni split of the accept level across them is conservative rather than exact. The same exact test, without the accept/harm rule, produces every p-value of the STEP revision.

Multiplicity. The policy was fixed in advance rather than chosen per result. A pre-registered candidate faces the per-candidate bar alone. A post-hoc finding must additionally clear a family-wise correction over every tested pair in its wave; for the modality-fusion wave, whose script declares its own exposure and its post-hoc third variant, that is $0.05/72 = 6.9\times10^{-4}$. A finding that clears the per-candidate bar but not the family-wise one is not adopted, but can be re-tested later as a pre-registered hypothesis of its own. Which bar applies is declared when the comparison is set up. Every delta and every p-value carries its split: selection quantities are validation, reported quantities are test, and neither is substituted for the other. Table 23 summarises every tested comparison in the project with a machine-readable record, 238 in total, of which 84 were accepted, by wave; the count is what a reader should weigh the accepted results against. Table 24 lists the design decisions the tests settled.

Historical stage selection and the revised policy. Under the historical rule the combined pool was rejected at stages 1–3 on validation, accepted at stage 4 on top-3 alone ($+0.75$, $p = 0.0095$) but not admitted because it failed the family-wise bar of its wave (6.9×10^{-4}) and its verdict moved with the gate's initialisation, and admitted at stages 5–6 (Table 15, last column). The revised pipeline applies the pool at all six stages regardless; these verdicts are supplementary statistical evidence, not eligibility requirements. The revision adds two families, both on test and corrected separately: the 12 comparisons that decompose STEP (ensemble − NN, STEP − ensemble and STEP − NN at six stages on two metrics; Bonferroni level $0.05/12 = 1.4\times10^{-3}$), of which 7 of the 12 STEP − NN comparisons survive (both metrics at stages 4–6 and top-3 at stage 3) and 5 of the 12 STEP − ensemble comparisons (both metrics at stages 5–6 and top-3 at stage 4); and the 48 prompted-run comparisons against STEP ($0.05/48$), all of which survive. The validation deltas in the same tables are reported, not tested against a threshold, and the axis ablations of Table 17 are descriptive. These families are not part of the 238-comparison ledger above.

wave	split	pairs	accepted	rejected	entered the historical system
Task 2: modality fusion	inner-val	72	25	47	yes: pool at stages 5–6 (now all stages)
Temporal pool data	inner-val	14	9	5	no
Temporal pool, config D, stages 5–6	inner-val	14	8	6	no (stage axis alone)
Temporal pool, config D, stages 1–4	inner-val	28	3	25	no
Task 1: temporal pool with context, stages 2–4	inner-val	15	3	12	no
Task 1: temporal pool with context, stages 5–6	inner-val	14	10	4	no (stage axis alone)
Context, uniform placement	inner-val (superseded)	6	6	0	no
Task 0c: context at audited capture time	inner-val	6	6	0	yes: context block
Config-D ladder, acceptance-test re-check	inner-val	6	6	0	yes: feature set
Task 1: vitals sequence with context	inner-val	2	1	1	no
Architectures on config D, held-out split	held-out partition val	24	1	23	no (baselines)
LLM arm vs NN 3-seed ensemble, test8198	test8198	24	0	24	no (comparison)
LLM within-arm, zero-shot → k=3, test8198	test8198	6	5	1	no (comparison)
LLM k selection on val820, pooled $n = 14{,}760$	val820 (pooled)	2	1	1	yes: k=3 for the LLM arm
Task 1: adopt-vs-production	inner-val	4	0	4	no
Provenance split	inner-val	1	0	1	no
all waves		**238**	**84**	**154**	

Table 23: Every acceptance-test comparison in the project with a machine-readable record, by wave: 238 paired comparisons, 84 accepted. The split is the one the test was run on; the last column says whether the wave's accepted candidates entered the pipeline before this revision; the revised STEP pipeline pools at every stage regardless of these verdicts, and its own test comparisons are counted separately (App. E). The full per-comparison ledger (deltas and p-values for every pair) is shipped with the code.

decision	alternative(s)	effect (p)	verdict
gated log-linear pool	ungated pool; residual MLP	gated $+0.62$ / $+0.92$ (0.020 / $8.2{\times}10^{-5}$); residual -0.19 / -0.04	gated selected (top-3 leg); residual rejected 20/20 cells
joint posterior last (identity init)	experts only	-3.27 / -2.07 (1.0×10^{-13} / 5.1×10^{-8})	experts-only rejected (harm); experts added to joint selected
3-seed geometric mean	single seed	val $+2.91$ / $+2.32$; test $+2.06$ / $+2.31$ (3.1×10^{-10} / 8.7×10^{-16})	selected
combined pool at stage 4	3-seed ensemble	$+0.49$ / $+0.75$ (0.109 / 0.010)	historically not admitted (family-wise bar); applied under the revised policy (App. C)
last-epoch checkpoint	best-validation epoch	top-1 -0.26 / -0.53 (stages 5 / 6; 0.445 / 0.120)	rejected
MLP	LSTM; transformer; two cascades (stage 6, held-out split)	-1.35 / -0.79; -1.49 / -1.14; -3.21; -2.91 (all $p < 0.01$)	all rejected with harm at stages 5–6
merged clinical vocabulary	provenance split (3,801 features)	-1.06 / -0.58 (0.005 / 0.090)	rejected (harmed top-1)
config-D features: explicit missingness; extended vitals; timed context	plain `zero_mask`; without; none	stage 5: $+2.52$ / $+1.83$; $+5.49$ / $+3.77$; stage 4 context $+5.76$ / $+4.73$ (all $p < 10^{-4}$)	selected (stage-6 encoding step on top-3 only)
LLM k=3	zero-shot; k=5 (val820)	k=3 $+2.20$ / $+1.80$; k=5 $+1.76$ / -1.21 (9.7×10^{-5})	k=5 rejected: harmed top-3
acceptance test, $\alpha_{\text{accept}} = 0.025$, $\alpha_{\text{harm}} = 0.05$	single top-1 test	reversed four decisions	design rule (harm leg is a safety screen)

Table 24: Why each design choice was made (historical acceptance-test decisions; STEP now pools at every stage, App. E). Effects are Δtop-1 / Δtop-3 in pp with the exact paired McNemar p; inner-validation acceptance test unless a split is stated. Where the driver is top-3 rather than top-1, the verdict says so.

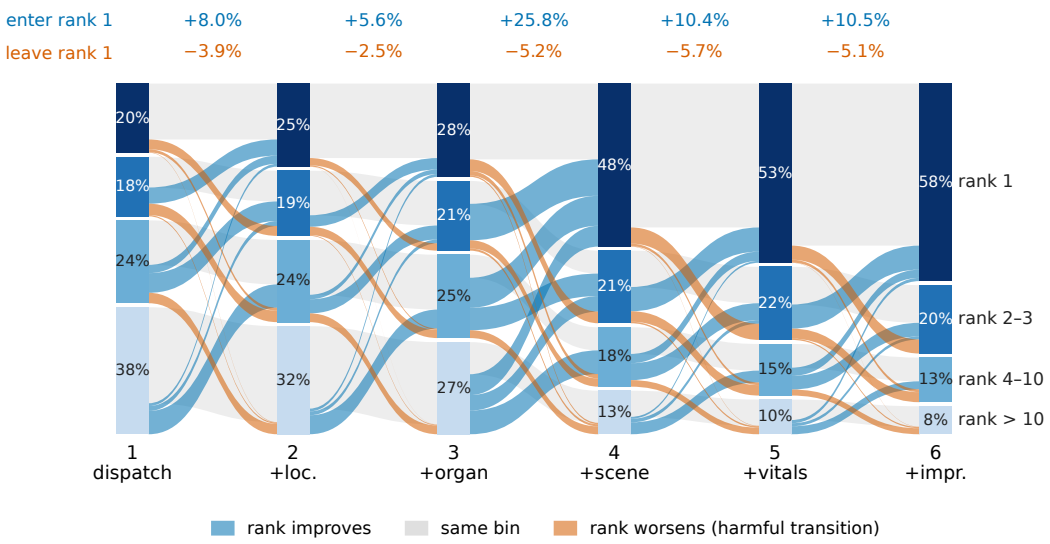

Figure 20: What the earlier posteriors carry. Rank of the true protocol under the three-seed ensemble (the pool's joint input) on the 8,198 balanced test records, binned, with the flows between consecutive stages: blue flows improve the bin, vermillion flows worsen it (a harmful transition), grey flows keep it; the printed shares are the records entering and leaving rank 1 at each transition (clean boundary; aggregates in the repository's `figures_3d.json`). Takeaway: at every transition a few per cent of records leave rank 1 while many more enter it, and most of those leaving stay in the top three, which is what a pool over the stage posteriors can recover and a hard reversion rule cannot separate from legitimate corrections.

F What Earlier Posteriors Carry: Diagnosis, a Training-Time Alternative and a Synthetic Study

All measurements of this appendix are on the feature dataset under the clean boundary; the diagnosis uses development data only (out-of-fold posteriors of the 66,402 fit rows and the fit-rows models' predictions on the 7,378 inner-validation rows), and test is read only for the final, frozen comparisons of Tables F and 30.

Transitions. At every transition the mean rises while a minority of records fall: 2.7% to 5.8% of the development records are ranked correctly by the earlier ensemble and incorrectly by the later one (transitions $1 \to 2$ to $5 \to 6$: 4.3, 2.7, 4.8, 5.8, 5.3%), against 5.5% to 25.5% that move the other way, and the true-class probability decreases on 31% to 43% of records. For 71% to 82% of the harmful flips the earlier top-1 is still inside the later top-3, which is why the top-3 increment of the pool is the more robust one. The stage models are strongly dependent (hit correlation between adjacent stages 0.42 to 0.78), yet 10.7% and 11.7% of the stage-5 and stage-6 errors are records some earlier stage ranked correctly and 13.3% and 13.7% are records some block expert ranked correctly (oracle upper bounds, never used for selection). Missingness of the new block is not the driver: at $5 \to 6$ the harm rate on records with an empty impression block (5.2%) equals the overall rate (5.3%).

Does the earlier posterior add information given the current one? Fitting on the out-of-fold posteriors and reading validation, a per-class scale and bias on the current ensemble alone lowers NLL at stage 6 from 1.747 to 1.630 and leaves top-1 unchanged (57.85 to 57.87); adding the previous stage's posterior as a second input of a static pool gives 58.02 / 1.603, all earlier stages 58.25 / 1.601, all earlier stages with per-class weights 58.54 / 1.589, the block experts 58.40 / 1.592, and both axes 58.74 / 1.573 (top-3 77.05 to 78.37). The same ordering holds at stages 4–5 and nothing beyond calibration appears at stages 2–3. About half of the NLL reduction is recalibration of the

current posterior; the accuracy gain and the other half are information the earlier and block-specific posteriors carry given the current one.

Hard retention cannot separate harm from correction. A logistic detector on inference-time quantities only (both stages' entropy, margin and maximum posterior, agreement, the later probability of the earlier top-1 and the reverse, the KL divergence between the two posteriors, and the number of non-zero entries in the new block) recognises a harmful flip against all other records with AUROC 0.86 to 0.93 on validation (maximum posterior alone: 0.47 to 0.74), but a harmful flip and a legitimate correction look alike, both being disagreements with a confident later model: restricted to disagreements the detector separates them with AUROC 0.66 to 0.80 (the confidence difference alone: 0.60 to 0.76). Reverting to the earlier prediction where the detector fires therefore blocks about as many corrections as it rescues: at the detector's own threshold the net top-1 change is $+0.01$, $+0.11$, -0.34, $+0.70$ and $+0.20$ % at transitions $1 \to 2$ to $5 \to 6$; at the 5% of validation records with the highest score -0.47, $+0.04$, -0.73, $+0.46$ and -0.22; at 10% -0.56 to -1.19; against oracle gains of $+2.7$ to $+5.8$. The soft pool, which reweights rather than switches, obtains $+0.8$ to $+1.0$ top-1 at stages 5–6 (Table 1); a hard, agreement-conditioned rule of the kind used for early exits (Kaya et al., 2019) is the wrong instrument here.

A training-time alternative: stage-congruent training. Because every harmful transition is a record the earlier network ranked correctly, a training-time analogue of the pool is to ask the stage-t network not to depart from the stage-$(t-1)$ posterior where that posterior was right, while leaving it free to revise where it was wrong. We add to the frozen recipe's cross-entropy a focal congruence term, $\lambda \mathbb{1}[\arg\max q = y]\,\mathrm{KL}(q \,\|\, p_\theta)$, where q is the out-of-fold three-seed posterior of stage $t-1$ on the same fit row (so no row is taught by a model that trained on it) and p_θ the stage-t network's posterior; this is the focal distillation of Yan et al. (2021) for regression-free model updates, transferred to a teacher that holds strictly less information than the student. $\lambda \in \{0.3, 1, 3\}$ and the plain-distillation variant (all rows weighted) were compared on validation with seed 0; $\lambda = 3$ with the focal weight was selected on validation top-1 at stages 4–6. Table F gives the three-seed result against the plain ensemble, against a self-distillation control that uses the stage's *own* out-of-fold ensemble as teacher (which isolates ensemble distillation from the earlier-stage information), and with the pool refitted on the congruent networks. The result is negative for accuracy and positive only for likelihood. The congruent single network gains 1.0 to 1.3 validation top-1 % at stages 4–6 over the plain one, but the congruent three-seed ensemble is -0.57, -0.45 and -0.20 top-1 % from the plain ensemble at stages 4–6 (test; $p = 0.10, 0.19, 0.58$; top-3 -0.33, -0.65, -0.06), the self-distillation control lands at the same place ($+0.16$, -0.41, $+0.09$; congruent minus self-distilled -0.73, -0.04, -0.28), and the pool refitted on the congruent networks is no better than the pool on the plain ones (-0.06, -0.83, -0.33 top-1). The congruence term does what it is designed to do at the margin, lowering the harmful-flip rate from 5.1–5.7% to 4.2–4.6%, but it lowers the correction rate by as much (25.8 to 24.2%, 10.4 to 9.0%, 10.5 to 9.7%), and it lowers test NLL by 0.03 to 0.19 nats (stage 6: 1.696 to 1.577), which is the smoothing any distillation from a three-seed teacher provides. The single-model gain is therefore an ensembling-like effect that the seed ensemble already captures, and the earlier-stage teacher adds nothing beyond a same-stage teacher: what the pool recovers is not recoverable by regularising the later network toward the earlier one.

stage	plain ens.	congruent ens.	Δ_1 [CI] (p)	Δ_3 (p)	ΔNLL	self-distilled ens. (Δ_1, p)	STEP on congruent (Δ_1 vs STEP on plain, p)
2	24.54 / 43.79	24.37 / 43.61	-0.17 [-0.51, 0.17] (0.38)	-0.18 (0.48)	-0.001	–	24.53 / 43.94 (-0.10, 0.59)
3	27.56 / 48.16	27.57 / 48.50	$+0.01$ [-0.35, 0.39] (1)	$+0.34$ (0.19)	-0.017	–	27.82 / 48.34 (-0.02, 0.95)
4	48.16 / 69.60	47.58 / 69.27	-0.57 [-1.22, 0.12] (0.1)	-0.33 (0.35)	-0.185	48.32 ($+0.16$, 0.64)	48.18 / 70.11 (-0.06, 0.85)
5	52.94 / 74.53	52.49 / 73.88	-0.45 [-1.13, 0.24] (0.19)	-0.65 (0.045)	-0.032	52.53 (-0.41, 0.2)	52.88 / 74.95 (-0.83, 0.0029)
6	58.34 / 78.56	58.15 / 78.49	-0.20 [-0.83, 0.48] (0.58)	-0.06 (0.87)	-0.119	58.43 ($+0.09$, 0.82)	58.98 / 79.34 (-0.33, 0.25)

Stage-congruent training on the corrected family (clean boundary, test split, 3-seed ensembles). Congruent: the stage MLP trained with the frozen recipe plus a focal congruence term ($\lambda = 3$) toward the stage-$(t-1)$ out-of-fold ensemble, weighted on records the earlier stage ranks correctly; self-distilled: the same term toward the stage's own out-of-fold ensemble (the ensemble-distillation control); STEP on congruent: the pool refitted with the congruent joint model. Paired McNemar p and bootstrap CIs on the 8,198 test records; descriptive.

The conjunction test. App. C shows that a pool with class-tied weights is additive in the experts' log-odds, so a non-additive interaction between blocks is not expressible within that class of pool, which is one reason the joint posterior is always an input. Whether the jointly trained model in fact recovers records whose block signals agree but whose decision-level combination fails is an empirical question. Let E be the test records an experts-only static pool (the block experts available

method	$3 \to 4$ harm / corr.	$4 \to 5$ harm / corr.	$5 \to 6$ harm / corr.	top-1 at 4 / 5 / 6	mean harm 4–6	mean corr. 4–6
plain ens.	5.17 / 25.77	5.66 / 10.44	5.07 / 10.48	48.16 / 52.94 / 58.34	5.30	15.56
PA product	2.61 / 20.64	2.20 / 8.16	1.63 / 7.37	45.11 / 51.07 / 56.81	2.15	12.06
static pool	4.33 / 24.73	4.05 / 9.25	3.23 / 8.84	48.11 / 53.31 / 58.92	3.87	14.27
STEP	4.51 / 24.91	3.72 / 9.19	2.96 / 8.56	48.24 / 53.71 / 59.31	3.73	14.22
congruent ens.	4.09 / 24.10	4.21 / 9.11	4.21 / 9.87	47.58 / 52.49 / 58.15	4.17	14.36
masked ens. (one model)	4.70 / 25.09	4.24 / 8.75	2.94 / 9.33	47.93 / 52.43 / 58.82	3.96	14.39
STEP on masked	4.31 / 25.18	3.55 / 8.76	2.89 / 8.64	48.45 / 53.66 / 59.40	3.58	14.19

Table 25: Harmful transitions (record ranked correctly at stage $t-1$ and incorrectly at t) and corrections (the reverse), in % of the 8,198 test records, for each method's own stage sequence under the clean boundary. Every method except the last two is a per-stage 3-seed ensemble with or without a pool; the masked ensemble is one network per seed trained over all six reveal stages with the unrevealed blocks masked, serving every stage.

at the stage, no joint posterior, fitted and selected as every other pool) ranks incorrectly, $S \subset E$ those whose true protocol is in the top-3 of at least two experts (the individual signals are present, the combination is missed) and $N \subset E$ those no expert has in its top-3. At stages 4–6 the jointly trained MLP ensemble recovers 29–33% of S against 13–15% of $E \setminus S$ and 6–7% of N; STEP recovers 24–31% of S. Two decision-level combiners that are more expressive than the class-tied pool, the gated pool and a per-class-weighted pool over the same expert posteriors, recover only 7–12% of S. Recovery is concentrated where the signature predicts it, and the two more expressive decision-level combiners we fitted do not obtain it. This is an empirical contrast between the jointly trained input and three decision-level rules on the same posteriors; it does not show that no decision-level rule could recover these records, and the additivity argument of App. C applies to class-tied pools only.

One block-masked network for all stages. If the pooling gain is finite-sample shrinkage toward lower-variance views, the same effect should be obtainable at training time by making one network predict well at every reveal stage: each fit row enters training in its six stage views (the stage-t matrix scattered into the stage-6 column layout, absent columns zero, plus a six-way stage indicator), so the network cannot over-rely on a block that may be absent. This is nested-block masking on the information axis; one network then serves all six stages with no pool and no extra inference cost. (Zeroing the later blocks on the stage-6 matrix is not equivalent: the 1,063 shared clinical columns carry merged symptom and impression counts at stage 6, and a first pilot that used that shortcut scored six % too high at stage 4; every stage view is taken from the stage-t matrix itself.) Preregistered hypotheses and results (Tables 25 and 1; three seeds, frozen recipe): (H1) the masked single network beats the per-stage single network at stages 4–6 by $+0.61$, $+0.29$ and $+0.81$ test top-1 % (only the last near significance); (H2) the masked three-seed ensemble is level with the per-stage ensemble (-0.23, -0.51, $+0.48$; all $p > 0.15$), so masking does not add accuracy beyond seed ensembling; (H3) the pool refitted on the masked networks adds $+0.52$, $+1.23$ and $+0.59$ top-1 over the masked ensemble at stages 4–6 and lands where STEP on the per-stage networks lands (48.45 / 53.66 / 59.40 against 48.24 / 53.71 / 59.31), so masking and pooling reach the same accuracy and do not add up; at stage 6 the pool's increment is smaller on masked bases ($+0.59$ against $+0.96$), consistent with a shared mechanism, at stage 5 it is not. What masking changes is the *consistency* of the stage sequence: harmful transitions fall from 5.17 / 5.66 / 5.07% to 4.70 / 4.24 / 2.94% at $3 \to 4$, $4 \to 5$ and $5 \to 6$ while corrections fall much less (25.8 / 10.4 / 10.5% to 25.1 / 8.8 / 9.3%), so the same accuracy is reached with about 40% fewer regressions at the two richest stages. Table 25 places every method on this harm–correction frontier: the fixed-weight product removes the most harm (2.2% mean at stages 4–6) but also a fifth of the corrections and 1.5 to 3 accuracy %; STEP lowers harm to 3.7% at the highest accuracy; the masked network reaches 4.0% with no pool; the congruence term sits between. Reducing harmful transitions is therefore a by-product of any shrinkage toward the earlier views, and the frontier, not the harm rate alone, is the quantity to report.

External cohort. Table 27 (§5.3) summarises this cohort at stages 5–6; the full stage range is Table F. The one-cohort limitation of earlier versions was addressed with a second NEMSIS 2024 extract that was never used for training or selection: 91,354 records drawn without protocol balancing, of which 90,759 fall in the 82 training protocols and share no PcrKey with the balanced

cohort. It is naturally distributed (protocol 36 alone is 53.8% of it, and 79 of the 82 protocols occur in it), its stage matrices were produced by a joint build with the balanced cohort followed by selection of the training family's columns by name, which reproduces the balanced rows byte for byte. Every model of Table 1 was retrained on the balanced fit rows only (bit-exact reproduction of the test numbers) and applied to it once (Table F). Three readings. Raw micro top-1 is low for every method (the balanced training prior mismatches the deployment prior: 24.42% for the ensemble at stage 6) and there STEP is 0.6 to 1.2 % *below* the ensemble at stages 4–6 ($p < 10^{-6}$), because the pool's per-class bias, fitted on balanced rows, moves mass away from the majority protocol. At the protocol level the increment transfers: macro top-1 rises by $+1.2$, $+1.3$ and $+1.9$ % at stages 4–6 (49.51 to 50.67, 55.36 to 56.62, 58.18 to 60.05). And once the label prior is accounted for, with $\log \pi_{\text{ext}}/\pi_{\text{train}}$ added to every method's log-posterior (π_{ext} estimated on one random half of the external cohort and scored on the other), STEP reaches 60.28 / 61.98 / 63.45 micro top-1 and 86.98 / 88.29 / 89.45 top-3 against 55.23 / 60.16 / 60.81 and 84.73 / 87.36 / 88.14 for the ensemble. The pooling gain therefore generalises to an independent, differently distributed cohort, and it does so only after the deployment prior is supplied: uncorrected, the class bias is a liability under prior shift, which is a deployment requirement rather than a property of the method. The one-network masked model transfers like the ensemble (macro top-1 within 0.7 % at stages 4–5, $+1.8$ at stage 6) and again halves the harmful-transition rate at $5 \to 6$ (3.4% to 2.1%).

stage	ens. micro top-1 / top-3	STEP micro top-1 / top-3	Δ_1 [95% CI] (p)	ens. macro top-1	STEP macro top-1	ens. prior-corr. top-1 / top-3	STEP prior-corr. to
1	9.30 / 15.51	8.91 / 15.27	−0.40 [-0.46, -0.33] (5.8e-38)	21.10	21.15	–	
2	9.58 / 17.17	9.07 / 16.88	−0.50 [-0.57, -0.43] (5.5e-45)	24.54	24.48	–	
3	10.06 / 19.87	9.77 / 19.12	−0.30 [-0.36, -0.23] (3e-18)	28.01	27.46	–	
4	19.49 / 37.18	18.31 / 33.49	−1.18 [-1.30, -1.04] (1.8e-63)	49.51	50.67	55.23 / 84.73	60
5	20.92 / 39.40	20.23 / 36.52	−0.68 [-0.82, -0.55] (1.6e-24)	55.36	56.62	60.16 / 87.36	61
6	24.42 / 47.52	23.24 / 41.92	−1.19 [-1.32, -1.04] (9.2e-57)	58.18	60.05	60.81 / 88.14	63

External evaluation on the naturally distributed NEMSIS 2024 cohort (90,759 records after removing the 82-class restriction's 80 records and the 515 records shared with the balanced cohort; protocol 36 alone is 54% of it), with the dataset family models trained on the balanced fit rows only and the pools fitted and selected exactly as in Table 1; no external record was used for training or selection. Micro top-1 / top-3 (dominated by the majority protocol), macro (per-protocol mean) top-1, and micro accuracy after a label-prior correction ($\log \pi_{\text{ext}}/\pi_{\text{train}}$ added to the log-posterior, with π_{ext} estimated on one random half of the external cohort and scored on the other, both folds pooled). Δ: STEP minus the 3-seed ensemble on micro top-1, paired, with percentile-bootstrap 95% CIs and exact McNemar p.

Year hold-out. The strongest generalisation test available in the public release is a different year. A NEMSIS 2023 cohort was built with the 2024 rules (single-protocol records; the protocol label is the index of the eProtocols.01 code in the protocol list, verified to agree with the 2024 mapping on 1.55 million single-protocol records; up to 1,000 records per protocol, 80,284 in all; delimiter-safe extraction; the unchanged vitals and context code) together with a naturally distributed 2023 sample of 100,000 single-protocol records, and every 2024 model was applied to both (Table F). A first pass had joined no vital-sign readings because of a field-separator escape in the extraction; it was caught by the vitals coverage check (0% against 86.7% after the fix, 99.4% of records with readings) and rerun; the numbers reported are from the corrected build. Across the year the balanced-cohort ensemble drifts by -1.7, -2.3, -3.0 top-1 % at stages 1–3 and -7.1, -6.9, -5.7 at stages 4–6, i.e. the drift sits in the symptom, vital-sign and impression blocks, and the pooling increment transfers at full size: $+0.98$ / $+1.01$, $+1.11$ / $+1.24$ and $+1.15$ / $+1.36$ top-1 / top-3 at stages 4–6 ($n = 80{,}284$, every $p < 10^{-26}$), with the static pool at $+0.68$ to $+0.88$ top-1 and the fixed product again harmful (-0.9 to -2.2). On the natural 2023 sample the picture of App. F's external cohort repeats: the class bias fitted on balanced rows works against the deployment prior, so the pool with bias loses micro top-1 and, more strongly, micro top-3 (-5.3 % at stage 6) while gaining macro top-1 ($+1.8$); the bias-free pool and the label-free prior correction are the remedy: with the class bias frozen the pool is ahead of the ensemble on the natural 2023 sample on every measure (stage 6: micro top-1 24.12 against 23.15, top-3 46.23 against 45.71, macro top-1 54.14 against 52.99; stage 5: 19.92 / 38.75 / 46.30 against 19.04 / 37.16 / 45.43), and after the EM prior estimate from the unlabelled 2023 posteriors it reaches 57.35 / 80.88 micro top-1 / top-3 at stage 6 against 52.33 / 80.90 for the corrected ensemble (55.51 / 77.02 against 37.94 / 72.61 at stage 4). In the reverse direction, a family trained on the 2023 balanced fit rows alone (in-year 2023 test: 59.36% top-1 at stage 6, 51.53% at stage 5) drifts on the 81,978 2024 records to 51.16% and 46.01% (-8.2 and -5.5 %), and again the pool adds $+1.00$ / $+1.24$ and $+1.00$ / $+1.26$ top-1 / top-3 at stages 6 and 5 ($p < 10^{-26}$), $+0.32$ / $+0.98$ at stage 4, the static pool $+0.77$ to $+0.91$ top-1, the fixed product -1.0 to -1.8. The

cohort (n)	st.	setting	ens. top-3 / top-5	STEP top-3 / top-5	Δ_3 / Δ_5 (STEP − ens.)
balanced 2024 test (8,198)	5	as trained	74.53/ 82.06	75.56/ 82.98	$+1.02$ / $+0.93$
natural 2024 (90,759)	5	as trained	39.40 / 55.59	36.52 / 50.42	-2.88 / -5.17
		τ=1	87.36 / 93.97	88.29 / 94.55	$+0.93$ / $+0.58$
natural 2023 (100,000)	5	as trained	37.16 / 53.29	34.64 / 48.37	-2.53 / -4.92
		τ=1	84.23 / 92.05	85.21 / 92.75	$+0.98$ / $+0.70$
balanced 2023, 24→23 (80,284)	5	as trained	67.43 / 75.85	68.67 / 76.97	$+1.24$ / $+1.13$
balanced 2024, 23→24 (81,978)	5	as trained	67.78 / 76.32	69.04 / 77.53	$+1.26$ / $+1.21$
balanced 2024 test (8,198)	6	as trained	78.56/ 85.09	79.82/ 86.36	$+1.27$ / $+1.27$
natural 2024 (90,759)	6	as trained	47.52 / 65.12	41.92 / 57.64	-5.60 / -7.49
		τ=1	88.14 / 94.67	89.45 / 95.51	$+1.32$ / $+0.85$
natural 2023 (100,000)	6	as trained	45.71 / 63.83	40.38 / 56.45	-5.34 / -7.38
		τ=1	86.02 / 93.59	87.18 / 94.50	$+1.16$ / $+0.91$
balanced 2023, 24→23 (80,284)	6	as trained	73.76 / 81.13	75.11 / 82.49	$+1.36$ / $+1.36$
balanced 2024, 23→24 (81,978)	6	as trained	72.83 / 80.59	74.06 / 81.68	$+1.24$ / $+1.10$

Table 26: Balanced-cohort ensemble and STEP applied unchanged (as trained) and after the τ=1 Bayes prior approximation (two-fold known prior; §5.3) to the real-world and multi-year samples: micro top-3 / top-5 (%) and the paired STEP − ens. difference at stages 5–6; all stages, top-1 and exact p in Table F.

increment is therefore symmetric across the two years and of the same size as within a year, while the joint models themselves lose 5 to 8 % to the year shift.

direction	st.	ens. top-1 / top-3	STEP top-1 / top-3	Δ_1 [95% CI] (p)	Δ_3 (p)	static Δ_1 / Δ_3	ens. / STEP macro top-1
2024 → 2023 balanced (80,284)	1	18.69 / 34.74	18.78 / 34.97	$+0.09$ [0.00, 0.18] (0.049)	$+0.23$ (0.00026)	$+0.06$ / $+0.15$	18.87 / 18.97
	2	22.23 / 39.93	22.35 / 40.14	$+0.12$ [0.03, 0.22] (0.012)	$+0.20$ (0.0032)	$+0.01$ / $+0.18$	22.34 / 22.49
	3	24.55 / 44.23	24.53 / 44.20	-0.02 [-0.13, 0.09] (0.7)	-0.02 (0.74)	$+0.08$ / -0.05	24.68 / 24.66
	4	41.08 / 62.41	42.06 / 63.41	$+0.98$ [0.82, 1.15] (1.3e-27)	$+1.01$ (5.1e-31)	$+0.68$ / $+0.56$	41.20 / 42.19
	5	46.07 / 67.43	47.19 / 68.67	$+1.11$ [0.94, 1.28] (1.8e-35)	$+1.24$ (3.8e-50)	$+0.88$ / $+0.95$	46.24 / 47.36
	6	52.69 / 73.76	53.84 / 75.11	$+1.15$ [0.96, 1.34] (3.2e-34)	$+1.36$ (2.8e-59)	$+0.81$ / $+1.18$	52.81 / 53.95
2024 → 2023 natural (100,000)	1	7.89 / 13.83	7.50 / 13.62	-0.39 [-0.46, -0.33] (1.3e-39)	-0.21 (7.1e-11)	-0.24 / -0.50	17.91 / 18.35
	2	8.43 / 15.64	7.94 / 15.33	-0.49 [-0.56, -0.43] (2.9e-49)	-0.31 (1.3e-12)	-0.60 / -0.34	20.82 / 21.05
	3	9.24 / 19.08	8.98 / 18.19	-0.26 [-0.33, -0.19] (1e-15)	-0.89 (4.4e-84)	-0.21 / -0.77	22.50 / 22.69
	4	18.26 / 35.25	17.12 / 31.84	-1.14 [-1.27, -1.01] (6.9e-67)	-3.41 (0)	-1.40 / -4.83	41.43 / 41.73
	5	19.04 / 37.16	18.59 / 34.64	-0.45 [-0.58, -0.32] (2.4e-12)	-2.53 (4.2e-184)	-0.63 / -3.37	45.43 / 46.39
	6	23.15 / 45.71	22.33 / 40.38	-0.82 [-0.96, -0.69] (4.6e-31)	-5.34 (0)	-1.00 / -6.17	52.99 / 54.80
2023 → 2024 balanced (81,978)	1	19.40 / 36.09	19.54 / 36.30	$+0.14$ [0.02, 0.25] (0.028)	$+0.22$ (9.9e-05)	$+0.14$ / $+0.20$	19.40 / 19.53
	2	22.18 / 39.94	22.20 / 40.17	$+0.02$ [-0.10, 0.13] (0.77)	$+0.23$ (0.00027)	-0.00 / $+0.30$	22.17 / 22.19
	3	24.40 / 44.10	24.57 / 44.08	$+0.17$ [0.06, 0.28] (0.0024)	-0.02 (0.83)	$+0.07$ / -0.15	24.40 / 24.57
	4	41.92 / 62.94	42.24 / 63.92	$+0.32$ [0.15, 0.48] (0.00014)	$+0.98$ (9.8e-34)	$+0.07$ / $+0.81$	41.91 / 42.23
	5	46.01 / 67.78	47.01 / 69.04	$+1.00$ [0.81, 1.18] (5.9e-27)	$+1.26$ (1.7e-48)	$+0.91$ / $+1.17$	46.00 / 47.00
	6	51.16 / 72.83	52.16 / 74.06	$+1.00$ [0.82, 1.17] (4.4e-27)	$+1.24$ (2.3e-50)	$+0.77$ / $+1.04$	51.15 / 52.15

Year hold-out. Models trained on one year's balanced fit rows only (clean boundary, unchanged recipe and pool selection) and applied once to the other year: a balanced 2023 cohort built with the same rules (single-protocol records, up to 1,000 per protocol, a naturally distributed 2023 sample, and, in reverse, the 2024 balanced cohort from a 2023-trained family. Micro top-1 / top-3, STEP minus the 3-seed ensemble (paired bootstrap CI, exact McNemar p), the static pool's increment, and macro top-1.

Distribution comparison at every stage. Table 26 gives, at stages 5–6, the comparison of §5.3 (the balanced-test rows are the macros of Table 1), and Table F extends it to all six stages and to top-1, on the four evaluation cohorts, as trained and after the τ=1 Bayes prior approximation with the two-fold known prior; Figure 4 (§5.3) draws the two natural cohorts against the balanced test split. The correction and the prior are computed CPU-only from the saved per-record logits; on the natural 2024 cohort the τ=1 top-1 / top-3 values reproduce the known-prior column of Table 27 exactly, and on the balanced cross-year cohorts, whose target prior is uniform to within a few records per protocol (the 2023 cohort has 7 protocols below 1,000 records, the smallest with 411), the correction moves every value by less than a point. Takeaway: at every stage the pool's class bias costs micro accuracy as trained on both natural cohorts and the τ=1 correction restores STEP's lead on every depth, most clearly from the scene stage on.

cohort	st.	setting	ens. top-1 / 3 / 5	STEP top-1 / 3 / 5	Δ_1 (p)	Δ_3 (p)	Δ_5
nat. 2024	1	as trained	9.30 / 15.51 / 19.00	8.91 / 15.27 / 18.67	-0.40 (5.8×10^{-38})	-0.24 (3.4×10^{-12})	-0.33
		τ=1	53.38 / 78.13 / 85.74	53.59 / 78.23 / 85.86	$+0.21$ (6.6×10^{-6})	$+0.10$ (0.018)	$+0.13$
	2	as trained	9.58 / 17.17 / 22.69	9.07 / 16.88 / 22.56	-0.50 (5.5×10^{-45})	-0.29 (4.1×10^{-11})	-0.14
		τ=1	54.00 / 79.53 / 87.07	54.65 / 79.70 / 87.18	$+0.65$ (3.1×10^{-23})	$+0.17$ (1.8×10^{-5})	$+0.12$
	3	as trained	10.06 / 19.87 / 25.38	9.77 / 19.12 / 26.97	-0.30 (3.0×10^{-18})	-0.75 (3.4×10^{-63})	$+1.59$
		τ=1	54.27 / 80.44 / 88.25	55.03 / 80.68 / 88.33	$+0.76$ (1.1×10^{-26})	$+0.23$ (6.0×10^{-7})	$+0.07$
	4	as trained	19.49 / 37.18 / 51.61	18.31 / 33.49 / 45.51	-1.18 (1.8×10^{-63})	-3.68 ($<10^{-300}$)	-6.10
		τ=1	55.23 / 84.73 / 92.31	60.28 / 86.98 / 93.56	$+5.05$ ($<10^{-300}$)	$+2.25$ (9.7×10^{-184})	$+1.25$
	5	as trained	20.92 / 39.40 / 55.59	20.23 / 36.52 / 50.42	-0.68 (1.6×10^{-24})	-2.88 (9.7×10^{-219})	-5.17
		τ=1	60.16 / 87.36 / 93.97	61.98 / 88.29 / 94.55	$+1.82$ (3.7×10^{-81})	$+0.93$ (1.0×10^{-52})	$+0.58$
	6	as trained	24.42 / 47.52 / 65.12	23.24 / 41.92 / 57.64	-1.19 (9.2×10^{-57})	-5.60 ($<10^{-300}$)	-7.49
		τ=1	60.81 / 88.14 / 94.67	63.45 / 89.45 / 95.51	$+2.64$ (2.3×10^{-137})	$+1.32$ (4.0×10^{-90})	$+0.85$
nat. 2023	1	as trained	7.89 / 13.83 / 17.59	7.50 / 13.62 / 17.19	-0.39 (1.3×10^{-39})	-0.21 (7.1×10^{-11})	-0.40
		τ=1	51.77 / 74.83 / 83.40	52.26 / 74.83 / 83.49	$+0.49$ (4.3×10^{-27})	-0.00 (0.98)	$+0.09$
	2	as trained	8.43 / 15.64 / 21.96	7.94 / 15.33 / 21.89	-0.49 (2.9×10^{-49})	-0.31 (1.3×10^{-12})	-0.08
		τ=1	52.52 / 75.76 / 85.25	53.35 / 75.75 / 85.39	$+0.83$ (4.2×10^{-44})	-0.01 (0.8)	$+0.14$
	3	as trained	9.24 / 19.08 / 24.70	8.98 / 18.19 / 27.21	-0.26 (1.0×10^{-15})	-0.89 (4.4×10^{-84})	$+2.50$
		τ=1	53.02 / 77.22 / 86.75	54.00 / 77.30 / 86.70	$+0.99$ (1.5×10^{-48})	$+0.08$ (0.11)	-0.04
	4	as trained	18.26 / 35.25 / 49.37	17.12 / 31.84 / 43.47	-1.14 (6.9×10^{-67})	-3.41 ($<10^{-300}$)	-5.90
		τ=1	52.76 / 81.71 / 90.24	57.30 / 83.59 / 91.61	$+4.53$ ($<10^{-300}$)	$+1.88$ (1.7×10^{-143})	$+1.37$
	5	as trained	19.04 / 37.16 / 53.29	18.59 / 34.64 / 48.37	-0.45 (2.4×10^{-12})	-2.53 (4.2×10^{-184})	-4.92
		τ=1	57.04 / 84.23 / 92.05	58.76 / 85.21 / 92.75	$+1.72$ (1.5×10^{-86})	$+0.98$ (2.2×10^{-56})	$+0.70$
	6	as trained	23.15 / 45.71 / 63.83	22.33 / 40.38 / 56.45	-0.82 (4.6×10^{-31})	-5.34 ($<10^{-300}$)	-7.38
		τ=1	57.84 / 86.02 / 93.59	60.36 / 87.18 / 94.50	$+2.53$ (1.9×10^{-144})	$+1.16$ (2.6×10^{-67})	$+0.91$
bal. 2023 (24→23)	1	as trained	18.69 / 34.74 / 44.54	18.78 / 34.97 / 44.57	$+0.09$ (0.049)	$+0.23$ (2.6×10^{-4})	$+0.03$
		τ=1	18.81 / 34.93 / 44.72	18.84 / 35.12 / 44.83	$+0.03$ (0.6)	$+0.20$ (0.0013)	$+0.11$
	2	as trained	22.23 / 39.93 / 50.09	22.35 / 40.14 / 50.59	$+0.12$ (0.012)	$+0.20$ (0.0032)	$+0.50$
		τ=1	22.32 / 40.10 / 50.25	22.26 / 40.20 / 50.68	-0.06 (0.23)	$+0.09$ (0.17)	$+0.43$
	3	as trained	24.55 / 44.23 / 54.61	24.53 / 44.20 / 54.72	-0.02 (0.7)	-0.02 (0.74)	$+0.11$
		τ=1	24.58 / 44.24 / 54.86	24.54 / 44.31 / 54.93	-0.04 (0.52)	$+0.07$ (0.33)	$+0.07$
	4	as trained	41.08 / 62.41 / 71.21	42.06 / 63.41 / 72.33	$+0.98$ (1.3×10^{-27})	$+1.01$ (5.1×10^{-31})	$+1.12$
		τ=1	41.12 / 62.47 / 71.27	42.02 / 63.50 / 72.40	$+0.90$ (1.5×10^{-23})	$+1.03$ (9.6×10^{-33})	$+1.13$
	5	as trained	46.07 / 67.43 / 75.85	47.19 / 68.67 / 76.97	$+1.11$ (1.8×10^{-35})	$+1.24$ (3.8×10^{-50})	$+1.13$
		τ=1	46.04 / 67.47 / 75.87	47.19 / 68.73 / 76.96	$+1.15$ (1.7×10^{-37})	$+1.26$ (1.5×10^{-51})	$+1.09$
	6	as trained	52.69 / 73.76 / 81.13	53.84 / 75.11 / 82.49	$+1.15$ (3.2×10^{-34})	$+1.36$ (2.8×10^{-59})	$+1.36$
		τ=1	52.72 / 73.79 / 81.15	53.92 / 75.21 / 82.55	$+1.20$ (1.8×10^{-37})	$+1.42$ (1.7×10^{-64})	$+1.41$
bal. 2024 (23→24)	1	as trained	19.40 / 36.09 / 45.69	19.54 / 36.30 / 45.97	$+0.14$ (0.028)	$+0.22$ (9.9×10^{-5})	$+0.28$
		τ=1	19.40 / 36.30 / 45.84	19.50 / 36.40 / 46.01	$+0.09$ (0.08)	$+0.10$ (0.1)	$+0.17$
	2	as trained	22.18 / 39.94 / 50.53	22.20 / 40.17 / 50.82	$+0.02$ (0.77)	$+0.23$ (2.7×10^{-4})	$+0.29$
		τ=1	22.50 / 40.08 / 50.50	22.55 / 40.25 / 50.78	$+0.05$ (0.4)	$+0.17$ (0.0064)	$+0.28$
	3	as trained	24.40 / 44.10 / 54.51	24.57 / 44.08 / 54.64	$+0.17$ (0.0024)	-0.02 (0.83)	$+0.13$
		τ=1	24.77 / 44.17 / 54.59	24.85 / 44.18 / 54.75	$+0.08$ (0.17)	$+0.01$ (0.94)	$+0.16$
	4	as trained	41.92 / 62.94 / 72.19	42.24 / 63.92 / 73.09	$+0.32$ (1.4×10^{-4})	$+0.98$ (9.8×10^{-34})	$+0.90$
		τ=1	41.99 / 62.99 / 72.19	42.33 / 63.92 / 73.05	$+0.34$ (3.6×10^{-5})	$+0.92$ (5.1×10^{-30})	$+0.86$
	5	as trained	46.01 / 67.78 / 76.32	47.01 / 69.04 / 77.53	$+1.00$ (5.9×10^{-27})	$+1.26$ (1.7×10^{-48})	$+1.21$
		τ=1	46.10 / 67.81 / 76.37	47.16 / 69.07 / 77.58	$+1.06$ (1.3×10^{-29})	$+1.26$ (2.4×10^{-48})	$+1.21$
	6	as trained	51.16 / 72.83 / 80.59	52.16 / 74.06 / 81.68	$+1.00$ (4.4×10^{-27})	$+1.24$ (2.3×10^{-50})	$+1.10$
		τ=1	51.21 / 72.87 / 80.60	52.30 / 74.12 / 81.72	$+1.09$ (8.3×10^{-32})	$+1.25$ (2.1×10^{-51})	$+1.12$

Distribution comparison, all six stages: the balanced-cohort 3-seed ensemble and STEP applied unchanged (as trained) and after the τ=1 Bayes prior approximation (two-fold known prior, as in Table 26; π_{train} is uniform for the 2024-trained family and the empirical 2023 balanced-cohort frequencies for the 2023-trained family) to the naturally distributed 2024 (90,759) and 2023 (100,000) samples and to the balanced cross-year hold-outs (80,284 and 81,978). Micro top-1 / top-3 / top-5 (%), STEP minus ensemble in points with the exact two-sided McNemar p for top-1 and top-3 (descriptive; the top-5 p values are in the repository's `dist_compare_r5.json`). On the natural 2024 cohort the τ=1 top-1 / top-3 values reproduce the known-prior column of Table 27 exactly. On the balanced cross-year cohorts the target prior is near-uniform, so τ=1 is close to the identity.

The prior-shift trade-off. Table 27 reports the remaining views of the naturally distributed 2024 cohort at stages 5–6 with the models of Table 1 applied unchanged. As trained, the pool *lowers* micro accuracy: at stage 6 micro top-1 falls from 24.42% to 23.24% and top-3 from 47.52% to 41.92% ($p < 10^{-50}$), because the pool's per-class bias, fitted on balanced rows, moves mass away from the majority protocol. What transfers is the protocol-level effect: macro top-1 rises from 58.18% to 60.05% at stage 6, and once the deployment prior is supplied ($\log \pi_{\text{ext}}/\pi_{\text{train}}$ added to every method's log-posterior, the τ=1 correction of Table 26) micro top-1 rises from 60.81% to 63.45%; the label-free EM estimate of the prior (Saerens et al., 2002) gives the same ordering. What is *exploratory* is the remedy, a refit with the class bias frozen at zero after these results had been inspected. On the balanced test split it keeps $+0.73$ / $+1.43$ of the stage-6 increment, and on the natural cohort it is ahead of the ensemble on micro top-1 (25.22% against 24.42%) and macro F1 (20.38 against

model	st.	balanced test	natural 2024 cohort, as trained			known prior	EM-estimated prior	
		Δ_1 / Δ_3 vs ens.	micro top-1 / top-3	macro top-1	macro F1 (top-1 / top-3)	micro top-1 / top-3	micro top-1 / top-3	macro F1
3-seed ensemble	5	+0.00 / +0.00	20.92 / 39.40	55.36	17.30 / 13.79	60.16 / 87.36	53.06 / 84.84	24.55
static pool	5	+0.37 / +0.82	20.11 / 35.86	56.55	17.12 / 13.47	61.64 / 88.02	57.54 / 85.95	23.57
STEP	5	+0.77 / +1.02	20.23 / 36.52	56.62	17.09 / 13.48	61.98 / 88.29	54.24 / 86.17	23.90
STEP, class bias frozen at 0	5	+0.38 / +0.84	21.65 / 40.97	56.13	17.76 / 14.00	–	62.06 / 86.23	17.85
3-seed ensemble	6	+0.00 / +0.00	24.42 / 47.52	58.18	20.04 / 15.60	60.81 / 88.14	55.38 / 85.45	28.48
static pool	6	+0.57 / +1.18	23.20 / 41.11	58.92	19.69 / 15.03	62.96 / 89.17	61.06 / 88.32	27.90
STEP	6	+0.96 / +1.27	23.24 / 41.92	60.05	19.58 / 15.01	63.45 / 89.45	60.41 / 88.41	26.98
STEP, class bias frozen at 0	6	+0.73 / +1.43	25.22 / 47.65	59.66	20.38 / 15.68	–	62.78 / 87.71	17.99

Table 27: The prior-shift trade-off on the naturally distributed NEMSIS 2024 cohort (90,759 records, one protocol is 54% of it; models and pools exactly as in Table 1). As trained, the pool's fitted class bias lowers micro accuracy under the deployment prior while macro top-1 (per-protocol mean) rises. "Known prior": $\log \pi_{\text{ext}}/\pi_{\text{train}}$ added to the log-posterior, prior estimated on one random half and scored on the other. "EM": prior estimated from the unlabelled posteriors (Saerens et al., 2002); it raises micro accuracy for every model and lowers macro F1 for the bias-free variant, because it moves rare-protocol predictions to the majority protocol. The indented row is *exploratory*: the bias was frozen after the external results had been inspected, so its external numbers are not a held-out test of that choice.

20.04); the natural 2023 sample repeats the pattern (above). Because the choice was made after the external cohort had been seen, those numbers are not a held-out test of it. Under an unknown prior the recommendation is a bias-free pool with an estimated prior; the EM correction trades macro recall for micro accuracy, and the paper reports both variants.

Prior-robust pooling (exploratory). This variant was developed after the external results above had been inspected, so its external numbers are not a held-out test of the choice; the naturally distributed 2023 sample of the year hold-out was built afterwards and repeats the pattern, but was analysed knowing it. The external loss on raw micro top-1 has a specific cause and a specific fix. The pool's per-class bias b (82 parameters) is fitted on balanced rows and absorbs the training prior; under prior shift it works against the deployment distribution. Refitting the static and the gated pool with b frozen at zero (Table 29) keeps most of the balanced-test gain ($+0.38$ / $+0.84$ at stage 5 and $+0.73$ / $+1.43$ at stage 6 for the gated pool, against $+0.77$ / $+1.02$ and $+0.96$ / $+1.27$ with the bias) and turns the external micro loss into a gain (stage 6: 25.22% against 24.42% for the ensemble and 23.24% with the bias; stage 5: 21.65 against 20.92 and 20.23), while macro top-1 stays above the ensemble (59.66 against 58.18). The bias-free pool also makes the deployment prior estimable without labels: the EM procedure of Saerens et al. (2002) applied to the unlabelled external posteriors recovers the external class frequencies to an L_1 error of 0.15–0.30 for the bias-free pools against 0.43–0.75 for the ensemble and 0.31–0.68 for the pools with bias, and the label-free correction then gives 59.5 / 62.1 / 62.8 micro top-1 at stages 4–6 (top-3 83.8 / 86.2 / 87.7) against 37.9 / 53.1 / 55.4 for the corrected ensemble. The recommendation for deployment under an unknown prior is therefore a bias-free pool with an EM-estimated prior; it costs about a quarter of the balanced-test top-1 increment and none of the top-3 increment.

Variance-aware gating. If the pool's gain is variance control, the natural gate signal is each input's estimation variance, which the three seeds of every input estimate directly (the mean KL divergence of the seed posteriors from their mean, the spread of the maximum posterior and the top-1 agreement rate). Replacing STEP's confidence summaries by these disagreement summaries, or adding them, on the identical inputs and recipe changes nothing at stages 1–5 (differences within ± 0.3 %, none significant) and adds $+0.34$ / $+0.50$ top-1 at stage 6 for the disagreement-only and the combined gate over the confidence gate ($p = 0.11$ and 0.012; combined gate $+1.28$ / $+1.28$ over the ensemble), even though the joint input's disagreement correlates 0.30–0.38 with its own error at stages 4–6. The gate is not where the increment lives; the static weights carry it.

Pool distillation into one network. Since the pooled posterior is the best estimate available and the masked network is the cheapest, distilling the former into the latter (the masked network's cross-entropy plus λ times the KL divergence from STEP's posterior on the same fit row's out-of-fold inputs, at every stage view; $\lambda = 1$) gives a single network that serves all six stages at the cost of one forward pass. The distilled single network beats the per-stage single network by $+1.81$, $+1.43$ and $+1.61$ top-1 % at stages 4–6 ($p < 10^{-3}$), which is two thirds of the single-network-to-STEP

		STEP			stage MLP (seed 0)			Qwen3-32B, k=3			MedGemma-27B, k=3		
k	stage	P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	1	19.1	20.7	16.0	19.0	20.4	16.3	12.4	15.0	10.5	11.4	15.6	10.4
	2	23.2	24.6	20.9	22.7	24.5	21.3	12.5	15.9	11.1	12.0	16.2	10.9
	3	26.9	27.8	24.5	25.3	27.3	23.8	13.2	16.3	11.1	13.2	16.4	11.3
	4	47.7	48.2	46.6	45.2	45.8	44.7	32.3	25.7	21.0	29.2	22.7	18.0
	5	53.5	53.7	52.2	50.0	50.3	49.0	32.2	27.1	23.0	29.5	24.8	20.1
	6	58.7	59.3	58.0	56.1	56.0	55.2	36.2	33.3	28.9	35.9	28.8	24.4
3	1	15.2	38.3	19.7	15.2	38.1	19.8	13.1	22.8	12.3	15.6	24.9	13.8
	2	15.9	43.9	22.1	15.8	43.4	22.1	12.5	24.6	12.6	15.4	26.1	14.6
	3	17.0	48.4	24.3	16.8	47.8	24.2	12.8	26.3	12.9	16.6	26.3	14.8
	4	25.4	70.5	36.8	25.2	67.3	36.0	26.6	36.0	21.3	27.3	36.9	21.6
	5	27.3	75.6	39.6	26.6	72.1	38.2	26.9	36.7	23.1	26.5	40.1	23.3
	6	30.1	79.8	42.9	30.2	76.0	42.1	30.0	42.7	27.7	30.8	45.3	26.9

Table 28: Macro-averaged precision (P), recall (R) and F1 over the 82 protocols on the 8,198 test records of the balanced cohort, in %, with "predicted positive" = "appears in the top k" (at k=1 this is ordinary multi-class macro P/R/F1; at k=3 every record contributes three predicted positives, so precision falls and recall equals top-3 accuracy). Because the cohort is balanced (100 test records per protocol), macro and support-weighted averages coincide to the first decimal. STEP and the stage MLP are clean-boundary models of Table 1; the prompted arms ran on the same family (Qwen3: mean over demonstration seeds 0–2; MedGemma: seed 0).

gap; three such networks ensembled are level with the per-stage ensemble on top-1 (-0.05, -0.13, $+0.12$) and ahead on top-3 at stage 4 ($+0.89$, $p = 0.006$), and the harmful-transition rate at $5 \to 6$ falls to 2.65% (single) against 5.07% for the per-stage ensemble. This is ensemble distillation (Lakshminarayanan et al., 2017) with a stronger teacher, not a new mechanism; its value is a deployable single model with the pool's consistency at a third of the ensemble's inference cost and none of the pool's.

Beyond top-k: precision, recall and F1. Tables 28 and 29 give, for every method, macro precision, recall and F1 of the top-1 prediction and under the paper's top-3 convention (predicted positive = in the shortlist), on the balanced test split and on the external cohort. Three readings change what the top-k numbers alone suggest. First, on the balanced split the pool's gain is a recall and F1 gain at top-1 (stage 6: macro F1 57.98 for STEP against 57.55 for the ensemble, macro recall 59.30 against 58.34) but not a precision gain under the top-3 convention: STEP's shortlists contain more spurious entries per protocol (macro precision 30.09 against 31.67), so its top-3 macro F1 is below the ensemble's (42.89 against 43.91) even though its top-3 accuracy is 1.3 % higher; the bias-free pool has the best macro F1 under both conventions (58.12 and 44.17), so the class bias is what buys top-3 hits at the price of shortlist precision. Second, the gradient-boosted model, level with the ensemble on top-1, has the highest top-3 macro F1 of the per-stage models (44.03), which the top-k table hides. Third, on the external cohort the label-free prior correction raises micro accuracy at the cost of macro recall (stage 6, bias-free STEP: micro top-1 25.2 to 62.8, macro recall 57.5 to 16.5, macro F1 20.4 to 18.0), because it moves rare-protocol predictions to the majority protocol; the uncorrected bias-free pool has the best external macro F1 of all methods (20.60 / 20.38 against 20.04 for the ensemble). Which variant to deploy therefore depends on the objective, and the paper reports both rather than choosing.

When does pooling earlier posteriors help? A synthetic study. Table 30 simulates two nested blocks: a 20-dimensional informative first block and a 300-dimensional second block of which a fraction f_2 is informative, revealed in that order to two per-stage MLPs trained and pooled under the clean protocol, over sample sizes 2,000 to 32,000, with and without redundancy of the earlier block, 50% missingness of the new block and a cross-block interaction in which the second block is informative only jointly with the first. Two regimes appear. When the added block is uninformative or weak and the sample is small, the later stage loses 4 to 24 % to the added dimensions (the peaking phenomenon), 5 to 29% of records flip from correct to incorrect, and the pool recovers most of the loss (static $+4$ to $+11$ at n = 2,000 and still $+4$ at n = 32,000 when the block is pure noise; the

st.	method	micro top-1	micro top-3	macro $F1_1$	macro $F1_3$	EM top-1	EM top-3	EM macro $F1_1$
4	ensemble	19.49	37.18	15.62	12.69	37.85	75.50	21.01
	static pool	18.22	32.50	15.28	12.25	46.80	82.27	18.91
	STEP	18.31	33.49	15.07	12.18	48.27	83.69	18.17
	static pool, no class bias	19.48	37.11	15.62	12.66	58.83	82.83	14.34
	STEP, no class bias	19.76	37.28	15.63	12.60	59.50	83.81	13.63
	masked ens.	18.87	35.98	15.27	12.56	43.50	79.95	21.49
5	ensemble	20.92	39.40	17.30	13.79	53.06	84.84	24.55
	static pool	20.11	35.86	17.12	13.47	57.54	85.95	23.57
	STEP	20.23	36.52	17.09	13.48	54.24	86.17	23.90
	static pool, no class bias	21.63	40.86	17.66	13.96	60.78	85.39	18.01
	STEP, no class bias	21.65	40.97	17.76	14.00	62.06	86.23	17.85
	masked ens.	20.61	38.67	17.30	13.61	49.76	83.22	24.60
6	ensemble	24.42	47.52	20.04	15.60	55.38	85.45	28.48
	static pool	23.20	41.11	19.69	15.03	61.06	88.32	27.90
	STEP	23.24	41.92	19.58	15.01	60.41	88.41	26.98
	static pool, no class bias	25.46	48.06	20.60	15.76	61.69	85.26	21.44
	STEP, no class bias	25.22	47.65	20.38	15.68	62.78	87.71	17.99
	masked ens.	23.66	43.46	19.46	14.91	50.05	84.50	27.84

Table 29: External cohort (90,759 naturally distributed records), stages 4–6: micro accuracy, macro F1 (top-1 and top-3 conventions), and the same after a label-free EM estimate of the deployment prior from the unlabelled external posteriors (Saerens et al., 2002). "No class bias": the pool refitted with its per-class bias frozen at zero.

n	f_2	interaction	acc. stage 1	acc. stage 2	harm	gain	static Δ	gated Δ	PA Δ	static w
2,000	0.00	no	69.8	45.5	28.6	4.4	$+11.12$	$+24.12$	$+14.25$	(0.27, 0.72)
2,000	0.05	no	69.8	65.0	17.4	12.6	$+6.50$	$+12.12$	$+9.00$	(0.26, 0.91)
2,000	0.30	no	69.8	99.4	0.2	29.9	-0.62	-0.88	-0.88	(0.21, 1.22)
2,000	0.30	yes	69.8	98.0	0.9	29.1	$+0.62$	$+0.75$	$+0.50$	(0.21, 1.22)
8,000	0.00	no	70.8	61.3	14.9	5.4	$+9.03$	$+9.31$	$+4.44$	(0.75, 0.20)
8,000	0.05	no	70.8	79.8	7.2	16.2	$+2.00$	$+1.69$	$+1.69$	(0.47, 0.52)
8,000	0.30	no	70.8	99.5	0.2	28.8	-0.03	$+0.00$	$+0.03$	(0.18, 1.24)
8,000	0.30	yes	70.8	99.6	0.3	29.0	$+0.06$	$+0.06$	$+0.06$	(0.17, 1.14)
32,000	0.00	no	69.7	65.6	8.6	4.4	$+4.23$	$+4.28$	$+1.94$	(0.96, 0.04)
32,000	0.05	no	69.7	82.0	5.0	17.3	$+0.63$	$+0.71$	$+0.71$	(0.25, 0.52)
32,000	0.30	no	69.7	99.7	0.1	30.1	-0.02	-0.02	-0.04	(0.16, 0.91)
32,000	0.30	yes	69.7	99.6	0.2	30.1	$+0.02$	$+0.02$	$+0.02$	(0.09, 0.94)

Table 30: Synthetic progressive reveal (two blocks, 20 classes; block 2 has 300 dimensions of which a fraction f_2 is informative; mean over two generator seeds; test top-1 in %). Harm and gain: records ranked correctly by stage 1 and not by stage 2, and the reverse. Δ: pooled model minus the stage-2 three-seed ensemble under the same clean protocol (static and gated pools fitted on out-of-fold posteriors and selected on validation; PA the fixed product with $w = (1/2, 1)$). Static w: the fitted stage weights. The full grid with redundancy and missingness is in the repository (`SYNTHETIC.md`).

gated pool up to $+24$), with the validation-fitted weight on the earlier stage rising as the sample grows and the new block stays noisy (0.27 at $n = 2{,}000$ to 0.96 at $n = 32{,}000$). When the added block is strongly informative the later stage is near-perfect, harmful flips vanish, validation returns weights close to the identity and the pool is inert (± 0.6 %), while at $n = 2{,}000$ the fixed products lose 1 to 2 %. Redundancy of the earlier block and missingness of the new one reduce both the loss and the recovery without changing the regime. Under the cross-block interaction the same pattern holds: the earlier posterior is worth $+0.6$ % at $n = 2{,}000$ and nothing once the joint model can learn the interaction. The synthetic study is consistent with finite-sample variance control on the new block as the mechanism, and identifies the failure mode as an unlearned or over-trusted earlier member; recovery of accuracy alone does not separate variance from optimisation, regularisation or representation effects, which the regularised-joint control below addresses in part. It is not clinical evidence.

Does regularising the joint predictor remove the gain? A reader may suspect that the pool merely compensates for an under-regularised joint model. Table 31 repeats the synthetic study with the stage-2 joint model tuned on the validation split over weight decay $\{0, 10^{-4}, 10^{-3}, 10^{-2}\}$, dropout $\{0.2, 0.5\}$ and early stopping on validation NLL, the selected configuration ensembled over three seeds and the static pool over the stage-1 and tuned stage-2 ensembles fitted and selected as

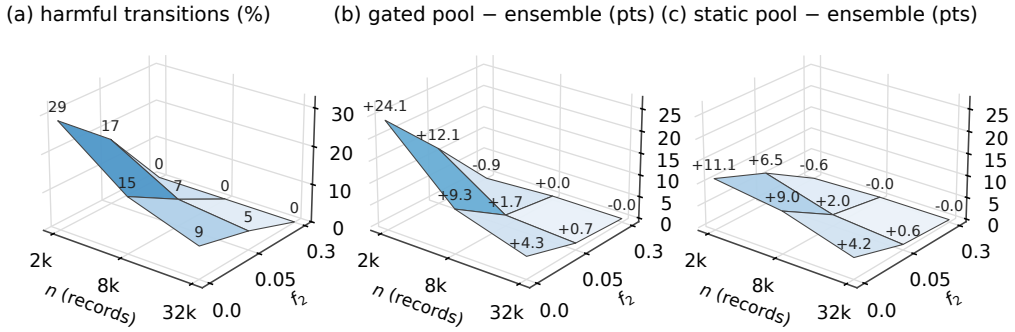

Figure 21: The synthetic regime map (Table 30: two blocks, 20 classes, block 2 has 300 dimensions of which a fraction f_2 is informative; mean over generator seeds). (a) Harmful transitions, records ranked correctly by the stage-1 model and not by the stage-2 model, as a function of training-set size n and f_2. (b, c) The gated and the static pool minus the stage-2 three-seed ensemble (test top-1, %) on the same grid; the values are those of Table 30. Takeaway: the gain tracks the harm rate, both are largest where the added block is high-dimensional, weakly informative and n is small, and both vanish where the block is strongly informative.

n	f_2	stage 1	stage 2 plain	stage 2 tuned	tuned (wd / dropout)	pool − plain ens.	pool − tuned ens.
2,000	0.0	70.2	43.4	46.5	0/0.5, 0.001/0.5	+14.37	+11.38
2,000	0.05	70.2	63.5	65.5	0.01/0.2, 0/0.5	+7.25	+6.38
2,000	0.3	70.2	99.4	99.2	0/0.2, 0/0.5	−0.75	−0.50
8,000	0.0	70.9	61.8	62.0	0.0001/0.5, 0.0001/0.5	+8.44	+8.16
8,000	0.05	70.9	80.3	80.1	0.01/0.5, 0.001/0.5	+1.88	+1.69
8,000	0.3	70.9	99.4	99.4	0/0.2, 0/0.5	+0.06	+0.00
32,000	0.0	69.9	66.3	66.4	0.0001/0.2, 0/0.2	+3.55	+3.46
32,000	0.05	69.9	82.9	82.9	0.001/0.2, 0.01/0.5	+0.44	+0.13
32,000	0.3	69.9	99.7	99.8	0.001/0.5, 0.001/0.2	−0.01	+0.00

Table 31: Does regularising the joint predictor remove the pooling gain? Same generator and protocol as Table 30 (two blocks, 20 classes; block 2 has 300 dimensions of which a fraction f_2 is informative). For the stage-2 joint model a grid over weight decay $\{0, 10^{-4}, 10^{-3}, 10^{-2}\}$, dropout $\{0.2, 0.5\}$ and early stopping on validation NLL is tuned on the validation split by top-1; the selected configuration (listed per repetition) is ensembled over three seeds and the static pool over the stage-1 and stage-2 ensembles is fitted and selected as before. Test top-1, mean over repetitions.

before. Tuning raises the joint model by up to three % at $n = 2{,}000$ and by nothing at $n \geq 8{,}000$, and the pool's gain over the tuned ensemble remains: $+11.4$ against $+14.4$ over the untuned one when the added block is pure noise at $n = 2{,}000$, $+6.4$ against $+7.3$ when it is weakly informative, $+8.2$ against $+8.4$ and $+1.7$ against $+1.9$ at $n = 8{,}000$, $+3.5$ against $+3.5$ for pure noise at $n = 32{,}000$, and $+0.1$ against $+0.4$ for the weakly informative block at $n = 32{,}000$, the one cell where tuning removes most of a small gain; where the block is strongly informative the pool is inert with or without tuning. Regularising the joint predictor is therefore not a substitute for pooling in this generator, which narrows, without closing, the alternative that the increment is an artefact of an under-tuned joint model.

A second, unrelated staged-feature task. Table 32 applies the protocol to UCI Covertype (581,012 records, 7 cover types, 54 features) with an artificial, documented reveal order of its natural feature groups: topography (elevation, aspect, slope), then distances to hydrology, roads and fire % with the three hillshades, then the four wilderness indicators, then the 40 soil-type indicators; continuous columns are standardised on the fit rows, subsamples of 2,000, 5,000 and 20,000 records (seed 0) are split 70/30 with a 15% inner validation, and per-stage MLPs ($[256, 128]$, 60 epochs,

n	stage	single	ensemble top-1 / top-3	static Δ_1 (p)	gated Δ_1 (p)	gated Δ_3 (p)
2,000	1 topography	66.00	65.00 / 99.17	+0.00 (1)	+0.00 (1)	+0.00 (1)
2,000	2 distances/hillshade	68.50	69.17 / 99.67	+0.00 (1)	−0.17 (1)	−0.17 (1)
2,000	3 wilderness	70.33	70.67 / 99.00	+0.00 (1)	−0.67 (0.39)	+0.17 (1)
2,000	4 soil	68.33	69.00 / 99.17	+0.00 (1)	+0.00 (1)	+0.33 (0.62)
5,000	1 topography	66.73	67.13 / 99.07	+0.13 (0.69)	+0.07 (1)	+0.00 (1)
5,000	2 distances/hillshade	72.47	72.00 / 99.53	+0.13 (0.69)	+0.00 (1)	+0.07 (1)
5,000	3 wilderness	73.67	73.47 / 99.67	−0.13 (0.8)	+0.13 (0.69)	+0.00 (1)
5,000	4 soil	74.67	74.13 / 99.40	+0.40 (0.33)	+0.47 (0.34)	+0.00 (1)
20,000	1 topography	68.77	68.73 / 98.93	+0.00 (1)	+0.07 (0.6)	+0.00 (1)
20,000	2 distances/hillshade	73.38	73.88 / 99.47	−0.27 (0.18)	+0.07 (0.78)	+0.05 (0.38)
20,000	3 wilderness	76.62	76.55 / 99.62	+0.18 (0.17)	+0.25 (0.2)	+0.07 (0.22)
20,000	4 soil	79.62	79.77 / 99.73	+0.35 (0.11)	+0.25 (0.26)	+0.02 (1)
2,000 (+500 noise cols)	1 topography	66.00	65.00 / 99.17	+0.00 (1)	+0.00 (1)	+0.00 (1)
2,000 (+500 noise cols)	2 distances/hillshade	68.50	69.17 / 99.67	+0.00 (1)	−0.17 (1)	−0.17 (1)
2,000 (+500 noise cols)	3 wilderness	70.33	70.67 / 99.00	+0.00 (1)	−0.67 (0.39)	+0.17 (1)
2,000 (+500 noise cols)	4 soil	69.83	71.33 / 98.00	+1.83 (0.14)	+2.83 (0.068)	+1.50 (0.022)
5,000 (+500 noise cols)	1 topography	66.73	67.13 / 99.07	+0.13 (0.69)	+0.07 (1)	+0.00 (1)
5,000 (+500 noise cols)	2 distances/hillshade	72.47	72.00 / 99.53	+0.13 (0.69)	+0.00 (1)	+0.07 (1)
5,000 (+500 noise cols)	3 wilderness	73.67	73.47 / 99.67	−0.13 (0.8)	+0.13 (0.69)	+0.00 (1)
5,000 (+500 noise cols)	4 soil	76.33	75.87 / 98.33	+0.47 (0.61)	+3.93 (3.3e-06)	+1.13 (1.5e-05)
20,000 (+500 noise cols)	1 topography	68.77	68.73 / 98.93	+0.00 (1)	+0.07 (0.6)	+0.00 (1)
20,000 (+500 noise cols)	2 distances/hillshade	73.38	73.88 / 99.47	−0.27 (0.18)	+0.07 (0.78)	+0.05 (0.38)
20,000 (+500 noise cols)	3 wilderness	76.62	76.55 / 99.62	+0.18 (0.17)	+0.25 (0.2)	+0.07 (0.22)
20,000 (+500 noise cols)	4 soil	78.83	79.98 / 98.90	+4.08 (2.6e-25)	+3.85 (2e-22)	+0.75 (1.2e-08)

Table 32: A second, unrelated staged-feature task: UCI Covertype (7 cover types) with an artificial, documented reveal order of its natural feature groups (topography; distances and hillshade; wilderness area; soil type), two subsample sizes, the paper's protocol (per-stage MLP $\times$ 3 seeds, block experts, out-of-fold pool fitting inside the fit rows, validation selection, one read of a 30 percent test split). Δ: pool minus the 3-seed ensemble, exact McNemar p (descriptive).

three seeds), block experts, out-of-fold pool fitting inside the fit rows, validation selection and a single test read follow the paper. On the unmodified task the pool is inert at every size (increments within ± 0.5 %, all $p > 0.1$; top-3 is above 98% from stage 1, so the task leaves no room there): the blocks are low-dimensional and the joint model is not in the finite-sample regime. To test the prediction rather than the dataset, the last block was padded with 500 sparse binary columns whose activation rate (1–5%) depends weakly on the class (a log-normal tilt with standard deviation 0.3), i.e. a high-dimensional, weakly informative block of the kind the synthetic study identifies. The stage-4 ensemble then loses 0.3 to 2.2 top-1 % to the added columns and the pool recovers more than it lost: $+2.8$ top-1 ($p = 0.07$, 600 test records) at $n = 2{,}000$, $+3.9$ ($p = 3.3 \times 10^{-6}$) and $+1.1$ top-3 at $n = 5{,}000$, and $+3.9$ ($p = 2 \times 10^{-22}$; static pool $+4.1$) and $+0.8$ top-3 at $n = 20{,}000$, while the harmful-transition rate at $3 \rightarrow 4$ falls from 10.9% to 4.7% and 9.8% to 4.7%. At the two smaller sizes the gate carries the gain and the static pool little of it ($+0.5$ at $n = 5{,}000$); at $n = 20{,}000$ the static pool carries it all, as on NEMSIS. This is one dataset with one artificial reveal order and one synthetic padding, so it shows the increment is not specific to NEMSIS and appears in the predicted regime; it does not show that natural staged blocks elsewhere sit in that regime.