FROM PROBABILITIES TO PRESCRIPTIONS: A UNIFIED DEPENDENCY-AWARE SET DECODER FOR MEDICATION AND HERB RECOMMENDATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Western medication and traditional Chinese medicine (TCM) herb recommendations differ in clinical theories, vocabularies, and safety knowledge, yet share a common set-valued prediction problem: mapping a clinical presentation to a variable-size prescription. Candidate-wise probabilities alone do not generally determine the Jaccard-optimal prescription, because the same marginals can arise from different candidate co-occurrence patterns. We therefore formulate both tasks as post-hoc set decoding and propose *Dependency-Aware Prescription Selection* (DAPS), a backbone-agnostic decoder for frozen probabilistic recommenders. DAPS uses two cross-fitted stages: the first calibrates candidate scores and constructs a provisional prescription, while the second refines these scores using set-conditioned context. Crucially, it separates behavioral co-selection and exclusion learned from recorded prescriptions from externally listed conflicts such as drug–drug interactions or herb incompatibilities. The refined probabilities are converted into a variable-size prescription through Jaccard-oriented cardinality selection, followed by a validation-selected global external-conflict budget. Across two electronic health record (EHR) and two TCM corpora, DAPS improves Jaccard in all 36 matched backbone–dataset evaluations, with average gains of 5.77 and 6.24 percentage points on the EHR and TCM corpora, respectively. Analyses characterize the sources and robustness of these gains, while highlighting the distinction between observational dependencies and external conflict knowledge.

1 INTRODUCTION

Medication recommendation predicts a prescription set from diagnoses, procedures, and longitudinal medication history in EHRs (Kim et al., 2024). Western methods encode visit sequences and drug graphs (Shang et al., 2019; Yang et al., 2021a), add molecular and safety priors (Yang et al., 2021b; 2023), or model health-status evolution (Kim et al., 2025). Traditional Chinese medicine (TCM) herb recommendation instead rests on syndrome differentiation and herb vocabularies, yet has the same form: symptoms, demographics, and history map to an unordered herb set. Its methods use symptom–herb and knowledge graphs (Jin et al., 2020; Yang et al., 2022; Yang & Ding, 2023), condition evolution (Liu et al., 2023), prescription knowledge (Dong et al., 2024; Zheng et al., 2025), or language-model fine-tuning (Zhou et al., 2024). Both literatures mainly improve candidate representations, leaving implicit how scores become a complete prescription set, including its cardinality.

Representative methods already incorporate rich relational information: GAMENet combines longitudinal memory with medication graphs (Shang et al., 2019); SafeDrug and MoleRec exploit molecular structure (Yang et al., 2021b; 2023); and PresRecST integrates syndrome differentiation with treatment planning (Dong et al., 2024). Generative approaches such as COGNet also condition prescription generation on previously selected medications (Wu et al., 2022). Thus, the missing ingredient is not dependence modeling itself. Our focus is the interface between candidate scores and the variable-size set used for evaluation (Dembczyński et al., 2012). For score-based backbones, a ranking leaves open how many candidates to retain and how to incorporate external conflict knowledge (curated lists of pairs not to be co-prescribed, e.g., drug–drug interactions or TCM herb incompatibilities). We ask whether explicit decoding improves these set-level decisions without encoder retraining. Appendix B reviews medication and herb recommendation approaches in detail.

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.

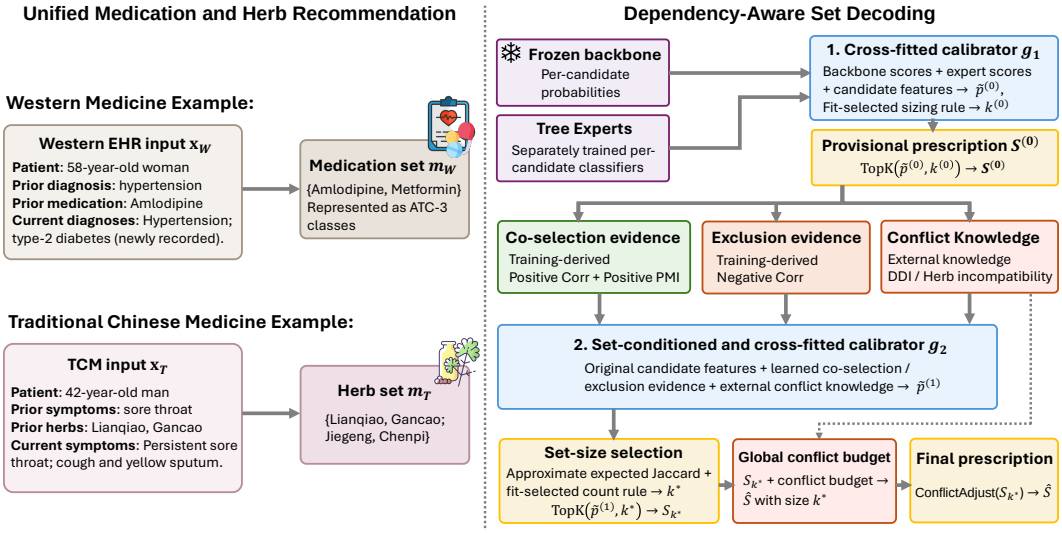


Figure 1: **Unified medication and herb recommendation with DAPS. Left**: Western EHR and TCM examples map pre-prescription information to medication and herb sets, respectively. **Right**: DAPS combines frozen backbone scores with tree-expert scores through cross-fitted calibration, refines a provisional set using co-selection, exclusion, and external-conflict channels, and implements approximate expected-Jaccard set sizing and the validation-selected global external-conflict budget.

Individually plausible candidates need not form the best predicted set. Consider a stylized case in which recorded prescriptions contain either $\{a, c\}$ or $\{b, c\}$: both a and b can receive appreciable marginal scores even though they are never selected together. Conversely, a moderately ranked candidate may frequently accompany members of a provisional set. The decision is therefore which candidates to retain together, not merely which ones rank highest. Such patterns motivate refining scores using candidate-to-set context, while selecting size explicitly to balance missed and extra members under Jaccard evaluation. Our analysis formalizes the underlying gap: a presentation need not determine one realized prescription, and even exact marginals need not identify the Jaccard-optimal set. Recorded co-selection and exclusion can help predict prescribing behavior, but do not establish pharmacological synergy or conflict. In the example, a and b could reflect different prescriber habits rather than incompatible treatments. We therefore separate observational signals from external conflict knowledge, using the former as context and the latter to constrain decoding.

These observations motivate *Dependency-Aware Prescription Selection* (DAPS), which adds an explicit, approximate set decision to any frozen probabilistic backbone that exports aligned candidate scores (Figure 1). Two cross-fitted calibrators combine backbone and tree-expert scores and then refine them using typed provisional-set summaries. These separate empirical co-selection and exclusion from the externally supplied conflict relations. Approximate expected Jaccard selects prescription size, and a validation-selected global external-conflict budget constrains final membership without updating the backbone. The budget controls an aggregate conflict rate against the supplied resource, not patient-specific clinical safety. This interface preserves domain-specific encoders, vocabularies, and conflict resources, while sharing only the decoding process, not clinical semantics.

Contributions. First, we unify medication and herb recommendation as variable-size set decoding process, characterize prescription uniqueness, and show why marginals alone cannot determine a general Jaccard-optimal decision. Second, we propose DAPS, a backbone-agnostic decoder combining cross-fitted calibration, typed dependency-aware refinement, approximate expected-Jaccard sizing, and a global external-conflict budget, while separating corpus association from external conflict knowledge. Third, we evaluate this interface on two EHR corpora and two TCM corpora, improving Jaccard in all 36 matched backbone–dataset pairs, by 5.77 and 6.24 Jaccard percentage points on average on the EHR and TCM corpora. Ablations, ranking and set-size analyses, and a held-out semantic audit examine the contributions of calibration, relation refinement, and cardinality selection to these gains. DAPS also exceeds the LLM plug-in GiantMed on every shared backbone.

2 Problem Formulation and Theoretical Analysis

Unified problem formulation. A case is $(\mathbf{x}, \mathbf{m})$, where the *presentation* $\mathbf{x} = (x_1, \ldots, x_N) \in \mathcal{X}$ contains all information recorded before prescribing, and the *prescription* $\mathbf{m} \in \{0,1\}^K$ is a multi-hot indicator over a vocabulary of K candidate medications. In Western EHRs, $\mathbf{x}$ consists of recorded diagnoses, procedures, and available medication history, while $\mathbf{m}$ indexes level-3 Anatomical Therapeutic Chemical (ATC) drug classes; in TCM, $\mathbf{x}$ consists of symptoms, demographics, and available history, while $\mathbf{m}$ indexes herbs. Both tasks predict a variable-size candidate set conditioned on the presentation, and both are evaluated by set overlap with the recorded prescription. Medication membership, rather than dose, timing, or route of administration, is the target, so a prescription and a prediction are compared as unordered sets. In the probabilistic analysis below, X denotes the random presentation, M denotes the random prescription set, and x denotes a realization of X. In the encoded case $(\mathbf{x}, \mathbf{m})$, M is realized as $\{j : \mathbf{m}_j = 1\}$, the support of $\mathbf{m}$. Two candidates in one prescription may substitute for each other, one may serve as an adjunct to the other, or the pair may be listed as conflicting by an external resource such as a drug–drug interaction or herb-incompatibility list. DAPS makes no assumption that presentations decompose into identifiable latent treatment demands. DAPS uses empirical co-selection and exclusion solely as predictive context, keeping externally maintained conflict knowledge distinct from these observational co-prescription patterns.

Why marginal probabilities are not yet a prescription decision. Recall that X is the random presentation and $M \subseteq [K]$ the random recorded prescription set. A frozen backbone estimates the candidate marginals $\eta_j(x) = P(j \in M \mid X = x)$, collected in the marginal vector $\eta(x) = (\eta_1(x), \ldots, \eta_K(x))$, whereas evaluation compares one predicted set S with the complete prescription. Appendix C instantiates this notation with one Western EHR and one TCM example. Proposition 1 states two reasons why candidate-wise probabilities, even exact ones, do not by themselves specify the final prescription decision for any single case; its proof is in Appendix C.1.

Proposition 1 (Limits of prescription prediction)**.** *Let $J(S, M) = |S \cap M|/|S \cup M|$ denote the Jaccard index, where $|\cdot|$ is set cardinality, with $J(\varnothing, \varnothing) = 1$.*

(i) (Prescription non-uniqueness). A presentation x does not uniquely determine the realized prescription whenever the conditional law $P(M \mid X = x)$ is non-degenerate, i.e., more than one prescription set has positive conditional probability, for example when two candidates are interchangeable given the rest of the set. Then even a predictor that knows the full conditional law cannot recover the realized prescription with probability one: the minimum exact-set error over deterministic f satisfies

$$R^*_{\text{exact}}(x) := \inf_f P(M \neq f(x) \mid X = x) = 1 - \max_{S \subseteq [K]} P(M = S \mid X = x) > 0. \tag{1}$$

(ii) (Marginal insufficiency). For $K \geq 3$, the marginal vector $\eta(x)$ does not in general determine the Jaccard-optimal prescription: there exist two conditional prescription laws p and p' on $2^{[K]}$ with identical candidate-wise marginals but different unique Jaccard-optimal sets,

$$\eta(p) = \eta(p'), \qquad S^*_J(p) \neq S^*_J(p'), \tag{2}$$

where $\eta(p) = (p(j \in M))_{j \in [K]}$ and $S^*_J(p) \in \arg\max_{S \subseteq [K]} \mathbb{E}_p[J(S, M)]$.

Remark 1 (Exact recovery and set-optimal prediction). By part (i), a mismatch with the recorded prescription need not indicate clinically incorrect treatment. By part (ii), identical candidate probabilities can yield different Jaccard-optimal sets because candidates co-occur differently.

Remark 2 (Dependence is predictive context, not pharmacological semantics). Proposition 1(ii) motivates information beyond candidate-wise marginals but assigns no clinical meaning to an observed association. Positive correlation or pointwise mutual information is thus treated only as empirical *co-selection*, and negative correlation only as empirical *exclusion*. Co-selection may be consistent with synergy, and exclusion with substitution or conflict, but these patterns may also reflect shared indications or prescriber preferences. DAPS keeps these channels separate from external drug–drug interaction (DDI) or TCM incompatibility resources, which alone define the global conflict budget.

Proposition 1(ii) needs dependence: under conditional independence, marginals fix the joint law and thus the Jaccard-optimal set. Theorem 1 derives this set and bounds the cost of estimated marginals.

Theorem 1 (Jaccard decisions under conditional independence). *Fix x, let $\eta = \eta(x)$ be its marginal vector, and suppose the indicators $Y_j = \mathbf{1}\{j \in M\}$, $j \in [K]$, are independent given $X = x$. For $p \in [0,1]^K$, let $U_p(S)$ be the expected Jaccard $\mathbb{E}[J(S,M)]$ when $Y_1, \ldots, Y_K$ are independent Bernoulli with means $p_1, \ldots, p_K$; thus $U_\eta(S) = \mathbb{E}[J(S,M) \mid X = x]$. Let $S_k(p)$ be the k candidates with the largest entries of p, ties broken arbitrarily, with $S_0(p) = \varnothing$. Then:*

(i) (Top-k structure). $\max_{S \subseteq [K]} U_\eta(S) = \max_{0 \le k \le K} U_\eta(S_k(\eta))$, so $S_{k^\star}(\eta)$ is Jaccard-optimal for every $k^\star \in \arg\max_k U_\eta(S_k(\eta))$. For $k \ge 1$, $U_\eta(S_k(\eta)) = \mathbb{E}[A_k]\,\mathbb{E}[1/(k + B_k)]$, where $A_k = |S_k(\eta) \cap M|$ and $B_k = |M \setminus S_k(\eta)|$ are independent Poisson–binomial counts.

(ii) (Plug-in regret). For any $\hat{p} \in [0,1]^K$, let $\widehat{S} = S_{\hat{k}}(\hat{p})$ with $\hat{k} \in \arg\max_k U_{\hat{p}}(S_k(\hat{p}))$. Then $\max_{S \subseteq [K]} U_\eta(S) - U_\eta(\widehat{S}) \le 2\|\hat{p} - \eta\|_1$.

Remark 3 (What the decision rule requires). Part (i) reduces the decision to one size per case: rank by probability, then maximize the Poisson–binomial expectation; a fixed size is optimal only when $k^\star$ is constant across cases. Part (ii) bounds the regret by the ℓ_1 error of the marginals, not by ranking quality alone. Combined with Pinsker's inequality (Appendix C.2), it bounds the expected regret under conditional independence by $\sqrt{2}\sum_j \sqrt{\Delta_j}$, where Δ_j is the excess binary log loss of $\hat{p}_j$; log-loss training thus controls it. Part (i) is the Jaccard counterpart of the F-measure probability-ranking result (Lewis, 1995; Waegeman et al., 2014). Under dependence the rule can pick the wrong size: in the example proving Proposition 1(ii), it loses $1/60$ in expected Jaccard under one of them.

These two results divide the decoder's task. Theorem 1 states what to do with accurate marginals, and Proposition 1(ii) states what marginals leave out. Proposition 2 in Appendix C.3 quantifies that departure without any independence assumption: it writes expected Jaccard as a second-order expansion plus an explicitly bounded remainder. The leading term depends only on the candidate-wise marginals; at second order, dependence enters through the covariance between selected and missed members and the variance of the missed count, both calculated relative to the proposed set.

3 A Unified Dependency-Aware Set Decoder

This section details DAPS (Figure 1), which turns candidate probabilities into a set decision. Its design follows Section 2: steps 1 and 2 estimate calibrated marginals and size sets with a fit-selected rule built on Theorem 1(i); step 2 also adds set context for the dependence that Proposition 1(ii) shows marginals miss; and step 3 is a conflict constraint outside both results. Each Western or TCM backbone keeps its representation, vocabulary, and training, then stays frozen. For presentation $\mathbf{x}$, let $p^B(\mathbf{x})$ and $p^E(\mathbf{x})$ be the backbone and auxiliary-expert probabilities. The pipeline is

$$\begin{aligned} \tilde{p}^{(0)} &= g_1\big(F(p^B, p^E, \mathbf{x})\big), & S^{(0)} &= \mathrm{TopK}\Big(\tilde{p}^{(0)}, k^{(0)}\Big), \\ \tilde{p}^{(1)} &= g_2\Big(F, C(S^{(0)}), D(S^{(0)})\Big), & S_{k^\star} &= \mathrm{TopK}\Big(\tilde{p}^{(1)}, k^\star\Big), \qquad \widehat{S} = \mathcal{B}_{\mathrm{conf}}(S_{k^\star}). \end{aligned} \tag{3}$$

Here $F = F(p^B, p^E, \mathbf{x})$ collects candidate features; calibrators g_1, g_2 produce initial and refined probabilities $\tilde{p}^{(0)}, \tilde{p}^{(1)}$. $\mathrm{TopK}(p,k)$ selects the top k candidates: $k^{(0)}$ defines the provisional set $S^{(0)}$, and $k^\star$ the refined set $S_{k^\star}$. $C(S^{(0)})$ collects typed candidate-to-set context, $D(S^{(0)})$ set-completion scores, and $\mathcal{B}_{\mathrm{conf}}$ is a size-preserving global conflict adjustment yielding the final prescription $\widehat{S}$. Dependence on $\mathbf{x}$ is suppressed. Sizes are case-specific; their rules and the conflict budget are selected on the validation (*fit*) partition. DAPS thus runs in three steps: (1) calibration and provisional sizing give $\tilde{p}^{(0)}$ and $S^{(0)}$; (2) set-conditioned refinement and final sizing give $\tilde{p}^{(1)}$ and $S_{k^\star}$; and (3) a global conflict adjustment gives $\widehat{S}$. Appendix D specifies implementation details and the data splits.

Step 1: calibration and provisional sizing. For candidate $j \in \{1, \ldots, K\}$, a separate gradient-boosted-tree expert trained on the backbone's training partition supplies $p_j^E(\mathbf{x})$. Features F_j combine p_j^B, p_j^E, their logits, normalized ranks, score margins, and pre-prescription context. Both calibrators are fit by binary log loss; under the working model of Theorem 1, excess log loss bounds the regret of the plug-in rule in part (ii) via Pinsker's inequality (Appendix C.2). Cross-fitting on the fit partition, with folds disjoint in evaluation units, keeps selection predictions out of fold. The first produces $\tilde{p}_j^{(0)} = g_1(F_j)$, and the fit-selected count search used for the final size in step 2, applied to $\tilde{p}^{(0)}$, yields the provisional size $k^{(0)}$. This provisional set $S^{(0)}$ solely furnishes the relation context.

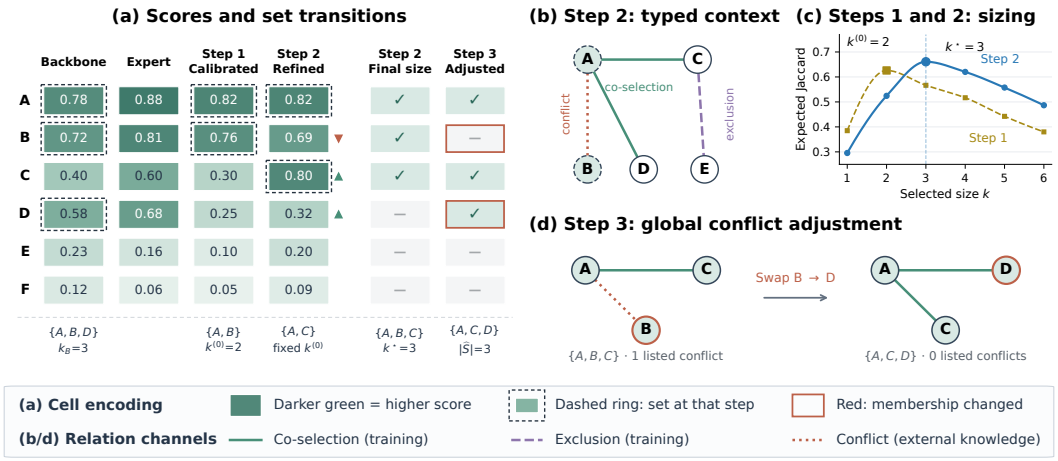


Figure 2: **Synthetic decoding example.** (a) Scores and membership across the three steps, named as in Table 4; step 2 is shown after refinement at the provisional size and after final resizing. (b) Relations inform step-2 refinement without hard constraints. (c) Independent expected Jaccard sets both sizes: the provisional $k^{(0)} = 2$ at step 1 from calibrated scores, and the final $k^\star = 3$ at step 2 from refined scores. (d) In step 3, a budget-triggered $B \to D$ swap removes a listed conflict at a utility cost, preserving size. Toy numbers illustrate decoding, not empirical effects or clinical safety.

Step 2: set-conditioned refinement and final sizing. Following Remark 2, training-prescription associations define co-selection and exclusion channels, not pharmacological labels:

$$R^{\mathrm{co}} = \{[\mathrm{Corr}(\mathbf{m})]_+, [\mathrm{PMI}(\mathbf{m})]_+\}, \qquad R^{\mathrm{excl}} = [-\mathrm{Corr}(\mathbf{m})]_+. \tag{4}$$

Here Corr and PMI denote $K \times K$ empirical correlation and pointwise mutual information matrices of candidate-presence indicators on training data; $[z]_+ = \max(z, 0)$ acts entrywise. External DDI or TCM incompatibility knowledge supplies the conflict channel R^{conf}; Western ATC-1/ATC-3 membership supplies the redundancy channel R^{red}. Summarizing these matrices against $S^{(0)}$ gives $C_j^r(S^{(0)})$, where $r \in \{\mathrm{co}, \mathrm{excl}, \mathrm{red}, \mathrm{conf}\}$ identifies the relation type; these features collectively form $C(S^{(0)})$, retaining the distinction between training-derived associations, taxonomy-based redundancy, and conflict knowledge from external resources for candidate rescoring (Appendix D.2).

Score refinement. Proposition 1(ii) shows that marginals miss dependence among candidates. DAPS represents this dependence relative to a proposed set through training-derived relation summaries and a supervised score refiner. These summaries serve as predictive features rather than estimates of case-conditional covariances:

$$\tilde{p}_j^{(1)} = g_2\Big(F_j, C_j^{\mathrm{co}}(S^{(0)}), C_j^{\mathrm{excl}}(S^{(0)}), C_j^{\mathrm{red}}(S^{(0)}), C_j^{\mathrm{conf}}(S^{(0)}), D_j(S^{(0)})\Big). \tag{5}$$

Besides these corpus-level summaries, $D_j(S^{(0)})$ holds the outputs of two set-completion models fitted on the training partition, which score candidate j given the presentation and the remainder of $S^{(0)}$; the typed relations enter one of them with fixed signs and case-dependent strengths (Appendix D.2). g_2 is boosted from $\mathrm{logit}\,\tilde{p}^{(0)}$, and the round is repeated once on the refined set. The redundancy channel is Western-only (fixed to zero for TCM). Figure 2 illustrates all three steps.

Loss-aware final size. For an admissible size $1 \le k \le 45$, let $S_k = \mathrm{TopK}(\tilde{p}^{(1)}, k)$; recorded prescriptions are nonempty, so omitting $k = 0$ loses nothing. Following Theorem 1(i), the base size rule maximizes estimated expected Jaccard against the random prescription set M:

$$k^{\mathrm{PB}} = \arg\max_k \widehat{\mathbb{E}}_{\mathrm{PB}}\left[\frac{|S_k \cap M|}{|S_k \cup M|}\right]. \tag{6}$$

Here $\widehat{\mathbb{E}}_{\mathrm{PB}}$ is a Poisson–binomial (PB) expectation using the refined probabilities as independent Bernoulli parameters, computed exactly by a recursion. Because the same probabilities rank and

Table 1: **Processed datasets used in the experiments.** The patient/case column counts MIMIC patients and independent TCM cases; input vocabulary combines diagnoses and procedures for MIMIC but symptoms for TCM. Averages report input-code and medication/herb set sizes per prescription.

Category	Corpus	Patients/ cases	Prescriptions	Input vocab	Output vocab	Avg. input size	Avg. output size
EHR	MIMIC-III (Johnson et al., 2016)	6,350	15,031	3,312	131	17.83	19.20
	MIMIC-IV (Johnson et al., 2023)	50,647	119,210	4,419	127	14.18	13.93
TCM	TCM-Onsite (Liu, 2024)	92,793	92,793	2,917	329	10.47	14.73
	TCM-Internet (Liu, 2024)	306,303	306,303	11,976	720	25.54	12.33

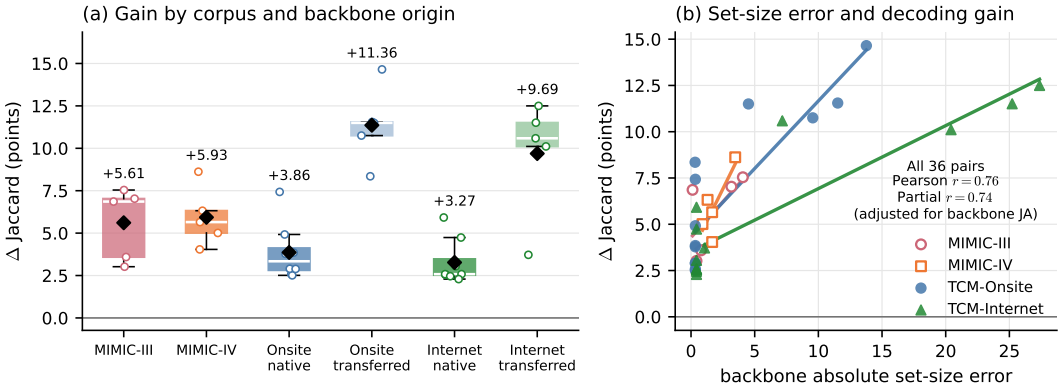


Figure 3: **Main comparison: DAPS gains and initial set-size errors.** Pairs are from Tables 2 and 3; color denotes corpus. (a) Jaccard gains by backbone group; diamonds mark the main-table mean gains, and pale boxes indicate transferred backbones. (b) Gains versus absolute mean set-size error, with descriptive within-corpus least-squares fits. The annotations report pooled correlations before and after adjusting for backbone Jaccard; the plotted coordinates are not regression residuals.

size the set, equation 6 is the plug-in rule of Theorem 1(ii) with $\hat{p} = \tilde{p}^{(1)}$; independence is a working model for sizing, not an assumption about the true prescription law. The executed size $k^\star$ adjusts this base rule: a fit-selected weight blends k^{PB} with the backbone count, and a scale and an offset follow. A soft-Jaccard source may replace the PB maximizer; the fit split chooses PB in 28 of the 36 matched evaluations. These adjustments absorb any residual miscalibration and dependence and carry no separate guarantee. They are all fixed on the fit partition before any testing (Appendix D.2).

Step 3: global external-conflict budget. Using only the external R^{conf}, this resource-based step ranks size-preserving within-prescription swaps by predicted-utility loss per unit of conflict removed and applies them globally, across all prescriptions, to meet the fit-selected pooled conflict budget, yielding $\widehat{S}$. This aggregate resource-based constraint neither bounds each prescription's conflict rate nor certifies patient-specific safety, as the budget controls the pooled rate of listed pairwise conflicts.

4 Experiments

Experimental Setup. We evaluate medication and herb recommendation on MIMIC-III/IV and two public TCM corpora from onsite and internet-hospital records (Liu, 2024); Table 1 summarizes the processed data. The standard split allocates $2/3$ to training, $1/6$ to validation (the fit partition), and $1/6$ to testing, by patient for MIMIC and by case for TCM, whose releases lack patient and practitioner identifiers. Matched decoder comparisons hold each backbone and evaluation cohort fixed. Experts and correlation/PMI channels use training data; calibration and decoder selection use the fit partition. We report Jaccard (JA), F_1, ranking PRAUC, listed conflict rates (DDI for MIMIC; Inc. for TCM), and mean predicted set size. Scores and conflict rates are percentages; set sizes are counts. Appendix E specifies metric aggregation, cohort filters, and protocol exceptions. Language-model evaluations use sampled cohorts and provide separate context in Appendix E.3, and the joint-training comparison trains backbone and decoder together. Main-table backbones are all evaluated locally; Appendix E marks the author-code, adapted, and paper-guided implementations.

Table 2: **Main comparison: medication recommendation on MIMIC-III/IV.** Five locally evaluated frozen backbones are each paired with GiantMed and DAPS on the same cohorts. Higher JA/F_1/PRAUC and lower DDI are better; # Med. denotes mean set size; bold marks the best values.

	MIMIC-III (Johnson et al., 2016)					MIMIC-IV (Johnson et al., 2023)				
Method	JA ↑	F_1 ↑	PRAUC ↑	DDI ↓	# Med.	JA ↑	F_1 ↑	PRAUC ↑	DDI ↓	# Med.
Medication recommenders evaluated locally, each frozen and paired with GiantMed (Lv et al., 2026) and DAPS										
SafeDrug (Yang et al., 2021b)	50.02	65.85	75.80	6.86	19.91	48.90	64.12	73.01	6.56	15.23
+ GiantMed	51.27	66.94	75.71	**5.66**	19.75	48.84	64.08	72.92	**6.27**	15.45
+ DAPS	56.88	71.70	81.64	6.67	20.72	55.22	69.68	79.69	6.51	14.66
DEPOT (Sun et al., 2022)	53.34	68.70	77.98	7.24	20.58	50.24	65.38	74.65	6.53	14.83
+ GiantMed	53.14	68.53	77.65	6.95	20.91	50.02	65.20	74.40	6.62	15.35
+ DAPS	56.93	71.75	81.69	7.07	20.97	55.25	69.71	79.75	6.52	14.62
MoleRec (Yang et al., 2023)	49.40	65.31	73.02	6.98	23.86	49.95	65.13	73.20	8.37	15.59
+ GiantMed	49.55	65.47	72.96	6.77	24.16	49.94	65.12	72.93	8.17	15.78
+ DAPS	56.94	71.74	81.68	6.85	20.79	55.60	70.00	79.69	8.25	14.67
VITA (Kim et al., 2024)	51.33	66.65	70.39	7.43	22.96	49.81	64.72	64.94	8.30	17.40
+ GiantMed	47.89	63.36	75.26	8.13	15.95	47.73	62.70	71.15	7.86	14.19
+ DAPS	**58.36**	**72.98**	**82.80**	7.40	21.07	**58.43**	**72.48**	**81.74**	8.06	15.34
MedAlign (Lv et al., 2025)	53.89	69.18	78.43	7.03	20.22	51.20	66.24	75.17	6.59	15.59
+ GiantMed	53.83	69.15	78.27	6.88	20.69	50.80	65.90	74.88	6.55	16.08
+ DAPS	56.91	71.72	81.66	6.93	20.98	55.24	69.70	79.71	6.53	14.67
Mean change over the five matched backbones										
Δ + GiantMed	−0.46	−0.45	+0.85	−0.23	−1.21	−0.55	−0.52	+1.06	−0.18	−0.36
Δ + DAPS	+5.61	+4.84	+6.77	−0.12	−0.60	+5.93	+5.20	+7.92	−0.10	−0.94

Main Comparison on Western EHRs. Table 2 compares five backbones, each paired with GiantMed and DAPS. Language-model results on separate sampled cohorts are in Appendix Table 19. In the matched decoder comparisons, DAPS improves Jaccard for all five backbones on both corpora. MedAlign rises from 53.89 to 56.91 on MIMIC-III and from 51.20 to 55.24 on MIMIC-IV. Averaged over three seeds, its MIMIC-III budget stage reduces DDI from 8.02 to 6.93 while reducing Jaccard from 57.12 to 56.98 (Tables 4 and 9), illustrating the overlap–conflict tradeoff. GiantMed reaches 53.83 on the identical MedAlign predictions; Appendix E.2 explains the differences from the published 56.63. Appendix E.4 provides further analysis. Published results of six earlier methods (Appendix Table 8) give literature context, not evaluations under our matched splits and protocols.

Main Comparison on TCM Corpora. Table 3 evaluates eight TCM-native rankers and five Western architectures trained on TCM data; sampled language-model evaluations appear in Appendix Table 20. In matched comparisons, DAPS improves Jaccard for every backbone on both corpora. Native rankers output scores without a set-size rule, and thresholding them at 0.5 under-predicts and lowers Jaccard for all eight on both corpora (Appendix E.2); we therefore decode them at the rounded training-split mean size, 15/12 herbs (corpus means 14.73/12.33). Their mean gains are 3.86/3.27 Jaccard points on TCM-Onsite/TCM-Internet. The adapted group gains 11.36/9.69 points on average and has a wider range of initial output sizes, from 5.26 to 39.80 herbs on TCM-Internet. Figure 3(a) summarizes these group differences; they do not imply that every adapted backbone gains more than every native backbone. Appendix E.4 details the comparisons. The results support one decoder interface across domains, motivating the analysis of initial set-size error.

Ablation Study on Decoder Components. Table 4 splits the decoder of Figure 2 into five cumulative sub-steps, fixing each dataset's backbone and split and averaging three seeds of the complete decoder; the gains quoted below use the two-decimal means in Appendix Table 9. Steps 1 and 2 each separate *calibration* from *sizing*: sub-step (a) changes only the scores and keeps the set size fixed (the backbone's own size for 1a, the provisional $k^{(0)}$ for 2a), so its gain reflects which candidates are ranked highest; sub-step (b) keeps those scores and changes only the size ($k^{(0)}$ for 1b, $k^\star$ for 2b), so its gain reflects how many are selected. Set-conditioned calibration (2a) is the largest single gain on TCM-Onsite (+1.14 Jaccard points), exceeding both 1a (+0.90) and 1b (+0.51), whereas independent calibration (1a) is the largest on MIMIC (+2.27 and +2.73). Final sizing (2b) adds

Table 3: **Main comparison: herb recommendation on TCM-Onsite/Internet.** Groups follow Table 2 and are named in the table. Inc. is the percentage of predicted herb pairs the classical *shibafan shijiuwei* list flags as incompatible, and # H. the mean prescription size against full-corpus means 14.73/12.33. Further rows (language-model baselines) are in Table 20; backbone adaptations are documented in Appendix E. Bold marks the best values within each named group and column.

	TCM-Onsite (Liu, 2024)					TCM-Internet (Liu, 2024)				
Method	JA ↑	F_1 ↑	PRAUC ↑	Inc. ↓	# H.	JA ↑	F_1 ↑	PRAUC ↑	Inc. ↓	# H.
Herb recommenders built for TCM, each frozen and paired with DAPS										
SMGCN (Jin et al., 2020)	44.42	59.51	65.49	1.44	15.00	26.58	39.84	42.74	0.00	12.00
+ DAPS	51.85	66.21	73.38	1.36	16.12	32.50	46.54	50.72	0.00	15.31
KDHR (Yang et al., 2022)	47.95	62.85	69.81	1.50	15.00	29.45	43.21	47.43	0.00	12.00
+ DAPS	51.80	66.16	73.36	1.37	15.94	32.50	46.54	50.72	0.00	15.26
SCEIKG (Liu et al., 2023)	48.59	63.37	70.46	0.97	15.00	29.89	43.74	48.24	0.00	12.00
+ DAPS	51.17	65.61	73.09	**0.96**	16.00	32.47	46.49	50.70	0.00	15.21
SMRGAT (Yang & Ding, 2023)	49.10	63.77	70.71	1.36	15.00	29.94	43.79	48.39	0.00	12.00
+ DAPS	51.61	66.00	73.21	1.36	15.96	32.49	46.52	50.69	0.00	15.21
PresRecST (Dong et al., 2024)	49.05	63.71	70.69	1.35	15.00	29.91	43.76	48.24	0.00	12.00
+ DAPS	**51.95**	**66.29**	**73.45**	1.32	15.90	32.49	46.51	50.71	0.00	15.18
TCMRGCL (Hu et al., 2025)	46.34	61.63	68.69	1.51	15.00	27.75	41.28	44.78	0.00	12.00
+ DAPS	51.26	65.74	72.92	1.34	16.25	32.49	46.52	50.70	0.00	15.36
PRDAGE (Wen et al., 2025)	47.57	62.56	69.71	1.48	15.00	30.30	44.08	48.91	0.00	12.00
+ DAPS	51.35	65.81	73.06	1.37	15.98	**32.59**	**46.63**	**50.89**	0.00	15.18
FMASH (Zheng et al., 2025)	48.70	63.51	70.64	1.41	15.00	30.02	43.88	48.41	0.00	12.00
+ DAPS	51.59	65.99	73.17	1.36	15.92	32.48	46.51	50.71	0.00	15.22
Mean change over the eight matched backbones										
Δ + DAPS	+3.86	+3.36	+3.68	−0.07	+1.01	+3.27	+3.58	+3.59	+0.00	+3.24
Western medication recommenders transferred to TCM, each frozen and paired with DAPS										
SafeDrug (Yang et al., 2021b)	41.60	57.69	68.07	1.09	26.18	22.32	35.62	46.35	0.09	37.65
+ DAPS	53.15	67.28	**74.69**	1.08	15.92	33.83	47.65	51.92	0.00	15.05
DEPOT (Sun et al., 2022)	42.44	58.38	67.56	1.15	24.22	23.69	37.19	45.85	0.14	32.85
+ DAPS	53.19	67.32	74.66	1.13	16.18	33.80	47.60	51.88	0.00	15.03
MoleRec (Yang et al., 2023)	38.26	54.45	63.45	0.90	28.43	21.30	34.29	44.30	0.21	39.80
+ DAPS	52.91	67.07	74.63	**0.89**	15.98	33.80	47.60	51.88	0.00	15.01
VITA (Kim et al., 2024)	14.34	21.11	25.29	1.23	14.98	43.60	54.60	45.56	0.01	13.49
+ DAPS	22.69	29.90	33.08	1.22	5.17	**47.32**	**58.33**	**62.85**	0.01	13.64
MedAlign (Lv et al., 2025)	41.74	56.88	68.78	1.81	10.15	23.25	34.53	49.03	0.01	5.26
+ DAPS	**53.24**	**67.37**	74.64	1.30	16.19	33.84	47.66	51.91	0.01	14.95
Mean change over the five transferred backbones										
Δ + DAPS	+11.36	+10.08	+7.71	−0.11	−6.90	+9.69	+10.52	+7.87	−0.00	−11.07

+0.07 to +0.37 on every dataset. Overall, step 2 adds +1.51, +1.13, +0.30, and +0.82 points after step 1's +1.41, +1.45, +2.93, and +3.66 (TCM-Onsite, TCM-Internet, MIMIC-III, MIMIC-IV); every seed standard deviation is at most 0.11. Step 3 lowers the listed conflict rate relative to step 2 on all four datasets, at no Jaccard cost on TCM and 0.14 and 0.41 points on MIMIC. Appendix Table 9 reports full set metrics. Count-policy and budget sensitivity appear in Appendix E.7 and Figures 5–6. Across six calibrator families, Jaccard varies by at most 0.48 points on TCM-Onsite and 0.34 on MIMIC-III, so the gain does not rely on any single learner (Appendix E.6 and Table 15).

Training Strategy: Post-hoc Decoding versus Joint Optimization. We compare frozen-backbone post-hoc DAPS with a joint model trained from scratch on MIMIC-III/IV. The joint model uses a trainable auxiliary head, soft provisional-set context, and a learned count predictor in place of the tree expert and discrete post-hoc decisions. Figure 4 shows higher JA and F_1 and lower DDI for post-hoc DAPS in both comparisons: JA is 56.91 versus 52.97 on MIMIC-III and 55.24 versus 52.32 on MIMIC-IV. Joint training falls below native MedAlign decoding on MIMIC-III (53.89) but exceeds it on MIMIC-IV (51.20). The two pipelines differ in architecture, objective, cardinality rule, and conflict control, and the joint model lacks the training-partition set-completion models used in step 2, so this is not a fully controlled ablation of the execution order (Appendix E.6 and Table 14).

Table 4: **Ablation study: five cumulative decoding sub-steps.** PresRecST (TCM) and MedAlign (MIMIC) are frozen. Each step separates calibration (a) from sizing (b). Test-split percentages over three decoder seeds; JA and F_1 means have one decimal and show their s.d. (two-decimal means and all s.d. are in Table 9); the frozen backbone row has no seed variation. Lower DDI or Inc. is better.

	TCM-Onsite			TCM-Internet			MIMIC-III			MIMIC-IV		
Configuration	JA ↑	F_1 ↑	Inc. ↓	JA ↑	F_1 ↑	Inc. ↓	JA ↑	F_1 ↑	DDI ↓	JA ↑	F_1 ↑	DDI ↓
Backbone	49.1±.00	63.7±.00	1.35	29.9±.00	43.8±.00	0.0025	53.9±.00	69.2±.00	7.03	51.2±.00	66.2±.00	6.59
+ 1a	50.0±.04	64.7±.03	1.49	30.4±.02	44.3±.02	0.0038	56.2±.02	71.1±.02	7.86	53.9±.01	68.6±.01	8.14
+ 1a–1b	50.5±.02	65.2±.02	1.38	31.4±.02	45.5±.02	0.0039	56.8±.02	71.7±.02	7.95	54.9±.01	69.4±.01	8.56
+ 1a–2a	51.6±.02	66.1±.00	1.33	32.1±.02	46.2±.02	0.0041	57.1±.06	71.8±.05	8.00	55.3±.01	69.8±.01	8.60
+ 1a–2b	52.0±.02	66.3±.02	1.36	32.5±.02	46.5±.03	0.0044	57.1±.11	71.9±.10	8.02	55.7±.02	70.1±.02	8.63
+ 1a–3	52.0±.02	66.3±.02	1.32	32.5±.02	46.5±.03	0.0027	57.0±.10	71.8±.09	6.93	55.3±.03	69.7±.03	6.53

Each row adds one sub-step to all rows above it (1a–2b applies 1a, 1b, 2a, 2b); (a) changes only scores at a fixed set size, (b) only the size. **1a**: cross-fitted calibration g_1 at the backbone's set size. **1b**: provisional size $k^{(0)}$ from $\tilde{p}^{(0)}$. **2a**: set-conditioned recalibration g_2 given $S^{(0)}$ at the fixed size $k^{(0)}$. **2b**: final size $k^\star$ from $\tilde{p}^{(1)}$. **3**: global conflict adjustment; 1a–3 is full DAPS. TCM-Internet Inc. has four decimal places.

Analysis of Set Size and Ranking Quality. Across the 36 main-table backbone–dataset pairs, larger initial mean set-size errors are associated with larger Jaccard gains ($r = 0.76$, partial $r = 0.74$ given backbone Jaccard; Figure 3(b)). This descriptive analysis uses evaluation reference sizes and is not causal. Consistent with Theorem 1(i), under an equal validation budget a fit-selected fixed size trails PB expected-Jaccard sizing by 0.5–5.8 points on every dataset, while PB, soft Jaccard, and a tuned global threshold differ by at most 0.4 (Appendix Table 16). The four pairs in Appendix Table 17 also gain size-independent PRAUC, so ranking improvements contribute too (Appendix E.8).

Further Analyses. A held-out TCM audit distinguishes co-selection from pharmacological compatibility (Appendix E.8; Table 18). Appendix E.5 reports complete five-sub-step metrics, paired-bootstrap uncertainty for four representative pairs, and fitting/inference costs (Tables 9, 12, and 13). Appendix E.6 reports all seven joint-training metrics (Table 14); both calibrators are also replaced with LightGBM, XGBoost, random forests, Extra Trees, logistic regression, or MLP while retaining the expert bank (Table 15). Count-policy sensitivity and conflict-budget tradeoffs appear in Appendix E.7 (Figures 5 and 6). Appendix E.8 collects ranking and set-size analyses, the relation audit, 1,000-case language-model baselines, and matched GiantMed comparisons on TCM (Tables 17–21). Main comparisons are detailed in Appendix E.4; Appendix E (Sections E.1–E.3) covers datasets, baseline adaptations, reproduction caveats, language-model evaluation procedures, and the provenance of every number.

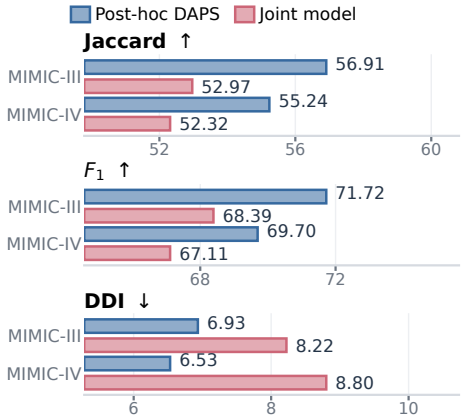


Figure 4: **Training-strategy comparison on MIMIC-III/IV.** Post-hoc DAPS (seed 1203) versus joint training; values are test percentages and axes are truncated. The joint model uses a learnable auxiliary head.

5 Conclusion

We unify medication and herb recommendation as variable-size set decoding over frozen candidate probabilities. DAPS combines cross-fitted calibration, typed dependency-aware refinement, approximate expected-Jaccard sizing, and a global external-conflict budget without retraining the backbone. Ablations attribute about half of the pre-adjustment gain on the TCM corpora, and 9–18% on the EHR corpora, to the set-conditioned step. Consistent Jaccard gains in all 36 matched backbone–dataset evaluations across four corpora support an explicit post-hoc selection stage, while the method remains a predictor of recorded prescribing behavior rather than a model that identifies treatment effects or establishes clinical benefit for individual patients in either domain. This modular decoder complements, rather than replaces, backbone innovations in Western medication and TCM herb recommendation, and it applies to every frozen backbone that outputs candidate probabilities.

REFERENCES

Anastasios Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International conference on learning representations*, 2024.

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4413–4421. IEEE, 2018.

Weiwei Cheng, Eyke Hüllermeier, and Krzysztof J Dembczynski. Bayes optimal multilabel classification via probabilistic classifier chains. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 279–286, 2010.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, 29, 2016.

Yan Ting Chok, Soyon Park, Seungheun Baek, Hajung Kim, Junhyun Lee, and Jaewoo Kang. Hiref: Leveraging hierarchical ontology and network refinement for robust medication recommendation. *arXiv preprint arXiv:2508.10425*, 2025.

Krzysztof Dembczyński, Willem Waegeman, Weiwei Cheng, and Eyke Hüllermeier. On label dependence and loss minimization in multi-label classification. *Machine Learning*, 88(1):5–45, 2012.

Krzysztof Dembczynski, Arkadiusz Jachnik, Wojciech Kotlowski, Willem Waegeman, and Eyke Hüllermeier. Optimizing the f-measure in multi-label classification: Plug-in rule approach versus structured loss minimization. In *International conference on machine learning*, pp. 1130–1138. PMLR, 2013.

Xin Dong, Chenxi Zhao, Xinpeng Song, Lei Zhang, Yu Liu, Jun Wu, Yiran Xu, Ning Xu, Jialing Liu, Haibin Yu, et al. Presrecst: a novel herbal prescription recommendation algorithm for real-world patients with integration of syndrome differentiation and treatment planning. *Journal of the American Medical Informatics Association*, 31(6):1268–1279, 2024.

Xuan Gao, Jiaying Sun, Jinjie Xu, and Weichao Xu. Prior-fused graph convolutional network for traditional chinese medicine formula recommendation. In *2026 6th International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA)*, pp. 1469–1475. IEEE, 2026.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Hailong Hu, Yaqian Li, Zeyu Zheng, Wenjun Hu, Riyang Lin, and Yanlei Kang. A traditional chinese medicine prescription recommendation model based on contrastive pre-training and hierarchical structure network. *Expert Systems with Applications*, 268:126318, 2025.

Yuanyuan Jin, Wei Zhang, Xiangnan He, Xinyu Wang, and Xiaoling Wang. Syndrome-aware herb recommendation with multi-graph convolution network. In *2020 IEEE 36th international conference on data engineering (ICDE)*, pp. 145–156. IEEE, 2020.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):160035, 2016.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Taeri Kim, Jiho Heo, Hongil Kim, Kijung Shin, and Sang-Wook Kim. Vita:'carefully chosen and weighted less' is better in medication recommendation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 8600–8607, 2024.

Taeri Kim, Jiho Heo, Hyunjoon Kim, and Sang-Wook Kim. Hi-dr: Exploiting health status-aware attention and an ehr graph+ for effective medication recommendation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 11950–11958, 2025.

Meelis Kull, Telmo Silva Filho, and Peter Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial intelligence and statistics*, pp. 623–631. PMLR, 2017.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems*, 32, 2019.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pp. 3744–3753. PMLR, 2019.

David D Lewis. Evaluating and optimizing autonomous text classification systems. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 246–254, 1995.

Zachary Chase Lipton, Charles Elkan, and Balakrishnan Narayanaswamy. Thresholding classifiers to maximize f1 score. *arXiv preprint arXiv:1402.1892*, 2014.

Jingjin Liu, Hankz Hankui Zhuo, Kebing Jin, Jiamin Yuan, Zhimin Yang, and Zhengan Yao. Sequential condition evolved interaction knowledge graph for traditional chinese medicine recommendation. *arXiv preprint arXiv:2305.17866*, 2023.

Sylvan Liu. Traditional Chinese medicine dataset (SFT). `https://huggingface.co/datasets/SylvanL/Traditional-Chinese-Medicine-Dataset-SFT`, 2024.

Hang Lv, Zixuan Guo, Zijie Wu, Yanchao Tan, Guofang Ma, Zhigang Lin, Xiping Chen, Hong Cheng, and Carl Yang. Medalign: Enhancing combinatorial medication recommendation with multi-modality alignment. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 6084–6092, 2025.

Hang Lv, Zixuan Guo, Yanchao Tan, Wanzi Shao, Hengyu Zhang, and Carl Yang. Awaken the giant: Activating llms via deep model guidance for boundary-aware medication recommendation. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 1030–1041, 2026.

Junyuan Shang, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun. Gamenet: Graph augmented memory networks for recommending medication combination. In *proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1126–1133, 2019.

Hongda Sun, Shufang Xie, Shuqi Li, Yuhan Chen, Ji-Rong Wen, and Rui Yan. Debiased, longitudinal and coordinated drug recommendation through multi-visit clinic records. *Advances in Neural Information Processing Systems*, 35:27837–27849, 2022.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Willem Waegeman, Krzysztof Dembczyński, Arkadiusz Jachnik, Weiwei Cheng, and Eyke Hüllermeier. On the bayes-optimality of f-measure maximizers. *Journal of Machine Learning Research*, pp. 3513–3568, 2014.

Xinyu Wang, Ying Zhang, Xiaoling Wang, and Jin Chen. A knowledge graph enhanced topic modeling approach for herb recommendation. In *International conference on database systems for advanced applications*, pp. 709–724. Springer, 2019.

Zhihua Wen, Yunchun Dong, Lihong Peng, Longxin Zhang, and Junfeng Yan. Prdage: a prescription recommendation framework for traditional chinese medicine based on data augmentation and multi-graph embedding. *PeerJ Computer Science*, 11:e2974, 2025.

Rui Wu, Zhaopeng Qiu, Jiacheng Jiang, Guilin Qi, and Xian Wu. Conditional generation net for medication recommendation. In *Proceedings of the ACM web conference 2022*, pp. 935–945, 2022.

Chaoqi Yang, Cao Xiao, Lucas Glass, and Jimeng Sun. Change matters: Medication change prediction with recurrent residual networks. *arXiv preprint arXiv:2105.01876*, 2021a.

Chaoqi Yang, Cao Xiao, Fenglong Ma, Lucas Glass, and Jimeng Sun. Safedrug: Dual molecular graph encoders for recommending effective and safe drug combinations. *arXiv preprint arXiv:2105.02711*, 2021b.

Kuo Yang, Xin Dong, Shuhan Zhang, Haibin Yu, Liqun Zhong, Lei Zhang, He Zhao, Yutong Hou, Xinpeng Song, and Xuezhong Zhou. Presrecrf: Herbal prescription recommendation via the representation fusion of large tcm semantics and molecular knowledge. *Phytomedicine*, 135: 156116, 2024.

Nianzu Yang, Kaipeng Zeng, Qitian Wu, and Junchi Yan. Molerec: Combinatorial drug recommendation with substructure-aware molecular representation learning. In *Proceedings of the ACM web conference 2023*, pp. 4075–4085, 2023.

Xiaoyan Yang and Changsong Ding. Smrgat: A traditional chinese herb recommendation model based on a multi-graph residual attention network and semantic knowledge fusion. *Journal of ethnopharmacology*, 315:116693, 2023.

Yun Yang, Yulong Rao, Minghao Yu, and Yan Kang. Multi-layer information fusion based on graph convolutional network for knowledge-driven herb recommendation. *Neural Networks*, 146:1–10, 2022.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 694–699, 2002.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, et al. Huatuogpt, towards taming language model to be a doctor. In *Findings of the association for computational linguistics: EMNLP 2023*, pp. 10859–10885, 2023.

Yutao Zhang, Robert Chen, Jie Tang, Walter F Stewart, and Jimeng Sun. Leap: learning to prescribe effective and safe treatment combinations for multimorbidity. In *proceedings of the 23rd ACM SIGKDD international conference on knowledge Discovery and data Mining*, pp. 1315–1324, 2017.

Xinhan Zheng, Xueting Wang, Ruotai Li, Huyu Wu, Haopeng Jin, Yehan Yang, and Guodong Shan. Fmash: Advancing traditional chinese medicine formula recommendation with efficient fusion of multiscale associations of symptoms and herbs. *arXiv preprint arXiv:2503.05167*, 2025.

Xingzhi Zhou, Xin Dong, Chunhao Li, Yuning Bai, Yulong Xu, Ka Chun Cheung, Simon See, Xinpeng Song, Runshun Zhang, Xuezhong Zhou, et al. Tcm-ftp: fine-tuning large language models for herbal prescription prediction. In *2024 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pp. 4092–4097. IEEE, 2024.

Wei Zhu, Wenjing Yue, and Xiaoling Wang. Shennong-tcm: A traditional chinese medicine large language model. `https://github.com/michael-wzhu/ShenNong-TCM-LLM`, 2023.

Appendix for "From Probabilities to Prescriptions: A Unified Dependency-Aware Set Decoder for Medication and Herb Recommendation"

APPENDIX CONTENTS

A AI Use, Ethics, Reproducibility, and Limitations. 14

A.1 AI use statement 14

A.2 Ethics statement 14

A.3 Reproducibility statement 15

A.4 Limitations and Future Directions 15

B Extended Related Work 16

B.1 Western medication recommendation 16

B.2 Herb and formula recommendation 16

B.3 Post-hoc calibration and set-valued decisions 17

B.4 Set Representations and Structured Decision Objectives 17

B.5 Language models as recommendation plug-ins 17

C Proofs of the Theoretical Results. 18

C.1 Proof of Proposition 1: Limits of Prescription Prediction 18

C.2 Proof of Theorem 1: Jaccard Decisions under Conditional Independence 19

C.3 Second-Order Decomposition of Expected Jaccard 20

D Step-by-Step Implementation of DAPS. 21

D.1 Step 1: Calibration and Provisional Sizing 21

D.2 Step 2: Set-Conditioned Refinement and Final Sizing 22

D.3 Step 3: Global Conflict Adjustment 24

E Experimental Details and Supplementary Analyses 24

E.1 Data, Evaluation Protocol, and Metrics 24

E.2 Backbone Implementations and Reproduction 25

E.3 Language-Model Evaluation and Result Provenance 29

E.4 Detailed Main Comparisons 30

E.5 Decoder Ablations, Uncertainty, and Runtime 31

E.6 Training Strategy and Calibrator Families 34

E.7 Cardinality and Conflict-Budget Sensitivity 36

E.8 Ranking, Relation Audits, and Supplementary Baselines. 37

A AI USE, ETHICS, REPRODUCIBILITY, AND LIMITATIONS

A.1 AI USE STATEMENT

AI assistance in the research workflow. We used AI assistants to inspect and modify experiment, data-preparation, and plotting code; search and organize related literature; and edit and restructure the manuscript. Their assistance also included feedback on the theoretical framework, checking definitions, mathematical claims and proof text, and discussing experimental methodology, interpretation, and presentation scope. The authors specified the research questions and scope and made the final scientific decisions. Generative AI was not used to fabricate or manually alter reported numerical values: result tables are populated directly from experiment outputs under the evaluation and validation rules documented in Appendix E.

Verification and author responsibility. The authors reviewed AI-assisted text and code and checked mathematical claims against their assumptions; complete proofs are provided in Appendix C. Implementation checks include comparisons with released checkpoints and matched published targets, while methods lacking required code or resources are identified as paper-guided implementations. Every reported DAPS result is checked against its run for backbone identity and split roles. Appendix E documents these checks and adaptations, and Appendix E.2 records unresolved reproduction discrepancies and known protocol defects. The authors take responsibility for the final manuscript, mathematical claims, experimental results, and accompanying artifacts, including all AI-assisted content.

Language models evaluated as baselines. Language models also appear as experimental subjects in the zero-shot and fine-tuned baseline comparisons and in the GiantMed plug-in (Lv et al., 2026). These evaluations are distinct from the author assistance disclosed above. Their prompts, decoding settings, scoring procedures, and execution provenance are documented in Appendix E, with locally executed results distinguished from quoted results in the tables.

A.2 ETHICS STATEMENT

Data access and privacy. This study uses existing observational prescription corpora. MIMIC-III and MIMIC-IV require PhysioNet credentialing and are not redistributed; the code release provides preprocessing code rather than patient records. The TCM corpora originate from publicly released structured collections, and our preparation removes prescriptions carrying leaked patient identifiers. The code release does not redistribute the processed TCM corpora. Neither TCM source supplies patient or practitioner identifiers, so case-level splitting cannot establish patient-disjoint or practitioner-disjoint evaluation. Appendix E documents the data sources, construction, and evaluation units so that these access conditions and scope limitations remain explicit.

Intended use and broader impact. The intended contribution is a framework for studying and auditing set-valued prescription prediction. Evaluation measures agreement with recorded medication or herb sets, rather than therapeutic efficacy or patient outcomes. The observational data and membership-only target therefore do not support individual treatment recommendations. In particular, the global external-conflict budget uses size-preserving swaps within individual prescriptions but constrains only the pooled corpus-level rate. It does not guarantee a bound for every prescription or account for every patient-specific contraindication, dose, or interaction. A potential benefit is to make the distinction between predictive association and externally documented conflict explicit when auditing existing recommenders. Any clinical application would require prospective evaluation, patient-specific assessment, and clinician oversight beyond the experiments reported here.

Prescribing preferences and generalization. Models trained on recorded prescriptions may inherit institutional preferences, uneven representation of patient groups, and unsafe prescribing habits. TCM-Onsite reflects one hospital's practice, and the held-out relation audit in Appendix E.8 illustrates that an externally identified incompatibility can nevertheless have positive observational association. Keeping external conflict separate from learned co-selection makes this issue visible, but does not eliminate bias or establish fairness. Evaluation across institutions and patient subgroups would be needed to assess transfer beyond the observed prescribing environments. We therefore interpret the reported gains as improvements in prescription prediction within the disclosed protocols.

A.3 REPRODUCIBILITY STATEMENT

Theory and experimental specification. Section 3 describes the decoder, and Appendix C provides complete proofs of Proposition 1, Theorem 1, and Proposition 2. Appendix D details every decoder stage, and Appendix E records dataset construction, evaluation units, split roles, hyperparameters, and backbone adaptations. The matched comparisons retain the same frozen backbone exports and evaluation cohorts while changing the decoder; Appendix E.5 documents controlled component analyses and the separation of validation-based selection from evaluation. These references specify both the method and the conditions under which its reported results can be assessed.

Code, artifacts, and access conditions. The code release contains the decoder and baseline experiment code, compatibility patches, table-generation code, and protocol notes. Checkpoints, model outputs, and logs are excluded from version control, and the code release is not a standalone redistribution of the processed datasets. Reproduction thus requires obtaining the source data under their access conditions and generating the corresponding outputs. Appendix E specifies the experimental protocol, and every reported DAPS result is checked against its run for backbone identity and split roles. This distinguishes available implementation materials from data and outputs that must be acquired or regenerated separately.

Baseline provenance and reproduction discrepancies. Quoted results, locally executed baselines, and paper-guided implementations are identified separately in the result tables and implementation notes. Appendix E.2 documents checkpoint and cohort incompatibilities, the unresolved GiantMed reproduction gap, and known baseline protocol defects; Appendix E records adaptations where released code or required knowledge resources were unavailable. These disclosures delimit the comparisons supported by the available artifacts and allow subsequent work to revisit individual discrepancies without conflating them with the matched decoder evaluations.

A.4 LIMITATIONS AND FUTURE DIRECTIONS

DAPS deliberately focuses on the decision layer between candidate probabilities and prescription sets. Strong candidate-wise predictions still leave open how to choose set membership and cardinality while accounting for dependencies and external conflicts. Freezing the backbone lets matched comparisons attribute changes to the decoder rather than to a different representation or predictor. It also enables reuse of trained recommenders without repeating backbone training. Our objective is therefore complementary to developing a new backbone: to provide a reusable set-decoding method whose value can be assessed across prediction architectures and both clinical domains. The complete prediction system is the backbone–decoder combination; Tables 2 and 3 report its absolute performance as well as improvements over each backbone. Readers interested in the strongest evaluated system can compare these complete configurations within each dataset and protocol, while the matched gains assess the broader utility of the decoding method.

We also investigate joint optimization rather than assuming post-hoc training is preferable. In the evaluated MIMIC-III/IV configurations, post-hoc DAPS achieves higher Jaccard and F_1 and lower DDI than the joint model (Figure 4; Appendix E.6). These results support the effectiveness of the modular design in our setting. They do not establish a general advantage of freezing: the comparison also changes the auxiliary architecture, training objective, cardinality rule, and conflict-control procedure. Jointly optimized decoders with alternative designs remain a promising direction, alongside improvements to the underlying backbones.

The modular interface makes calibration, relation refinement, sizing, and conflict adjustment explicit while leaving the backbone's internal reasoning unchanged. The learned relations capture prescribing associations, and the reported gains concern agreement with recorded prescriptions rather than treatment effects or clinical benefit. A natural next step is to connect this decoding interface with disease–treatment–outcome models and evaluate the resulting predictions using outcome-aware objectives alongside prescription overlap.

The current formulation selects membership from a fixed candidate vocabulary and uses pairwise relation summaries. Higher-order relations, dosage and duration, alternative set objectives, and interfaces for name-generation models offer concrete extensions. The global external-conflict budget provides an explicit aggregate constraint based on the supplied knowledge resource; patient-specific contraindications and per-prescription safeguards would require additional information and con-

straints. Broader knowledge coverage and prospective, clinician-supervised evaluation could help assess these extensions beyond the retrospective prediction setting studied here.

B Extended Related Work

B.1 Western medication recommendation

Medication recommendation predicts a set of drugs from the current clinical state and, for longitudinal models, previous visits. Published comparators in Appendix Table 8 include the reverse-time attention model RETAIN (Choi et al., 2016), the sequence generator LEAP (Zhang et al., 2017), and GAMENet's graph-augmented memory (Shang et al., 2019). Later systems add explicit inductive biases: MICRON models changes between adjacent visits (Yang et al., 2021a); SafeDrug incorporates molecular structure and a DDI controller (Yang et al., 2021b); DEPOT reduces temporal and demographic popularity bias (Sun et al., 2022); MoleRec represents medications through molecular substructures (Yang et al., 2023); VITA selects and weights useful historical visits (Kim et al., 2024); COGNet decodes medications autoregressively (Wu et al., 2022); and HiRef hierarchically refines clinical representations (Chok et al., 2025). MedAlign aligns multiple EHR views before prediction (Lv et al., 2025); its checkpoint and preprocessing are also distributed with GiantMed (Lv et al., 2026).

These methods primarily improve estimation of $P(\mathbf{m} \mid \mathbf{x})$ through a new encoder, graph, memory, or loss. DAPS instead holds that estimator fixed and changes the set decision induced by its marginal scores. This distinction is also why our paired comparisons attach DAPS only to backbones for which exact fit/evaluation probability exports are available; historical rows copied from prior comparison tables are contextual baselines, not plug-in experiments. The two research directions are complementary: improved representations can supply better candidate scores, while an explicit decoder addresses how those scores translate into a variable-size prescription. Holding the predictor fixed also permits a matched assessment of the decision layer without confounding its contribution with a change in the underlying architecture.

B.2 Herb and formula recommendation

Herb recommendation maps a symptom presentation to an unordered prescription. Early work uses knowledge-enhanced topic modeling (KGETM) (Wang et al., 2019) and symptom–herb graph propagation (SMGCN) (Jin et al., 2020). KDHR adds herb attributes and multi-layer fusion (Yang et al., 2022), while SMRGAT introduces residual attention over symptom–symptom, herb–herb, and bipartite graphs (Yang & Ding, 2023). SCEIKG models condition evolution and exclusion knowledge (Liu et al., 2023); PresRecST decomposes prediction into syndrome, treatment-method, and herb stages (Dong et al., 2024); and PresRecRF fuses relational and semantic representations (Yang et al., 2024). TCMRGCL learns contrastive graph views (Hu et al., 2025), PRDAGE combines graph, semantic, attribute, and sequence features (Wen et al., 2025), and FMASH couples graph propagation with degree-ordered state-space modeling and multi-scale herb information (Zheng et al., 2025).

Generative approaches instead fine-tune domain language models: TCM-FTP adapts large language models (LLMs) to herb-set generation (Zhou et al., 2024), and PFGCN combines a fine-tuned language model with graph convolution on private clinical records (Gao et al., 2026). Our LLM study therefore includes general models alongside the domain models ShenNong-TCM (Zhu et al., 2023) and HuatuoGPT2 (Zhang et al., 2023), both zero-shot and under a common TCM-FTP-style fine-tuning recipe. We use *herb recommendation* as the umbrella term because every evaluated system ultimately selects an unordered herb set; "formula recommendation" and "prescription prediction" in prior work refer to closely related output objects.

The two TCM corpora do not ship every knowledge graph, molecular descriptor, semantic embedding, intermediate syndrome label, or official split assumed by these papers. Consequently, the experiments separate published numbers from our common-protocol implementations and disclose which modules survive on each corpus. Appendix E gives this information method by method; a missing resource is never silently replaced with a fabricated one.

B.3 POST-HOC CALIBRATION AND SET-VALUED DECISIONS

Post-hoc calibration can repair distorted neural-network probabilities without retraining the predictor (Zadrozny & Elkan, 2002; Guo et al., 2017), while conformal prediction constructs sets with distribution-free coverage under its exchangeability conditions (Vovk et al., 2005; Angelopoulos & Bates, 2023). DAPS has a different target: it directly selects a set for expected Jaccard and does not claim conformal coverage. The closest decision-theoretic precedent is F-measure maximization, where the optimal multilabel set depends on the full score vector rather than a fixed per-label threshold, and plug-in decisions can approach the Bayes rule when their probability estimates are accurate (Dembczyński et al., 2012; Waegeman et al., 2014). Our cardinality stage instantiates that principle for Jaccard and combines it with cross-fitted calibration and a typed, set-conditioned refinement.

The distinction between probability estimation and decision making also appears in threshold selection. Lipton et al. (2014) show that F_1-optimal thresholding depends on the attainable score; calibrated probabilities alone do not justify a universal threshold of 0.5. This motivates treating the decision rule as a separate design choice, rather than assuming that improved probabilities automatically yield the best discrete predictions. Their thresholding analysis does not directly give a per-prescription Jaccard rule: DAPS evaluates candidate set sizes using an approximate expected-Jaccard criterion after context-aware score refinement.

Calibration families and risk objectives. Beta calibration provides a parametric correction for binary probabilities (Kull et al., 2017), while Dirichlet calibration extends this approach to multiclass probability vectors (Kull et al., 2019). These methods motivate separating the probability estimator from a corrective map, but do not by themselves select a variable-size multilabel prescription. Our calibrators also use candidate and provisional-set context. Appendix E.6 therefore compares alternative learners in this context-rich calibration role; it is not a comparison against implementations of beta or Dirichlet calibration. Conformal risk control provides expected-risk guarantees for monotone losses under its assumptions (Angelopoulos et al., 2024). Our fit-selected conflict budget is an empirical decoding constraint, without such a distribution-free guarantee.

B.4 SET REPRESENTATIONS AND STRUCTURED DECISION OBJECTIVES

Deep Sets characterizes permutation-invariant set functions (Zaheer et al., 2017), and Set Transformer models interactions among set elements using attention (Lee et al., 2019). These works address how to represent unordered collections. Our candidate-to-set summaries similarly avoid a prescription order, but serve as typed context for a frozen-backbone decoder rather than a replacement encoder. Permutation invariance alone does not distinguish co-selection from pharmacological compatibility, nor determine the set size that optimizes the evaluation objective.

Probabilistic classifier chains explicitly represent label dependence and connect multilabel prediction to Bayes decision making (Cheng et al., 2010). This complements the distinction between marginal scores and set-optimal decisions established in Section 2. The Lovasz–Softmax loss offers a tractable surrogate for intersection-over-union optimization in semantic segmentation (Berman et al., 2018), demonstrating a training-time approach to a non-decomposable overlap objective. DAPS instead retains the backbone and performs an approximate Jaccard-oriented decision after calibration. Neither a richer set representation nor an overlap-oriented objective supplies the external semantics needed to interpret medication conflicts.

Dembczynski et al. (2013) explicitly compare a probabilistic plug-in rule with structured loss minimization for multilabel F-measure optimization. Their separation of probability estimation from utility-based inference provides a precedent for studying the decision stage as an algorithmic contribution in its own right. DAPS adopts this separation for a different objective and interface: frozen candidate scores, empirical relation context, and an external-conflict budget. The consistency results for their F-measure algorithm do not transfer automatically to our approximate Jaccard decoder; our matched-backbone experiments assess the practical value of this modular design.

B.5 LANGUAGE MODELS AS RECOMMENDATION PLUG-INS

GiantMed is closest in placement: it freezes a medication recommender and asks an LLM to adjudicate candidates near a probability boundary (Lv et al., 2026). DAPS occupies the same post-hoc

slot but treats the residual error as a set-decision problem rather than a missing-knowledge problem. This predicts the pattern in Appendix Figure 7: LLM adjudication helps most when a weak backbone under-prescribes, whereas calibration and cardinality selection remain useful across stronger backbones. Because released GiantMed results and our rerun differ, the paper reports our local rerun and preserves the audit trail in Appendix E.2 instead of substituting the published number.

The comparison also separates two ways to introduce auxiliary information. LLM adjudication uses a language model to reconsider candidate choices, whereas DAPS combines a training-fitted expert with explicit relation features and a fit-selected set decision. Both retain the original backbone, but the source of auxiliary evidence and the output-selection mechanism differ. This makes matched exports and evaluation cohorts important when interpreting gains: they compare complete post-hoc configurations, not the isolated value of language-model knowledge. The accompanying size and ranking analyses distinguish changes in predicted membership counts from changes in candidate ordering.

C Proofs of the Theoretical Results

Worked examples of the unified formulation. The following toy cases instantiate the notation in Section 2; they illustrate the prediction interface and do not propose treatment for any patient.

Western EHR example. Suppose the recorded pre-prescription presentation $X = x$ contains diagnoses for hypertension and type-2 diabetes, a procedure history, and medications from earlier visits. For an illustrative vocabulary ordered as lisinopril, losartan, metformin, glipizide, aspirin, and furosemide, the recorded target might be represented as $M = \{\text{losartan}, \text{metformin}, \text{aspirin}\}$, or equivalently by the multi-hot vector $(0, 1, 1, 0, 1, 0)$. A frozen medication recommender returns six candidate marginals $\eta(x) \in [0, 1]^6$. The set decoder must then select one prediction $S \subseteq [6]$, including its cardinality, from those scores. Different prescriptions may be recorded for presentations with the same encoded diagnoses and history, so x need not determine the realized M.

TCM example. Suppose $X = x$ records sore throat, cough, yellow sputum, demographics, and any available earlier prescription history. For an illustrative vocabulary ordered as *Jinyinhua*, *Lianqiao*, *Jiegeng*, *Kuxingren*, *Gancao*, and *Chenpi*, a recorded herb set might be $M = \{\textit{Lianqiao}, \textit{Jiegeng}, \textit{Gancao}, \textit{Chenpi}\}$, represented by $(0, 1, 1, 0, 1, 1)$. A frozen herb recommender again returns one marginal score per candidate, and the same set-decoding problem determines the membership and size of S. The shared mathematical interface does not equate the clinical meanings, vocabularies, or external knowledge resources of the two domains.

C.1 Proof of Proposition 1: Limits of Prescription Prediction

For part (i) of Proposition 1, fix $X = x$. A deterministic prediction $f(x)$ matches the recorded prescription with probability $P(M = f(x) \mid X = x)$, which is maximized by choosing a most probable set; hence $R^*_{\text{exact}}(x) = 1 - \max_S P(M = S \mid X = x)$. If the conditional law $P(M \mid X = x)$ is non-degenerate, two distinct sets $S_1 \neq S_2$ have positive conditional probability. Whichever set attains the maximum, at least one of them is excluded, so $\max_S P(M = S \mid X = x) \leq 1 - \min\{P(M = S_1 \mid X = x), P(M = S_2 \mid X = x)\} < 1$ and $R^*_{\text{exact}}(x) > 0$. Averaging over X, the overall minimum exact-set error $\inf_f P(f(X) \neq M) = \mathbb{E}_X[R^*_{\text{exact}}(X)]$ is positive whenever non-uniqueness holds with positive probability.

For a concrete example of part (i), fix a presentation $X = x$ and suppose

$$P(M = \{a, c\} \mid X = x) = 0.7, \qquad P(M = \{b, c\} \mid X = x) = 0.3,$$

so that a and b are interchangeable given $S = \{c\}$. The best exact-match prediction is $\{a, c\}$, but $R^*_{\text{exact}}(x) = 1 - 0.7 = 0.3 > 0$. Even a predictor that knows these probabilities exactly cannot always recover the recorded prescription from x alone. This is a conditional error for this presentation; the overall minimum error averages the corresponding conditional errors over X.

For part (ii), fix one presentation and let $K = 3$. Order four nonempty outcomes as $001, 011, 101, 111$ and consider

$$p = (0.4, 0.1, 0.2, 0.3), \qquad p' = (0.3, 0.2, 0.3, 0.2).$$

Both laws have candidate marginals $(0.5, 0.4, 1)$ but are different joint distributions. Because candidate 3 occurs surely, a Jaccard-optimal prediction must contain it. The expected Jaccard values for the four remaining predictions are

S	$\{3\}$	$\{1,3\}$	$\{2,3\}$	$\{1,2,3\}$
$J_p(S)$	$13/20$	$19/30$	$17/30$	$19/30$
$J_{p'}(S)$	$37/60$	$13/20$	$7/12$	$19/30$.

Hence p has the unique optimum $\{3\}$, whereas p' has the unique optimum $\{1,3\}$. Appending zero-probability coordinates proves the claim for every $K > 3$. Thus identical candidate marginals can imply different optimal set decisions under Jaccard, completing the proof.

Link to Theorem 1. The law p' factorizes: $P(Y_1 = a, Y_2 = b)$ equals the product of the marginals 0.5 and 0.4 for every $a, b \in \{0, 1\}$, so p' is the conditionally independent law with these marginals, whereas p puts extra mass on $Y_1 = Y_2 = 0$ and is positively dependent. The rule of Theorem 1(i) depends only on the marginals, so it returns $\{1, 3\}$ under both laws. This is optimal under p' but costs $13/20 - 19/30 = 1/60$ expected Jaccard under p: the ranking is still correct, and the error lies in the size.

C.2 Proof of Theorem 1: Jaccard Decisions under Conditional Independence

Throughout, x is fixed, expectations are conditional on $X = x$, and $f_k(a, b) = a/(k + b)$, so that $J(S, M) = f_k(A, B)$ for a nonempty S of size k.

Part (i): exchange argument. Let S be nonempty with $|S| = k$, and take $i \in S$ and $j \notin S$ with $\eta_j \geq \eta_i$. Let $S' = (S \setminus \{i\}) \cup \{j\}$. Condition on every indicator except Y_i and Y_j, and let $A_0 = |(S \setminus \{i\}) \cap M|$ and $B_0 = |M \setminus (S \cup \{j\})|$. Then $J(S, M) = f_k(A_0 + Y_i, B_0 + Y_j)$ and $J(S', M) = f_k(A_0 + Y_j, B_0 + Y_i)$. The two values coincide when $Y_i = Y_j$, so

$$\mathbb{E}[J(S', M) - J(S, M) \mid A_0, B_0] = \big[P(Y_i = 0, Y_j = 1) - P(Y_i = 1, Y_j = 0)\big]\,\Delta,$$

with $\Delta = f_k(A_0 + 1, B_0) - f_k(A_0, B_0 + 1)$, where the conditional probabilities equal the unconditional ones by independence. The bracket is $\eta_j(1 - \eta_i) - \eta_i(1 - \eta_j) = \eta_j - \eta_i \geq 0$, and $\Delta \geq 0$ because $(A_0 + 1)/(k + B_0) \geq A_0/(k + B_0) \geq A_0/(k + B_0 + 1)$. Hence $U_\eta(S') \geq U_\eta(S)$. Repeating such exchanges turns any set of size k into S_k without decreasing expected Jaccard, so S_k is optimal among sets of size k, and an optimal set is found among $S_0, \ldots, S_K$. For $k \geq 1$, A and B are sums of independent indicators over disjoint candidate sets, so they are independent Poisson–binomial counts and $\mathbb{E}[A/(k + B)] = \mathbb{E}[A]\,\mathbb{E}[1/(k + B)]$.

Part (ii): plug-in regret. Recall that $U_p(S)$ denotes expected Jaccard when the indicators are independent with marginals $p \in [0, 1]^K$. $U_p(S)$ is affine in each p_j, with slope $\mathbb{E}[J \mid Y_j = 1] - \mathbb{E}[J \mid Y_j = 0]$. For nonempty S with $|S| = k$, the slope is $\mathbb{E}[1/(k + B)] \in [0, 1/k]$ if $j \in S$, and $-\mathbb{E}[A/((k + B')(k + B' + 1))] \in [-1/(k + 1), 0]$ if $j \notin S$, where B' omits j and $A \leq k \leq k + B'$. For $S = \varnothing$, $U_p(S) = \prod_j (1 - p_j)$ has slopes of magnitude at most one. Changing one coordinate at a time therefore gives $|U_p(S) - U_q(S)| \leq \|p - q\|_1$ for every S. Let $S^\star$ maximize U_η and note that $\widehat{S}$ maximizes $U_{\hat{p}}$ over all subsets by part (i) applied to the independent law with marginals $\hat{p}$. Then

$$U_\eta(S^\star) - U_\eta(\widehat{S}) \leq \big[U_\eta(S^\star) - U_{\hat{p}}(S^\star)\big] + \big[U_{\hat{p}}(S^\star) - U_{\hat{p}}(\widehat{S})\big] + \big[U_{\hat{p}}(\widehat{S}) - U_\eta(\widehat{S})\big] \leq 2\|\hat{p} - \eta\|_1,$$

since the middle bracket is nonpositive. Keeping the slope bounds gives the sharper $\|\hat{p} - \eta\|_1\,(1/\max(1, |S^\star|) + 1/\max(1, |\widehat{S}|))$. If conditional independence holds for almost every x, averaging over X bounds the expected regret by $2\,\mathbb{E}_X\|\hat{p}(X) - \eta(X)\|_1$.

Link to log-loss calibration. Let $\Delta_j = \mathbb{E}_X[\mathrm{KL}(\mathrm{Ber}(\eta_j(X)) \,\|\, \mathrm{Ber}(\hat{p}_j(X)))]$ be the excess binary log loss of $\hat{p}_j$ over the true marginal. Pinsker's inequality gives $\mathrm{KL}(\mathrm{Ber}(a)\|\mathrm{Ber}(b)) \geq 2(a - b)^2$, and Jensen's inequality then gives $\mathbb{E}_X|\hat{p}_j(X) - \eta_j(X)| \leq \sqrt{\Delta_j/2}$. Hence the expected regret is at most $\sqrt{2}\sum_j \sqrt{\Delta_j}$. The bound is loose when K is large, but under the working model it shows that the quantity the DAPS calibrators minimize controls the regret of the plug-in rule. Ranking quality alone does not control it, because a correct ranking with miscalibrated magnitudes can still produce the wrong size.

Relation to the implemented rule. DAPS ranks candidates by its calibrated probabilities $\tilde{p}^{(1)}$ and, with source $s_i = \arg\max_{1\le k\le 45} \widehat{J}_{\mathrm{PB}}(S_k)$ in equation 9, uses the size $k_i = \mathrm{clip}_{[1,45]}(\mathrm{round}\{a[(1-w)b_i + ws_i] + o\})$, where b_i is the backbone count and (w, a, o) are fit-selected (Appendix D.2). With $w = a = 1$ and $o = 0$ this is exactly the plug-in rule $\widehat{S}$ of part (ii) with $\hat{p} = \tilde{p}^{(1)}$, restricted to $1 \le k \le 45$; the restriction $k \ge 1$ is without loss because recorded prescriptions are nonempty, so $U(S_0) = 0$. Parts (i)–(ii) therefore justify this base rule under the working model. The blend, scale, offset, and the alternative soft-Jaccard source (selected in 8 of 36 matched evaluations) absorb residual miscalibration and dependence, are chosen on fit data only, and carry no separate guarantee. Table 16 isolates the size rule empirically.

C.3 Second-Order Decomposition of Expected Jaccard

The following result quantifies where dependence enters set-level utility; it is referenced from Section 2.

Proposition 2 (A second-order decomposition of expected Jaccard)**.** *Fix a presentation x and a nonempty prediction $S \subseteq [K]$ of size k; all moments below are conditional on $X = x$. Let $A = |S \cap M|$, $B = |M \setminus S|$, $\mu_A = \mathbb{E}[A]$, $\mu_B = \mathbb{E}[B]$, $d = k + \mu_B$, and $Z = B - \mu_B$. Then*

$$\mathbb{E}[J(S, M)] = \underbrace{\frac{\mu_A}{d} - \frac{\mathrm{Cov}(A, B)}{d^2} + \frac{\mu_A \,\mathrm{Var}(B)}{d^3}}_{U_2(S)} + R(S), \tag{7}$$

where $U_2(S)$ depends only on candidate marginals and pairwise conditional covariances. The remainder $R(S)$ captures higher-order terms; its exact expression and bound are given below. No independence assumption is required.

Expansion and remainder bound. Fix x and S as in the proposition, with all expectations conditional on $X = x$. Since $d + Z = k + B \ge k > 0$, the following finite algebraic identity holds without requiring $|Z| < d$:

$$\frac{1}{d+Z} = \frac{1}{d} - \frac{Z}{d^2} + \frac{Z^2}{d^3} - \frac{Z^3}{d^3(d+Z)}.$$

Multiplying by A and taking expectations gives

$$\mathbb{E}\left[\frac{A}{k+B}\right] = \frac{\mu_A}{d} - \frac{\mathbb{E}[AZ]}{d^2} + \frac{\mathbb{E}[AZ^2]}{d^3} - \mathbb{E}\left[\frac{AZ^3}{d^3(d+Z)}\right].$$

Now $\mathbb{E}[AZ] = \mathrm{Cov}(A, B)$ and $\mathbb{E}[AZ^2] = \mu_A \,\mathrm{Var}(B) + \mathbb{E}[(A - \mu_A)Z^2]$. Hence the exact remainder is

$$R(S) = \frac{\mathbb{E}[(A - \mu_A)Z^2]}{d^3} - \mathbb{E}\left[\frac{AZ^3}{d^3(d+Z)}\right].$$

The triangle inequality, $A \ge 0$, and $d + Z \ge k$ yield

$$|R(S)| \le \frac{\mathbb{E}[|A - \mu_A|Z^2]}{d^3} + \frac{\mathbb{E}[A|Z|^3]}{kd^3}. \tag{8}$$

This bound need not be small for arbitrary prescription distributions. All moments exist because the vocabulary is finite.

Which dependencies enter the approximation? Write $Y_i = \mathbf{1}\{i \in M\}$, $\eta_i = \mathbb{E}[Y_i]$, and $\Sigma_{ij} = \mathrm{Cov}(Y_i, Y_j)$, still conditional on $X = x$. Then

$$\mu_A = \sum_{i\in S} \eta_i, \qquad \mu_B = \sum_{j\notin S} \eta_j,$$

$$\mathrm{Cov}(A, B) = \sum_{i\in S}\sum_{j\notin S} \Sigma_{ij}, \qquad \mathrm{Var}(B) = \sum_{j\notin S} \eta_j(1-\eta_j) + 2\sum_{\substack{i<j \\ i,j\notin S}} \Sigma_{ij}.$$

Let $U_{2,\mathrm{ind}}(S)$ be the same second-order approximation after setting off-diagonal covariances to zero while retaining the marginals. The dependence correction is exactly

$$U_2(S) - U_{2,\mathrm{ind}}(S) = -\frac{\sum_{i\in S}\sum_{j\notin S}\Sigma_{ij}}{d^2} + \frac{2\mu_A \sum_{i<j,\, i,j\notin S}\Sigma_{ij}}{d^3}.$$

Cross-boundary dependence and dependence among omitted candidates thus enter with different coefficients. This is a comparison of second-order approximations, not an exact difference between dependent and independent expected Jaccard. In particular, $U_{2,\mathrm{ind}}$ is not the full Poisson–binomial expectation used by DAPS.

Implication for selecting a candidate set. For a finite family $\mathcal{S}$ of nonempty candidate sets, define $U(S) = \mathbb{E}[J(S, M)]$, and suppose an estimate $\widehat{U}_2$ satisfies $\max_{S\in\mathcal{S}} |\widehat{U}_2(S) - U_2(S)| \le \delta$. Let $\varepsilon = \max_{S\in\mathcal{S}} r(S)$, where $r(S)$ is the right-hand side of the proposition's remainder bound. If $S^* \in \arg\max_{S\in\mathcal{S}} U(S)$ and $\widehat{S} \in \arg\max_{S\in\mathcal{S}} \widehat{U}_2(S)$, then

$$U(S^*) - U(\widehat{S}) \le 2(\varepsilon + \delta).$$

Indeed, insert $\widehat{U}_2(S^*)$ and $\widehat{U}_2(\widehat{S})$ between the two utilities; their difference is nonpositive by the definition of $\widehat{S}$, and each approximation error is at most $\varepsilon + \delta$. For top-k candidates under a fixed ranking, this controls the sizing decision only within that family, not across all subsets. This conditional result does not supply an estimator or a finite-sample guarantee for δ, nor does it assert that the remainder bound is small.

D Step-by-Step Implementation of DAPS

This section expands the pipeline in Section 3; DAPS denotes the complete learned decoder with the global external-conflict budget. Dataset and backbone-specific preparation is specified separately in Appendix E. Input alignment is preparation, not an additional decoding step. The three steps below match Section 3 and Figure 2: calibration and provisional sizing; set-conditioned refinement with final cardinality selection; and global conflict adjustment. Step 2 first updates scores at the provisional count and then reselects the count from the refined scores.

Preparation: frozen probability exports and input alignment. For each split, the decoder takes a probability matrix in $[0, 1]^{n\times K}$, a binary target matrix, and evaluation-unit identifiers. The record order and vocabulary must match the exported probabilities, and all exports are aligned to one dense candidate order; when visit indices are present, the records are filtered to the visits actually scored by the backbone, and a mismatch between probability and record counts is rejected. Previously prescribed medications are read from earlier visits, never from the current target. A supplied binary prediction matrix preserves the backbone's native decoding, including autoregressive models for which a 0.5 threshold would not reproduce its predictions. Otherwise, thresholding supplies the base set. Its size is retained as an input to the cardinality search.

The training split fits the auxiliary expert bank and observational relations. The fit split supplies calibration targets and selects decoding settings. The evaluation split is used to score the frozen pipeline; its labels are not used to construct candidate features or select settings.

D.1 Step 1: Calibration and Provisional Sizing

Auxiliary experts and candidate features. The auxiliary bank trains one gradient-boosted binary classifier per output medication on training records. The reported bank uses 120 trees for each of two configurations with 15 and 31 leaves and different random seeds. Available inputs include current condition and procedure indicators, previous medications, recency-weighted histories, changes in condition/procedure sets, and visit summaries. Inputs without a corresponding history remain absent or zero according to the corpus build. The expert outputs complement the frozen backbone; they do not update it.

For each probability source p_j, candidate features include p_j, its logit after clipping to $[10^{-5}, 1 - 10^{-5}]$, normalized within-case rank, and its product with previous-set membership. Further features

describe the backbone score's distance from its selection boundary, previous membership, base and previous set sizes, visit index, condition/procedure counts, graph channels, and candidate identity. The feature tensor has shape $n \times K \times d$ and is flattened to nK candidate examples for binary calibration. Candidate identity is the final feature and is treated as categorical by the LightGBM implementation. Previous-set graph channels and current-set relation channels are distinct: the former are available before the provisional prescription is constructed.

First-round cross-fitted calibration. The first binary LightGBM calibrator uses 200 trees, learning rate 0.035, 15 leaves, minimum child size 200, feature subsampling 0.85, and L_1/L_2 penalties $0.5/5.0$. Five folds are assigned by evaluation-unit identifier modulo five. All visits of an identified patient share a fold; case-based corpora use the corresponding case identifier. For fold v, fit g_1^{-v} on the other four folds and write

$$\tilde{p}_{ij}^{(0)} = g_1^{-v(i)}(F_{ij})$$

for each held-out candidate. These out-of-fold scores support fit-split cardinality selection. A separate g_1^{full} is trained on the complete fit split and produces evaluation scores. This separates training a calibrator from measuring its fit-split predictions, while retaining all fit examples for the final model.

Provisional set construction. The first-round out-of-fold probabilities select a cardinality configuration by the search in Appendix D.2. For case i, this produces $k_i^{(0)}$ and $S_i^{(0)} = \mathrm{TopK}(\tilde{p}_i^{(0)}, k_i^{(0)})$. The same fit-selected configuration is applied to first-round evaluation probabilities. The provisional set is a feature-construction device: it is not the final output and is not a ground-truth prescription. Its membership, size, and relation summaries are recomputed separately for each case.

D.2 Step 2: Set-Conditioned Refinement and Final Sizing

Training-only typed relation matrices. Let $Y \in \{0,1\}^{n_{\text{tr}} \times K}$ be the training prescription matrix, $\hat{p}_j = n_{\text{tr}}^{-1} \sum_i Y_{ij}$, and $\hat{p}_{j\ell} = n_{\text{tr}}^{-1} \sum_i Y_{ij} Y_{i\ell}$. Pearson correlation is computed between nonconstant columns. Constant-column correlations and all diagonals are set to zero. PMI uses

$$\mathrm{PMI}_{j\ell} = \log \frac{\hat{p}_{j\ell} + 10^{-9}}{\hat{p}_j \hat{p}_\ell + 10^{-9}},$$

with entries zeroed when the pair occurs fewer than five times. The learned views are positive Pearson correlation, positive PMI, and the magnitude of negative Pearson correlation. They are kept as three nonnegative matrices, with the first two representing co-selection and the third exclusion.

These are corpus-level training statistics, not estimates of each presentation-specific conditional covariance in the theory. Their role is predictive context whose interpretation respects the distinction established in Section 2. External DDI or TCM incompatibility information supplies a separate conflict matrix. ATC-1 and ATC-3 membership supplies Western redundancy channels; these channels are zero for TCM rather than inferred from herb names. Held-out semantic audit labels are not relation-training inputs.

Set-conditioned features and second-round calibration. For each learned relation matrix R, candidate j, and nonempty provisional set S, the relation context contains

$$C_j^R(S) = \left(\frac{1}{|S|} \sum_{\ell \in S} R_{j\ell},\ \min_{\ell \in S} R_{j\ell},\ \max_{\ell \in S} R_{j\ell} \right).$$

Empty-set summaries are zero. With zero diagonals, membership of j does not create a self-relation, although the denominator remains $|S|$ as in the implementation. Additional channels include EHR affinity mean/maximum, external conflict density/count, ATC redundancy, provisional membership and size, and first-round rank. The concatenated tensor retains candidate identity as its last feature.

The second calibrator uses the same 200-tree configuration, with its own fold-specific random seeds. It is cross-fitted over the same five unit-disjoint folds and then refitted on all fit examples. Fit features are built from first-round out-of-fold scores; evaluation features use g_1^{full}. The procedure is sequential cross-fitting, not a fully nested refit of both stages for every outer fold: the first-stage out-of-fold feature tensor and the fit-selected provisional count rule are reused when cross-fitting g_2. Evaluation labels remain excluded throughout. The resulting $\tilde{p}^{(1)}$ supplies the final score order and the second cardinality search.

Training-partition set-completion models. The dependency features $D_j(S)$ in equation 5 come from two models fitted on training prescriptions only, with case features x given by the presentation codes. Both are two-layer networks (width 512, dropout 0.2) trained for 30 epochs with AdamW (learning rate 10^{-3}, weight decay 10^{-4}, batch 256); each is averaged over three seeds. (i) A *masked set-completion model* estimates $P(y_j{=}1 \mid x,\, y_{-j})$. In each batch a random 10–60% of candidates is hidden; the visible part of the true prescription is corrupted by dropping 25% of its members and adding frequency-weighted false candidates, and the binary cross-entropy is taken on hidden candidates only. At decoding time S is the visible set and the candidates are split into eight fixed groups that are hidden in turn, so candidate j never observes its own membership. It contributes the conditional score, the set-free score obtained by hiding every candidate, their logit difference, and the conditional rank. (ii) A *typed denoising model* estimates $P(y \mid x, S, \hat{p})$. Its training sets are provisional prescriptions built on the training partition from two-fold cross-fitted marginal networks $\hat{p}$ (top-k with $k = \mathrm{round}(u \sum_j \hat{p}_j)$, $u \sim \mathcal{U}[0.8, 1.2]$). Each typed relation R_r defined above—positive correlation and rescaled positive PMI (co-selection), negative correlation (exclusion), the external conflict indicator, and, on MIMIC, ATC-3 redundancy—adds $\sigma_r\ \mathrm{softplus}(\gamma_r(x))_j\ (R_r S)_j/|S|$ to the logit of candidate j, with $\sigma_r = +1$ for co-selection and -1 otherwise, so a relation's sign is fixed while its strength depends on the case. It contributes its probability, the logit difference to the current round's scores, its rank, and one channel per typed relation. The second calibrator g_2 receives these features together with the relation context above and is boosted from $\mathrm{logit}\,\tilde{p}^{(0)}$ as the LightGBM initial score. A second round rebuilds all set features from $S^{(1)} = \mathrm{TopK}(\tilde{p}^{(1)}, k^{(1)})$, where $k^{(1)}$ is selected by the cardinality search below, and is boosted from the first round's scores; the final $k^\star$ is selected on the second-round scores. No validation or evaluation label enters either model.

Final cardinality selection. Order candidate probabilities as $p_{(1)} \geq \cdots \geq p_{(K)}$ and let S_k be the first k candidates. Under the independent Bernoulli approximation, $Z_k = \sum_{j>k} m_{(j)}$ has a Poisson–binomial distribution and is independent of the number of selected true positives. Hence

$$\widehat{J}_{\mathrm{PB}}(S_k) = \left(\sum_{j \leq k} p_{(j)} \right) \sum_{z=0}^{K-k} \frac{\Pr(Z_k = z)}{k + z}. \tag{9}$$

The outside-count distribution is accumulated from the end of the candidate order using the recurrence $u_{\mathrm{new}}(z) = (1-p)u(z) + pu(z-1)$, starting from $u(0) = 1$. This evaluates an expected ratio, rather than merely a ratio of expectations. It does not assume the true prescription law is identified by the marginals.

The search compares three size sources: the maximizer of equation 9, the soft-Jaccard maximizer of $\sum_{j \leq k} p_{(j)} / (k + \sum_{j>k} p_{(j)})$, and the marginal sum $\sum_j p_j$. For a source value s_i and backbone count b_i, the searched rule is

$$k_i = \mathrm{clip}_{[1,45]}(\mathrm{round}\{a[(1-w)b_i + ws_i] + o\}),$$

where $w \in \{0, 0.1, \ldots, 1\}$, $a \in \{0.900, 0.925, \ldots, 1.100\}$, and $o \in \{-1, -0.5, 0, 0.5, 1\}$. All reported vocabularies exceed the cap of 45. The cap exceeds every recorded TCM prescription (maximum 21 herbs on TCM-Onsite and 40 on TCM-Internet) and is exceeded by only 10 of 15,031 MIMIC-III visits (0.07%; maximum 53) and 27 of 119,210 MIMIC-IV visits (0.02%; maximum 65), whereas mean predicted sizes lie between 14.7 and 21.1 (Table 9). The source and parameters maximizing mean fit-row Jaccard are retained; ties retain the first configuration visited. The complete search is repeated on second-round out-of-fold probabilities because refinement can change both rank and scale. The selected configuration is then frozen for evaluation. Across the 36 reported matched evaluations, the fit split selects the Poisson–binomial expected-Jaccard source in 28 cases and soft Jaccard in eight; the marginal-sum source is never selected. Thus expected Jaccard is the central size criterion, implemented within a validation-selected family that also includes simpler size estimates. This selection frequency does not by itself measure the performance advantage of one source over another; Table 16 reports that comparison with all other components fixed.

D.3 Step 3: Global Conflict Adjustment

For a collection of predicted sets $\mathcal{S} = (S_i)_i$ and symmetric binary conflict matrix A, the budget uses the pooled pair rate

$$R(\mathcal{S}) = \frac{\sum_i \sum_{j<\ell,\, j,\ell \in S_i} A_{j\ell}}{\max\{1, \sum_i \binom{|S_i|}{2}\}}.$$

The ceiling is $\tau = \max\{0.99 R(\mathcal{S}^B_{\text{fit}}), 10^{-6}\}$. If the learned fit predictions already meet the ceiling, the selected budget configuration is inactive and is also left inactive on evaluation. Otherwise, the search considers five utility definitions (rank, probability, logit, mean rank/probability, and mean rank/normalized-logit), candidate-pool extensions 20/35/60, and risk powers $0.75/1/1.25$.

For removing $j \in S$ and adding $\ell \notin S$, the reduction in flagged pairs is

$$\Delta R_{j\to\ell} = \sum_{t \in S\setminus\{j\}} A_{jt} - \sum_{t \in S\setminus\{j\}} A_{\ell t}.$$

Only positive reductions are eligible; swaps are ranked by $(u_j - u_\ell)/(\Delta R_{j\to\ell})^\gamma$. The implementation precomputes up to four sequential swaps per case, then uses a global priority queue while respecting each case's swap order. It stops when the required pooled conflict reduction is achieved or no precomputed swap remains. Membership counts are preserved. Consequently, the procedure controls the pooled rate but does not require every prescription to be conflict-free or below a per-prescription threshold. Validation Jaccard selects the utility, pool, and risk-power configuration. The returned rate is recorded; the finite swap search is not a proof that every requested ceiling is feasible. The selected settings transfer to evaluation without using its labels.

Output and ablations. The headline prediction is the learned set after the selected budget procedure. The same run also retains the intermediate calibrated and pre-budget sets. Ranking metrics use refined probabilities; set overlap and conflict metrics use the emitted membership set. Only the complete learned set is reported as DAPS. Appendix E.2 records baseline-specific exceptions rather than assuming all exported arrays were produced by identical preprocessing.

E Experimental Details and Supplementary Analyses

E.1 Data, Evaluation Protocol, and Metrics

Evaluation Protocol and Split Roles. This section records the evaluation protocol behind every result table. Except for quoted rows and explicitly disclosed protocol exceptions, model selection is performed on a fit/validation split and the resulting configuration is frozen before evaluation. Overlap and ranking metrics for multi-visit cases are first averaged within a patient and then across patients; the single-visit TCM corpora reduce to the ordinary per-case mean. DDI/Incompatibility matrices are symmetrised before scoring, and rates count unordered pairs in the predicted set.

Dataset Construction, Preprocessing, and Evaluation Units. MIMIC-III and MIMIC-IV follow the processed diagnosis, procedure, and ATC-3 spaces used by the corresponding backbone build. Patient identifiers define disjoint train, fit, and evaluation partitions. The two TCM corpora are derived from the public structured collection of Liu (2024); neither release exposes practitioner identifiers, so individual cases are the splitting unit with reconstructed histories supplying within-case context rather than verified cross-case patient linkage. We remove leaked patient identifiers, vendor-coded granule formulations, and bulk preparations whose gram values are incompatible with a decoction. Processed forms such as *sheng banxia* and *jiang banxia* remain distinct medications. Same-base pairs used in Section 4 are held out as evaluation annotations and never enter model features.

TCM Safety Knowledge and Held-Out Relation Annotations. Table 18 is computed from the training partition alone; its semantic pair annotations are audit metadata, not learned labels. The TCM safety matrix is an exact-name projection of the classical list: it contains *sheng gancao/haizao*, but does not expand *renshen* to the processed form *hongshen*. The latter pair therefore tests the semantic interpretation of an observational sign, not the operation of the global external-conflict budget.

Training-Only Symptom Vocabulary Construction. For free-text symptoms, ordinary Chinese tokenizers fragment clinical compounds (e.g., *zigong jiliu*, uterine fibroid). We therefore mine cohesive n-grams and reject fragments by neighbour dominance, which is more stable than boundary entropy on templated notes. This changes the onsite symptom vocabulary from $2{,}033$ to $2{,}917$ and adds 0.88 backbone Jaccard before DAPS is applied. Vocabulary induction uses training data only; the resulting dictionary is then frozen for fit and evaluation cases.

Evaluation Metrics and Aggregation. For a target set M_i and prediction S_i, the per-prescription overlap metrics are $J_i = |S_i \cap M_i|/|S_i \cup M_i|$ and $F_{1,i} = 2|S_i \cap M_i|/(|S_i| + |M_i|)$. Empty-denominator cases, which require both sets to be empty, receive zero rather than the value one used in Proposition 1; every DAPS prediction is nonempty. Longitudinal reports average visit-level metrics within each patient and then average across patients; case-based reports use the per-case mean. PRAUC is implemented as per-prescription average precision from the complete score vector, then aggregated by the same evaluation unit. Direct name-generation baselines do not supply a complete candidate score vector and therefore have no reported PRAUC. Mean predicted set size describes decoding behavior and is interpreted against the reference size for the corresponding cohort, rather than optimized as a smaller-is-better metric.

DDI and TCM incompatibility use symmetric external adjacency matrices and count unordered pairs. Their reported rate pools flagged pairs over all predicted pairs, with a zero-pair denominator guarded by one. This differs from averaging per-case pair rates, and it is the pooled quantity used by the budget in Appendix D.3. Percentages in tables multiply rates by 100.

E.2 Backbone Implementations and Reproduction

Western Medication Baselines and Evaluation Protocol. Table 2 uses the MedAlign preprocessing and ATC-3 output space. The split is $2/3$ training, $1/6$ fit, and $1/6$ evaluation by patient. The released MedAlign checkpoint is evaluated directly on MIMIC-III; incompatible or missing checkpoints for the other backbone/build pairs are retrained with their released losses and architectures. The exact exported probability arrays, not rounded table cells, feed both GiantMed and DAPS. VITA retains its published requirement of at least two visits, giving a distinct cohort; absolute scores across backbones are interpreted with this cohort difference in mind. The paragraphs below document checkpoint compatibility and the GiantMed implementation.

Meaning of the provenance labels. "Quoted" means that the table cell is transcribed from a cited paper and no local probability export exists. "Author code" means that we execute a released implementation, with any necessary compatibility patch disclosed below. "Paper-guided" means that no usable implementation was released and we implemented the described architecture ourselves; it is not an exact reproduction claim. "Adapted" means that the original modality or supervision does not exist in the target corpus and a stated analogue is used. Only locally executed rows yield frozen probabilities to which GiantMed or DAPS can be attached.

Table 5: **Implementation provenance for Western-EHR methods.** Methods and plug-ins in Tables 2 and 8 are documented with departures.

Method	Provenance	Implementation and departures
RETAIN; LEAP; GAMENet; MICRON	Quoted	Cells transcribed from GiantMed's comparison table (Lv et al., 2026) are used for context because exact fit/evaluation marginal exports under our split are unavailable. The architectural roles are respectively reverse-time attention, sequence generation, graph memory, and medication-change modeling. No DAPS result is inferred from their rounded cells.
COGNet	Quoted	Official code is publicly available. The row in Table 8 retains all five metrics from the same comparison reported by Lv et al. (2026); we do not mix those published cells with validation metrics or an incomplete local evaluation.

Method	Provenance	Implementation and departures
HiRef	Quoted	Jaccard, PRAUC, and F_1 are quoted from the paper, which releases no code and omits DDI and set size; the unreported diagnostics remain blank rather than being mixed with a separate implementation.
MedAlign	Author code / checkpoint	On MIMIC-III, renaming the checkpoint keys for the query and medicine encoders makes the released weights load and gives 53.89 Jaccard versus 53.90 reported. MIMIC-IV has no compatible released checkpoint and is retrained with the released architecture and objective on the common build.
DEPOT	Author code, minimally patched	The checkpoint contains 13 surplus tensors from another revision and is loaded non-strictly with no missing tensor. The released forward pass indexes a five-field previous visit as if it had three fields and reads the admission identifier as medication history; truncating the record to its three model inputs is the minimal semantic repair.
MoleRec	Author code, auxiliary inputs rebuilt	The release omits `idx2SMILES.pkl` and `substructure_smiles.pkl`. We regenerate the drug–SMILES mapping from the common ATC build. In the released `use_embedding=True` path, substructure graphs are never read, so only a correctly sized placeholder list is supplied; no unrecoverable ordering is invented.
SafeDrug	Author code, retrained	The shipped checkpoint has a $1958/1430/112$ vocabulary and cannot load against either common output space. We retrain it. Its MPNN embedding is a plain attribute absent from the `state_dict`; reseeding immediately before model construction recreates the embedding used during training and makes repeated checkpoint evaluation deterministic.
VITA	Author code, configurable constants	Diagnosis offsets, padding widths, and paths hard-coded for one MIMIC build are parameterized without changing the model. We retain the published requirement of at least two visits, which yields a distinct cohort. Gradient clipping at 1.0 is used because the released loop diverges to NaN in our environment.
GiantMed	Released plug-in, rerun	We run its boundary-adjudication pipeline on the exact frozen arrays used by DAPS, using Qwen3-8B. For VITA we additionally export the per-step maximum required to create a nonempty boundary band; the native step-mean score is retained for VITA and DAPS. The published and rerun MedAlign results differ, so the manuscript reports our local rerun and audits the gap in Appendix E.2.
Prompted LLMs	Our evaluation	Qwen3-8B, Qwen3-32B, Qwen3.5-27B, DeepSeek-V4-Flash, GLM-4.7-Flash, LLaMA3.1, and Aloe-Beta receive the same structured ICD/ATC prompt and are evaluated on the same deterministic 1,000-visit sample. They produce names rather than complete marginals, so PRAUC is undefined and no post-hoc decoder is attached.

Western Checkpoint Compatibility and Cohort Definitions. The released MedAlign checkpoint is compatible with the MIMIC-III preprocessing used here; the other backbone/build pairs require the retraining or compatibility adjustments listed in Table 5.

VITA retains its published requirement of at least two visits, giving $5{,}441$ patients and $14{,}122$ visits on MIMIC-III, compared with $6{,}350$ patients and $15{,}031$ visits in the general processed cohort. Decoder comparisons within each backbone use its corresponding exports and cohort; absolute comparisons across backbones should account for this population difference.

GiantMed Implementation and Matched Evaluation. Loading the released MedAlign checkpoint through its evaluation path gives Jaccard 53.89, close to the published 53.90. Our GiantMed refinement run on those predictions gives 53.83, compared with the published 56.63. The locally executed backbone and plug-in are scored on the same split, and Table 2 reports that matched run. The published refinement figure is provided as a separate numerical reference, not as a locally reproduced result.

The local refinement rerun retains the released boundary settings $\delta = 0.5$, $\epsilon = 0.2$, top-3 retrieval, and generation configuration. We supplied probability export, repaired the retrieval script's undefined variable, and implemented response merging and scoring where the released pipeline did not provide runnable steps. The authors confirmed diagnosis-only Jaccard retrieval and identified the middle sixth as their refinement split. They also reported that incomplete boundary responses were introduced in a later code reorganization and were absent from the version used for the paper. These clarifications are summarized with their permission.

Additional checks did not resolve the numerical difference. Greedy decoding gives 53.83 versus 53.81 at the published temperature; seven merge variants give 52.50–53.91; and Qwen3-32B gives 53.85. All $2{,}279$ responses were parsed, with 74 omitted decisions among $17{,}041$ boundary decisions (0.43%). On the author-identified split, refinement gives 53.51 versus a 53.63 backbone. The released backbone trainer evaluates the final sixth, whereas the authors identified the middle sixth for refinement; whether the published backbone reference was re-scored on that same split remains unconfirmed. We therefore base the decoder comparison on our matched local runs and do not infer an error in the published result from this discrepancy. For reference, MedAlign with DAPS reaches 56.91 on the matched run, which is not below the published GiantMed figure of 56.63 either, although the two numbers are not evaluated under a verified common split.

Adapting Western Backbones to TCM. The adapted-backbone block in Table 3 is deliberately a domain-transfer stress test, not a claim that Western drug representations are native TCM models. Symptoms replace diagnosis/procedure indicators and herbs replace ATC-3 outputs; training, fit, and evaluation roles otherwise remain unchanged. On TCM, MedAlign preserves its ID/structure/text three-view alignment and entropic optimal-transport fusion. Its molecular path becomes herb $\rightarrow$ classical formula $\rightarrow$ formula composition, and its drug-leaflet embedding becomes TF–IDF/SVD of descriptions from the TCM knowledge corpus. A TCM case is one consultation, so its longitudinal GRU collapses to one batched step.

The Western-to-TCM block in Table 3 uses the same implementation and DAPS settings. Herbs have no single molecular Simplified Molecular Input Line Entry System (SMILES) representation. For DEPOT, MoleRec, and SafeDrug, the molecular-only global encoder is therefore replaced by a projection of the available herb-by-classical-formula membership matrix; herbs with no formula membership receive singleton columns, as detailed below.

For DEPOT, MoleRec, and SafeDrug this substitution changes only the molecular/global representation branch: the visit encoder, objective, and decoder remain those of the Western model. The formula-membership substitution is triggered by a constant molecular embedding rather than by the dataset name. For the onsite vocabulary, 136 of 329 herbs occur in no retained classical formula; each receives a singleton column so a mask-gated output layer can still emit it. For VITA, the change is instead temporal. The internet corpus contains a prescription at every reconstructed visit, whereas the onsite source contains longitudinal symptoms but only one record-level prescription. Thus onsite visit selection is meaningful, but medication reuse from a selected past visit is unavailable and is disabled rather than filled with the current label.

Table 6: **TCM adaptations of the local Western backbones in Table 3.**

Backbone	TCM adaptation
MedAlign	Retains the ID/structure/text alignment and optimal-transport fusion; replaces SMILES/BRICS with herb/formula/composition and drug-leaflet text with TF–IDF/SVD herb descriptions. The one-case visit encoder is batched.

Backbone	TCM adaptation
DEPOT	Retains its debiased temporal predictor and training loss. Its drug-structure input receives the formula-membership projection with singleton columns for herbs absent from retained formulas.
MoleRec	Retains its local compositional and medication-scoring modules. The global molecular encoder alone is replaced by the formula-membership projection; placeholder SMILES are not treated as biological structure.
SafeDrug	Retains its dual-branch combination and incompatibility objective. The constant molecular branch is replaced by the same formula projection, while the formula mask continues to gate herb outputs.
VITA	Uses reconstructed patient visits and retains visit selection, weighting, and the published step-mean probability convention. On onsite data, only symptom history is visit-specific, so prior-medication reuse is unavailable.

TCM-Native Baseline Implementations. The eight literature-derived TCM encoders share one PyTorch trainer so their rows differ in the encoder rather than in optimization. Inputs and targets are multi-hot sets; the loss is binary cross-entropy (BCE; plus the published contrastive auxiliary loss for TCMRGCL); Adam uses learning rate 10^{-3}, batch size 256, and no weight decay. Training allows at most 60 epochs and stops after ten epochs without improvement in validation P@5, retaining the best validation checkpoint. Embedding width is 64 at input and the shared graph stack uses 128/256 hidden units. Symptom and herb graphs are constructed from the training split only, with co-occurrence thresholds 5 and 40; SMRGAT uses its published unthresholded graphs. PRDAGE's order permutations act as rare-label reweighting. When a released corpus lacks herb attributes or symptom embeddings, the corresponding learned table is used and the missing resource is not synthesized.

Table 7: **Method-by-method implementation of the TCM baselines.** "Full text" and "abstract" describe the source available for our implementation, not the quality of the underlying paper or its reported results.

Method	Basis	Implemented components and unavoidable departures
SMGCN	Full text; paper-guided	Separate symptom and herb graph convolutional networks (GCNs), a bipartite MLP, and additive syndrome fusion.
KDHR	Full text and released code; common-trainer implementation	Adds herb attributes and multi-layer graph fusion to SMGCN. If a new corpus has no aligned attribute matrix, the attribute branch is absent and a learned herb table is used; the missing attributes are not inferred from herb names.
SMRGAT	Full text; paper-guided	Residual graph attention is applied to the symptom, herb, and bipartite relations. Following the paper, its graphs are unthresholded. The original 23-dimensional herb properties and symptom word2vec vectors are unavailable for the two TCM corpora, so learned embeddings replace only these initial features.
SCEIKG	Full text; partial adaptation	The attention-based heterogeneous graph propagation and published mutual-exclusion regularizer are retained. Its condition-transition module requires a sequence of labeled pre/post-treatment visits and BERT condition text; the implementation retains the static portion when these resources are unavailable.
PresRecST	Abstract; paper-guided interpretation	The described progressive syndrome $\rightarrow$ treatment-method $\rightarrow$ herb structure is implemented as two latent residual refinement stages. The self-curated knowledge graph and intermediate syndrome/treatment labels are not released with our benchmarks. This is our operational reading of the abstract, not a reproduction.

Method	Basis	Implemented components and unavoidable departures
TCMRGCL	Full text; paper-guided	Two augmented bipartite views (truncated singular value decomposition (SVD) and edge dropout), shared GCN propagation, and an InfoNCE auxiliary objective are implemented. The paper pretrains and freezes the contrastive encoder; to isolate encoder differences under the common protocol, we optimize its contrastive term jointly with prediction and report that departure.
PRDAGE	Full text; paper-guided	Retains the three graph views, long short-term memory (LSTM) prediction head, BCE objective, and rare-label herb-order augmentation. Sentence-BERT symptom vectors and herb property vectors are used only when aligned resources exist; otherwise learned tables replace those semantic inputs.
FMASH	Full text; paper-guided	Implements graph propagation followed by degree-sort, bidirectional sequence modeling, and unsorting. Bidirectional Mamba is used when `mamba_ssm` is available; otherwise the recorded backend is a bidirectional gated recurrent unit (GRU). No evaluated herb vocabulary has aligned UniMol features, so the molecular half of multi-level fusion is absent.
KGETM	Abstract; diagnostic implementation, not tabled	Without the paper's external knowledge graph, the described model reduces to a low-rank symptom–topic–herb factorisation. Because that removes the component that distinguishes KGETM, we retain the implementation for diagnosis but do not present it as a faithful benchmark row.
PresRecRF	Full text; diagnostic implementation, not tabled	Its herb–symptom–protein graph, BERT/ChatGPT semantics, and dose target are absent. Only the representation-fusion shell and herb-prediction objective survive; we therefore do not label that reduced model a PresRecRF reproduction in the result tables.
Zero-shot LLMs	Our evaluation	GPT-5.6-sol, Claude Opus 5, DeepSeek-V4-Flash, GLM-4.7-Flash, Qwen3.5-27B, Qwen3-32B, Qwen3-8B, ShenNong-TCM, and HuatuoGPT2-7B receive the same structured symptoms, frozen candidate vocabulary, and training-set interquartile size guidance. API and local models use the same 1,000 evaluation cases.
Fine-tuned LLMs	TCM-FTP recipe, rerun	ShenNong-TCM, HuatuoGPT2-7B, the Qwen models, GLM-4.7-Flash, and DeepSeek-V4-Flash are trained with the common TCM-FTP-style answer-only LoRA recipe below. These are applications of the recipe to current base models, not reproductions of TCM-FTP's original base checkpoints.

TCM Split and Decoding Protocol. For Table 3, native rankers decode at the rounded training-split mean prescription length (15 on TCM-Onsite, 12 on TCM-Internet) so set metrics are comparable across rankers; PRAUC still uses the complete score vector. These rankers define no set-size rule. Thresholding their scores at 0.5 instead selects 8.6–12.0 herbs on average with Jaccard 37.8–47.2 on TCM-Onsite, and 3.3–5.2 herbs with Jaccard 14.0–22.5 on TCM-Internet, below the fixed-count row of every backbone in Table 3. The main TCM corpora use training, fit, and evaluation proportions of $2/3, 1/6, 1/6$.

E.3 Language-Model Evaluation and Result Provenance

Language-Model Sampling, Prompting, and Scoring. Every LLM is scored on the same fixed sample of 1,000 evaluation cases per corpus. Zero-shot models receive identical structured inputs and candidate herb vocabularies, sample at temperature 0.7, and are vocabulary-filtered before scoring. ShenNong omits the full candidate enumeration where its 2,048-token context would otherwise

Table 8: **Literature comparison: published medication-recommendation results.** These rows provide context. RETAIN, LEAP, MICRON, GAMENet, and COGNet are transcribed from the comparison table of Lv et al. (2026); HiRef's overlap and ranking metrics are taken from its paper. These methods were not evaluated under our common split, so the rows are not directly comparable with the matched local experiments in Table 2. A dash denotes an unreported metric from its source.

	MIMIC-III					MIMIC-IV				
Method	JA ↑	F_1 ↑	PRAUC ↑	DDI ↓	# Med.	JA ↑	F_1 ↑	PRAUC ↑	DDI ↓	# Med.
RETAIN (Choi et al., 2016)	48.22	64.58	75.44	8.15	18.95	40.27	55.83	66.11	8.09	10.93
LEAP (Zhang et al., 2017)	46.28	63.90	74.56	7.56	18.78	40.09	55.17	58.80	7.24	11.75
MICRON (Yang et al., 2021a)	49.95	65.92	75.24	7.42	18.64	43.64	58.16	67.26	7.19	13.31
GAMENet (Shang et al., 2019)	49.14	65.03	73.98	8.01	24.36	43.05	57.97	67.54	7.92	15.26
COGNet (Wu et al., 2022)	51.34	67.06	76.77	7.84	24.17	41.31	56.60	64.60	5.96	19.40
HiRef (Chok et al., 2025)	48.32	63.90	73.78	–	–	49.89	65.00	75.92	–	–

truncate the case. Direct generations provide ranked names but no calibrated marginal for every herb, so their PRAUC is reported as not applicable.

Sample Size and Data-Use Constraints. We restrict language-model evaluation to this fixed sample because generative inference is far more expensive than decoding stored marginals: every case requires a long structured prompt with the candidate vocabulary, the evaluation covers more than a dozen zero-shot and fine-tuned models, and TCM-Internet alone has 51,051 test cases. The sample is drawn once and shared by all models, and the frozen backbone and its DAPS decoding are re-scored on the same cases (Tables 19 and 20). Their sample Jaccard lies within 0.9 points of the corresponding full-test values, and the DAPS gain over the backbone keeps its sign on all four corpora. Data-use terms further determine which models are evaluated. The PhysioNet credentialed data use agreement for MIMIC-III/IV prohibits sharing records with third parties, including submission to external online services, so MIMIC visits are processed only by open-weight models running on local hardware. The TCM corpora are publicly released, so for them we additionally prompt the API models GPT-5.6-sol and Claude Opus 5 in the zero-shot setting.

Language-Model Fine-Tuning and Execution Exceptions. Fine-tuning follows TCM-FTP: two epochs, two herb-order permutations, LoRA rank 16 with $\alpha = 32$, learning rate 10^{-4}, answer-only loss, and maximum sequence length 768. Evaluation is greedy on the same 1,000 cases. DeepSeek-V4 uses quantized low-rank adaptation (QLoRA) and a deterministic 1,000-case training subset because the full native-quantized backward pass is computationally prohibitive. Qwen3.5-27B and GLM-4.7-Flash are evaluated through native parameter-efficient fine-tuning (PEFT) because the vLLM mixed-attention adapter path did not reproduce the native model's behavior in a matched audit.

E.4 DETAILED MAIN COMPARISONS

Detailed Analysis of the Western Main Comparisons. Table 2 contains five matched backbone pairs on each Western corpus. Their mean Jaccard gains are 5.61 points on MIMIC-III and 5.93 on MIMIC-IV. These means summarize the evaluated pairs; they do not constitute a single cohort-controlled comparison among all encoders, because VITA retains its multi-visit filter. On the MedAlign pairs, the gains are $3.02/4.04$ Jaccard points and $3.23/4.54$ PRAUC points. The simultaneous improvement in set overlap and ranking motivates the separate analysis in Appendix E.8, while the component tables below distinguish calibration, refinement, and budget effects under the same frozen exports.

Language-model results in Appendix Tables 19 and 20 use a deterministic sample rather than the full backbone evaluation cohort; their before/after fine-tuning comparison is matched within that sample. Table 8 separately supplies published literature context rather than locally paired decoder evaluations, and Appendix E.2 identifies these provenance differences and details the GiantMed evaluation. Among the published results in Table 8, the highest Jaccard is 51.34 (COGNet) on MIMIC-III and 49.89 (HiRef) on MIMIC-IV, and the reported DDI rates span 5.96–8.15. The matched MedAlign–DAPS configuration reaches 56.91 and 55.24 Jaccard with DDI 6.93 and 6.53

under our split. Because the published rows use their own preprocessing and splits, this places DAPS in a competitive range rather than establishing a controlled margin over those methods.

Detailed Analysis of the TCM Main Comparisons. Table 3 separates eight native TCM encoders from five Western architectures adapted to herb recommendation. The native group uses fixed backbone counts of 15 and 12 on TCM-Onsite and TCM-Internet; its DAPS gains average $3.86/3.27$ Jaccard points. The adapted group includes both under-predicting and over-predicting baseline sets and gains $11.36/9.69$ points on average. These are architectures trained on the corresponding TCM data, not unchanged Western checkpoints applied to herb labels. Their representation substitutions and history construction are detailed in Appendix E.2.

The two main tables contain $10 + 16 + 10 = 36$ matched backbone–dataset pairs. All aggregate claims and figures use exactly these 36 pairs.

E.5 DECODER ABLATIONS, UNCERTAINTY, AND RUNTIME

Ablation Study: Protocol and Complete Five-Sub-Step Metrics. We use MedAlign on MIMIC-III/IV and PresRecST on both TCM datasets. We select MedAlign because it provides a strong, reproducible EHR backbone with aligned candidate-level probability exports and a released checkpoint compatible with our MIMIC-III preprocessing. We select PresRecST as a representative domain-native TCM model that explicitly follows the syndrome–treatment-method–herb pipeline and can be evaluated consistently on both TCM corpora. These backbones are selected for reproducibility and domain representativeness, rather than for the magnitude of their test-set DAPS gains. Within each dataset, all six rows in Table 9 share the same frozen backbone export and evaluation cohort, and each cell averages three seeds (1203, 1204, 1205) of the complete decoder, which reseeds the tree experts, both calibrators, and the dependency models; the frozen backbone row has no seed variance. Sub-step 1a cross-fits the calibrator g_1 and keeps the backbone's own set size; 1b selects the provisional count $k^{(0)}$ from $\tilde{p}^{(0)}$; 2a recalibrates with g_2 given context built from the step-1 set $S^{(0)}$, not the original backbone's set, while keeping $k^{(0)}$; 2b reselects $k^\star$ from the refined scores; and step 3 applies size-preserving swaps under the global external-conflict budget. No test label enters model, stage, or hyperparameter selection.

Table 9 reports every set metric exported by these analysis runs. They do not export rank-only PRAUC, which is not inferred from binary sets or borrowed from differently configured runs. The TCM backbone rows use fixed counts of 15/12 herbs; all stage-wise gains are computed against the backbone row of the same controlled run rather than rounded main-table cells.

Across the four datasets, steps 1–2 improve overlap mainly through recall, which rises by 4.36, 7.85, 4.10, and 1.81 points over the backbone (TCM-Onsite, TCM-Internet, MIMIC-III, MIMIC-IV). Precision rises on TCM-Onsite, MIMIC-III, and MIMIC-IV ($+1.26$, $+1.25$, and $+5.38$) but falls slightly on TCM-Internet (-0.97). The calibration sub-steps leave the size unchanged by construction; sizing is not a uniform push toward larger sets: DAPS enlarges the fixed TCM sets from 15 to 15.86 and from 12 to 15.22 herbs and MIMIC-III sets from 20.22 to 21.00, but shrinks MIMIC-IV sets from 15.59 to 14.68, where the precision gain is largest. Steps 1 and 2 also raise the listed conflict rate on every dataset, most visibly from 6.59 to 8.63 on MIMIC-IV, because higher-overlap sets contain more listed pairs. Step 3 then lowers the rate while leaving every set size unchanged, which isolates the conflict adjustment from the overlap gains of the earlier steps.

Shuffled provisional-set control. Step 2 adds set context and the training-partition models of Appendix D.2, whose outputs also carry case-level information that does not depend on $S^{(0)}$. To separate the two, we rerun step 2 with every provisional set ($S^{(0)}$ and the second-round $S^{(1)}$) replaced by the provisional set of a randomly permuted case of the same split, keeping all models, features, and capacity unchanged. Table 10 reports the step-2 gain under both inputs. The part that depends on each case's own set is $+0.43$ and $+0.38$ Jaccard points on TCM-Onsite and TCM-Internet, and $+0.16$ and $+0.07$ on MIMIC-III and MIMIC-IV; across all backbones it averages $+0.58$, $+0.35$, $+0.02$, and $+0.06$. The remainder of the step-2 gain is case-level information supplied by the training-partition models. The set-dependent part is thus largest on the TCM corpora, whose prescriptions follow formula structure, and small on MIMIC.

Table 9: **Ablation study: complete five-sub-step set metrics.** All sub-steps use the same frozen backbone export and evaluation cohort. Values are mean ±s.d. over three seeds of the complete decoder (seeds 1203, 1204, 1205), so the frozen backbone rows have zero s.d.; metrics are test percentages except mean set size. Conflict is DDI on MIMIC and incompatibility on TCM, shown to four decimals for TCM-Internet. All sub-steps are cumulative; the shaded row 1a–3 is full DAPS.

Configuration	JA	F_1	Precision	Recall	Conflict	Set size
PresRecST / TCM-Onsite						
Backbone	49.05±.00	63.71±.00	63.64±.00	64.60±.00	1.35±.00	15.00±.00
+ 1a (calibration)	49.95±.04	64.66±.03	64.58±.03	65.56±.04	1.49±.00	15.00±.00
+ 1a–1b (provisional sizing)	50.46±.02	65.16±.02	62.84±.23	68.90±.28	1.38±.00	16.36±.13
+ 1a–2a (set-cond. calibration)	51.60±.02	66.07±.00	63.72±.24	69.85±.28	1.33±.01	16.36±.13
+ 1a–2b (final sizing)	51.97±.02	66.30±.02	64.90±.07	68.96±.06	1.36±.01	15.86±.03
+ 1a–3 (conflict budget)	51.97±.02	66.30±.02	64.90±.07	68.96±.06	1.32±.00	15.86±.03
PresRecST / TCM-Internet						
Backbone	29.91±.00	43.76±.00	44.42±.00	45.83±.00	0.0025±.0000	12.00±.00
+ 1a (calibration)	30.42±.02	44.33±.02	45.00±.02	46.42±.02	0.0038±.0002	12.00±.00
+ 1a–1b (provisional sizing)	31.36±.02	45.51±.02	42.06±.04	52.98±.01	0.0039±.0001	15.35±.01
+ 1a–2a (set-cond. calibration)	32.13±.02	46.25±.02	42.72±.04	53.87±.01	0.0041±.0002	15.35±.01
+ 1a–2b (final sizing)	32.49±.02	46.53±.03	43.45±.08	53.68±.07	0.0044±.0002	15.22±.05
+ 1a–3 (conflict budget)	32.49±.02	46.53±.03	43.45±.08	53.68±.07	0.0027±.0000	15.22±.05
MedAlign / MIMIC-III						
Backbone	53.89±.00	69.18±.00	69.10±.00	71.89±.00	7.03±.00	20.22±.00
+ 1a (calibration)	56.16±.02	71.11±.02	70.99±.02	73.92±.03	7.86±.03	20.22±.00
+ 1a–1b (provisional sizing)	56.82±.02	71.65±.02	69.92±.16	75.95±.21	7.95±.02	21.09±.11
+ 1a–2a (set-cond. calibration)	57.05±.06	71.84±.05	70.10±.17	76.15±.23	8.00±.02	21.09±.11
+ 1a–2b (final sizing)	57.12±.11	71.90±.10	70.35±.14	75.99±.14	8.02±.04	21.00±.04
+ 1a–3 (conflict budget)	56.98±.10	71.79±.09	70.23±.14	75.89±.12	6.93±.00	21.00±.04
MedAlign / MIMIC-IV						
Backbone	51.20±.00	66.24±.00	63.64±.00	72.58±.00	6.59±.00	15.59±.00
+ 1a (calibration)	53.93±.01	68.62±.01	65.91±.01	75.20±.02	8.14±.01	15.59±.00
+ 1a–1b (provisional sizing)	54.86±.01	69.36±.01	68.29±.02	73.77±.02	8.56±.01	14.67±.01
+ 1a–2a (set-cond. calibration)	55.34±.01	69.78±.01	68.70±.02	74.24±.02	8.60±.00	14.67±.01
+ 1a–2b (final sizing)	55.68±.02	70.07±.02	69.02±.03	74.39±.03	8.63±.01	14.68±.01
+ 1a–3 (conflict budget)	55.27±.03	69.72±.03	68.69±.01	74.01±.06	6.53±.00	14.68±.01

Table 10: **Shuffled provisional-set control for step 2.** Jaccard gain of step 2 over step 1 when the set context is built from the case's own provisional set (*own set*) or from the provisional set of a randomly permuted case of the same split (*shuffled set*); models, features, and capacity are unchanged. Their difference is the part of the gain that depends on the case's own set. Main pairs: mean ±s.d. over three seeds; last column: mean over all backbones of the dataset (seed 1203; number of backbones in parentheses).

Dataset	Backbone	Own set	Shuffled set	Set-dependent	All backbones
TCM-Onsite	PresRecST	+1.51±.03	+1.07±.03	+0.43±.01	+0.58 (13)
TCM-Internet	PresRecST	+1.14±.02	+0.76±.01	+0.38±.01	+0.35 (13)
MIMIC-III	MedAlign	+0.30±.12	+0.14±.08	+0.16±.09	+0.02 (5)
MIMIC-IV	MedAlign	+0.82±.02	+0.74±.03	+0.07±.01	+0.06 (5)

Table 11: **Dataset descriptives behind the step-2 gap.** Training-split statistics of the recorded prescriptions (rows 1–6) and fit-split residual dependence after step-1 calibration (rows 7–8; the pairwise correlation of $\mathbf{m} - \tilde{p}^{(0)}$ relative to that of $\mathbf{m}$, over outputs with frequency at least 2%). Positive PMI is averaged over pairs inside each prescription; the covariance ratio averages, over the 50 most frequent input codes and the 30 most frequent outputs, the mean absolute pairwise covariance given the code relative to the unconditional one. No test data are used.

Statistic	TCM-Onsite	TCM-Internet	MIMIC-III	MIMIC-IV
Output labels K (herbs; ATC-3 classes)	329	720	131	127
Mean positive PMI within a prescription	0.23	0.41	0.11	0.19
Common outputs with a partner of corr < -0.05	99%	92%	45%	73%
Pairwise covariance kept given one frequent input code	1.00	1.04	0.91	0.87
Visits with a previous prescription	–	–	53%	58%
Current medications already in the previous visit	–	–	63%	66%
Pairwise $\lvert\mathrm{corr}\rvert$ kept after step-1 calibration	0.65	0.73	0.45	0.28
Pairs with residual $\lvert\mathrm{corr}\rvert > 0.2$ after step 1	2.7%	1.7%	0.1%	0.1%

Table 12: **Uncertainty analysis: paired bootstrap of DAPS gains.** With 10,000 resamples, we report full DAPS (steps 1–3, as in the main tables) minus its matched backbone in Jaccard percentage points, with a percentile 95% confidence interval. Resampling uses patients for MIMIC and cases for TCM, which lacks reliable patient IDs.

Backbone	Dataset	Base	DAPS	Δ Jaccard [95% CI]	Units
PresRecST	TCM-Onsite	49.05	51.95	+2.90 [+2.77, +3.02]	15,466 cases
PresRecST	TCM-Internet	29.91	32.49	+2.57 [+2.50, +2.64]	51,051 cases
MedAlign	MIMIC-III	53.89	56.91	+3.02 [+2.71, +3.33]	1,059 patients
MedAlign	MIMIC-IV	51.20	55.24	+4.03 [+3.87, +4.20]	8,442 patients

Why the set-conditioned step contributes more on TCM. Step 2 adds $+1.51$ and $+1.13$ Jaccard points on TCM-Onsite and TCM-Internet but $+0.30$ and $+0.82$ on MIMIC-III and MIMIC-IV (Table 4). Table 11 collects descriptive statistics of the four corpora that are consistent with this gap; they are computed on training or fit data only and do not enter the method. Three properties of the EHR corpora stand out. First, the presentation already explains much of the co-prescription structure: more than half of the MIMIC visits carry a previous prescription, about two thirds of the current medications already appear in it, and conditioning on a single frequent diagnosis or procedure code removes 9–13% of the pairwise medication covariance, whereas conditioning on a single symptom leaves the herb covariance unchanged. Accordingly, the calibrated marginals of step 1 retain only 28–45% of the pairwise dependence among medications on MIMIC but 65–73% among herbs, and pairs with residual correlation above 0.2 are more than ten times rarer on MIMIC. Second, the EHR labels are level-3 ATC classes, which merge interchangeable drugs of one class into a single label; the exclusion structure that remains between classes is weaker (45–73% of common classes have a negatively correlated partner, versus 92–99% of common herbs). Third, herb prescriptions carry stronger within-set co-selection (positive PMI about twice that of MIMIC), consistent with formula-based prescribing that is not determined by individual symptoms. The shuffled-set control above points the same way: the part of the step-2 gain that depends on the case's own set is $+0.38$ to $+0.43$ on TCM and at most $+0.16$ on MIMIC. On the EHR corpora, the larger share of the total gain comes from calibration (step 1a adds $+2.27$ and $+2.73$), whose scores the frozen backbones leave less well calibrated; once calibrated given the presentation, little dependence is left for set context to explain. These statistics describe the corpora as processed here and do not identify the clinical mechanisms behind either prescribing pattern.

Uncertainty of matched improvements. Because full DAPS is applied to exactly the same frozen test predictions as each backbone, we quantify uncertainty with paired rather than independent resampling. Table 12 reports patient-level bootstrap intervals for MIMIC and case-level intervals for TCM. The latter are necessarily weaker: the released TCM data do not provide reliable patient identifiers, so these intervals cannot account for repeated prescriptions from the same patient. All four 95% intervals exclude zero; across these representative pairs, the estimated gains range from 2.57 to 4.03 Jaccard points and equal the main-table gains of the same runs.

Table 13: **Runtime analysis: DAPS on Western and TCM backbones.** Fit time includes expert and feature construction, both calibrators, the training-partition dependency models, and validation-only cardinality-policy and conflict-budget selection. Application time covers full frozen evaluation-split decoding. Both use 32 CPU workers and one GPU for the dependency models.

Backbone	Dataset	Fit cases	Test cases	Fit time	Apply time	ms/case
MedAlign	MIMIC-III	2,264	2,279	5.28 min	8.99 s	3.94
PresRecST	TCM-Internet	51,050	51,051	222.93 min	2300.04 s	45.05

Table 14: **Training-strategy comparison and joint-model ablations: complete test metrics.** Post-hoc DAPS is compared with four joint configurations on each MIMIC dataset. Metrics are percentages except mean medication count. Joint variants remove typed relation channels, the auxiliary expert head, or both; plain retains learned sizing and is not native MedAlign.

Configuration	JA	F_1	PRAUC	Precision	Recall	DDI	# Med.
MIMIC-III							
Post-hoc DAPS	56.91	71.72	81.66	70.22	75.76	6.93	20.98
Joint: full	52.97	68.39	77.50	68.36	70.69	8.22	20.09
Joint: no relations	53.11	68.51	77.50	67.91	71.41	8.50	20.41
Joint: no expert	53.79	69.14	78.35	68.82	71.74	8.47	20.15
Joint: neither (plain)	53.97	69.27	78.39	67.87	73.03	8.50	20.87
MIMIC-IV							
Post-hoc DAPS	55.24	69.70	79.71	68.69	73.97	6.53	14.67
Joint: full	52.32	67.11	76.03	66.90	70.31	8.80	14.35
Joint: no relations	52.29	67.08	76.01	66.42	70.74	8.66	14.55
Joint: no expert	51.81	66.67	75.21	66.81	69.25	8.85	14.01
Joint: neither (plain)	51.92	66.79	75.39	66.61	69.73	8.83	14.18

Runtime. Table 13 separates one-time fitting and validation-only selection from deployment. Once fitted, DAPS decodes a MIMIC-III prescription in 3.94 ms and a 720-label TCM-Internet prescription in 45.05 ms on average. The corresponding peak process memory is 1.77 and 57.01 GiB, respectively; the latter reflects the substantially larger corpus and label space. Fitting dominates the cost: it takes 5.28 minutes on MIMIC-III and 222.93 minutes on TCM-Internet, which has 22.5 times as many fit cases and 720 candidate herbs. This is a one-time cost per backbone and dataset, whereas deployment stays in milliseconds per case and requires no backbone retraining.

E.6 Training Strategy and Calibrator Families

Post-hoc versus Joint Training. Table 14 gives all seven test metrics underlying Figure 4. The post-hoc configuration freezes MedAlign and fits tree experts and calibrators separately. The full joint configuration trains from scratch with a trainable auxiliary head and typed relation channels. Its count predictor is supervised on training data, and test sets are decoded by top-k at the rounded predicted count. Joint training uses a DDI-dependent loss schedule, whereas post-hoc DAPS applies a fit-selected pooled conflict budget through size-preserving swaps. Checkpoints are selected on fit data, not test labels. These are complete-configuration comparisons: the auxiliary architecture, objectives, sizing and conflict-control rules, and optimization differ. The results therefore do not isolate execution order or freezing alone. Figure 4 compares the full configurations; Table 14 additionally reports three joint-model ablations on each dataset. The *no relations* variant removes typed relation channels, *no expert* removes the trainable auxiliary head, and *plain* removes both. Each variant is trained separately with the same dataset-specific training settings and fit-based checkpoint selection. All retain the learned count predictor and joint training objective, including DDI-dependent loss control; plain is therefore not the native MedAlign baseline. These are component ablations within the joint architecture, not additional training strategies or cumulative stages of the post-hoc decoder.

The joint-component results are dataset-dependent. On MIMIC-III, plain has the highest joint-model Jaccard (53.97), whereas the full joint model has the lowest joint-model DDI (8.22). On MIMIC-IV, the full joint model has the highest Jaccard (52.32), with no relations close behind (52.29). Thus adding the auxiliary head and relation channels does not uniformly improve joint-model overlap.

Table 15: **Ablation study: calibrator-family sensitivity of DAPS.** Both calibrators use the listed family; the LightGBM rows are the main-table configuration. Families without an initial-score interface (random forests, Extra Trees, logistic regression, MLP) receive the previous round's logit as an input feature instead of as a boosting offset. Scores and conflict rates are percentages; set sizes are counts. Conflict denotes incompatibility on TCM and DDI on MIMIC. Bold marks the largest unrounded Jaccard in each setting.

Calibrator	JA ↑	F_1 ↑	PRAUC ↑	Conflict ↓	Set size
PresRecST / TCM-Onsite					
LightGBM	51.95	66.29	73.45	1.32	15.90
XGBoost	51.87	66.21	73.39	1.32	15.89
Random forest	51.88	66.25	73.49	1.32	16.04
Extra Trees	51.48	65.91	73.30	1.32	15.60
Logistic regression	51.80	66.18	73.28	1.32	15.89
MLP	**51.96**	66.28	73.41	1.32	15.69
MedAlign / MIMIC-III					
LightGBM	**56.91**	71.72	81.66	6.93	20.98
XGBoost	56.90	71.72	81.53	6.93	21.09
Random forest	56.58	71.47	81.48	6.93	20.84
Extra Trees	56.57	71.47	81.36	6.93	20.88
Logistic regression	56.58	71.48	81.38	6.93	20.69
MLP	56.78	71.64	81.44	6.93	21.25

Post-hoc DAPS achieves higher Jaccard and F_1 and lower DDI than every evaluated joint variant on both datasets. This comparison supports the reported modular configuration, while the differing objectives do not establish a universal advantage of post-hoc training.

Scope and controlled configuration. We replace the learner used by both cross-fitted calibrators with LightGBM, XGBoost, random forests, extremely randomized trees (Extra Trees), logistic regression, or a multilayer perceptron (MLP), using PresRecST on TCM-Onsite and MedAlign on MIMIC-III. The batch fixes the backbone exports, data partitions, base seed, expert bank, relation construction, and cardinality and safety search procedures within each setting. The latter procedures are rerun on each learner's fit predictions, so selected counts and swap configurations can differ. This tests sensitivity of the complete decoder to the calibrator implementation, rather than changing the backbone or freezing the final decisions across learners.

LightGBM treats candidate identity as categorical; the other tree learners use its integer encoding. Logistic regression and MLP omit that column and standardize the remaining features with statistics fitted on their training fold. Consequently, the comparison includes these learner-specific input adaptations and should not be interpreted as an architecture-only comparison with identical feature encodings. Tree ensembles use 200 estimators; LightGBM and XGBoost use learning rate 0.035, and the forest variants use minimum leaf size 200. Logistic regression uses $C = 1$ and at most 400 iterations. The MLP uses hidden widths $(64, 32)$, at most 40 epochs, and internal early stopping. These fixed configurations are sensitivity probes, not an equally extensive hyperparameter search for every family.

Completed runs and reporting. Table 15 reports the final refined, budgeted prediction of every family in both settings; all ten non-LightGBM runs completed, and the LightGBM rows are the main-table runs of Tables 2 and 3. Each family replaces both calibrators, including the boosted second round; families without an initial-score interface receive the first-round logit as an input feature instead. The expert bank and dependency models are shared across families.

Interpretation. Jaccard ranges from 51.48 to 51.96 on TCM-Onsite and from 56.57 to 56.91 on MIMIC-III, so the six families span only 0.48 and 0.34 points, and each family stays at least 2.43 and 2.68 points above its backbone (49.05 and 53.89). Logistic regression stays inside both bands, which places the map from candidate features to a calibrated probability close to linear once provisional-set context is supplied. No family wins in both settings: MLP leads narrowly on TCM-Onsite and LightGBM on MIMIC-III. Conflict rates coincide because the fit-selected budget binds in every run,

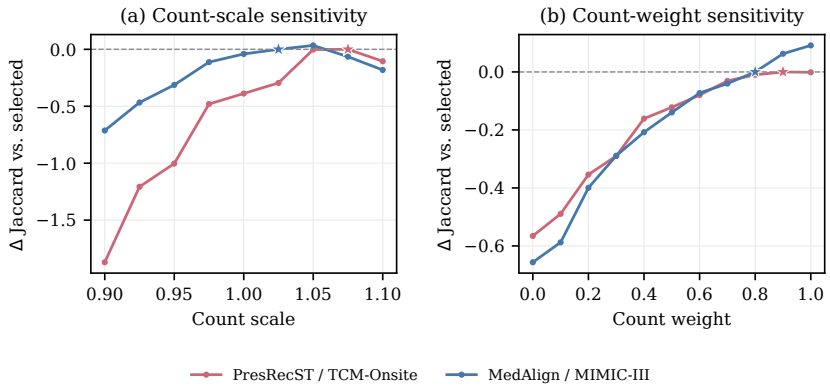


Figure 5: **Hyperparameter sensitivity: count scale and count weight.** Results use MedAlign on MIMIC-III and PresRecST on TCM-Onsite. Frozen-test Jaccard changes for (a) count scale and (b) count weight, relative to the validation-selected settings (stars).

while output counts still differ with each learner's scores. These are descriptive comparisons, not significance tests.

E.7 Cardinality and Conflict-Budget Sensitivity

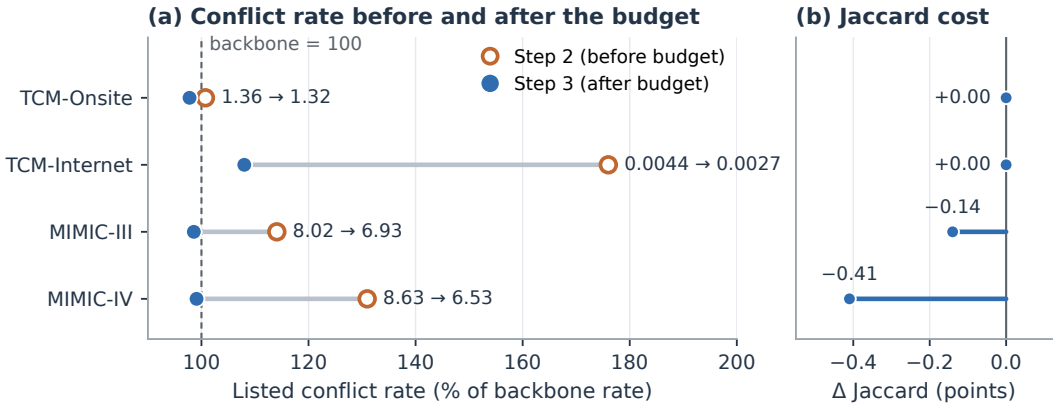


Figure 6: **Budget analysis: global conflict-budget control on all four datasets.** Values are the step-2 and step-3 rows of Table 9. (a) Listed conflict rate (DDI for MIMIC, Inc. for TCM) relative to each backbone's rate, before (hollow) and after (filled) the budget; raw rates are annotated. (b) Jaccard change caused by the same step.

Global External-Conflict Budget Analysis. Relative to step 2 in Table 9, the global external-conflict budget (step 3) lowers the listed conflict rate on all four datasets: from 1.36 to 1.32 on TCM-Onsite, 0.0044 to 0.0027 on TCM-Internet, 8.02 to 6.93 on MIMIC-III, and 8.63 to 6.53 on MIMIC-IV. The final rate falls below the backbone rate on TCM-Onsite and both MIMIC datasets but not on TCM-Internet (0.0025). The Jaccard cost is not visible at the reported precision on either TCM corpus and is 0.14 and 0.41 points on MIMIC-III/IV. The TCM-Internet exception reflects how the budget is set rather than a failed search: the ceiling is 0.99 times the backbone's rate on the fit partition (Appendix D.3), so the evaluation rate is not guaranteed to fall below the backbone's, and the remaining gap is 0.0002 percentage points. The rate is also computed per predicted herb pair, and DAPS predicts 15.22 herbs per prescription against the backbone's fixed 12, so each of its prescriptions contains more pairs that could be flagged.

Hyperparameter Sensitivity: Count Scale and Count Weight. The scale and weight sweeps in Figure 5 vary one coordinate around the final (step-2b) validation-selected configuration on the

Table 16: **Equal-budget cardinality rules.** All rules act on the same refined scores after step 2 and before the conflict adjustment, use the same fit labels, and choose their setting by fit-set Jaccard; each test set is scored once. The global threshold is searched over $0.01, \ldots, 0.99$ and the fixed size over $k = 1, \ldots, 45$; the three count estimators share the search in Appendix D.2. PresRecST is used on TCM and MedAlign on MIMIC. Size is the mean predicted set size.

	TCM-Onsite		TCM-Internet		MIMIC-III		MIMIC-IV	
Rule	JA	Size	JA	Size	JA	Size	JA	Size
Fixed top-k	51.45	16.00	31.53	14.00	54.51	21.00	49.89	14.00
Global threshold	51.74	15.38	32.10	14.25	57.00	20.63	55.54	14.59
Marginal sum	51.49	15.39	31.99	13.60	56.75	21.18	55.22	14.41
Soft-Jaccard count	51.96	16.10	32.43	15.11	57.02	20.98	55.62	14.90
PB expected Jaccard	51.95	15.90	32.49	15.18	56.94	21.15	55.66	14.67

Table 17: **Plug-in comparison: DAPS on matched frozen backbones.** Size-independent PRAUC separates ranking quality from cardinality correction.

Backbone	Dataset	Jaccard			PRAUC		
		base	+DAPS	Δ	base	+DAPS	Δ
MedAlign	MIMIC-III	53.89	56.91	$+3.02$	78.43	81.66	$+3.23$
MedAlign	MIMIC-IV	51.20	55.24	$+4.04$	75.17	79.71	$+4.54$
MedAlign	TCM-Onsite	41.74	53.24	$+11.50$	68.78	74.64	$+5.86$
MedAlign	TCM-Internet	23.25	33.84	$+10.59$	49.03	51.91	$+2.88$

refined scores while holding its probability source and offset fixed. Their test curves are therefore descriptive sensitivity checks, not additional hyperparameter searches.

Both curves change little near the selected settings, but shrinking sets is costlier than enlarging them. Over count scales 0.90–1.10, test Jaccard stays within 0.71 points of the selected point on MIMIC-III and within 1.87 points on TCM-Onsite; the largest losses come from shrinking to scale 0.90 (-0.71 and -1.87), whereas enlarging to 1.10 costs only 0.18 and 0.10 points. The count weight interpolates between the backbone's count (weight 0) and the adaptive estimate of DAPS (weight 1); using the backbone's count alone costs 0.66 and 0.57 points, and the loss shrinks as the weight increases toward the selected values of 0.8 and 0.9. The validation-selected settings lie within 0.09 points of the best test value on every grid, so validation-only selection leaves little test-set gain unrealized.

Equal-Budget Cardinality Rules. Table 16 replaces only the size rule, keeping the refined scores and fit labels of each dataset fixed, and gives every rule the same validation budget. A fit-selected fixed size is lowest on all four datasets, trailing PB expected Jaccard by 0.50, 0.96, 2.42, and 5.77 Jaccard points (TCM-Onsite, TCM-Internet, MIMIC-III, MIMIC-IV); the marginal sum trails by 0.19–0.50 points. PB expected Jaccard and soft-Jaccard counting differ by at most 0.08 points, a tuned global threshold by at most 0.39, and PB is within 0.08 points of the best rule on every dataset. The gain therefore comes from choosing each prescription's size by the evaluation loss rather than from the particular estimator; PB is retained because it is never materially worse than the alternatives.

E.8 RANKING, RELATION AUDITS, AND SUPPLEMENTARY BASELINES

Set-Size and Ranking Analysis on Matched Backbones. This is the experiment behind the set-size and ranking analysis in Section 4: the same frozen export, split and seed, with only the decoder changed. Table 17 gives the matched pairs and Figure 7 compares decoder gains against backbone strength, including the GiantMed results.

On the two TCM corpora, MedAlign predicts 10.15 herbs on TCM-Onsite and 5.26 on TCM-Internet, below the corresponding evaluation means 14.66 and 12.43. In Table 21, GiantMed raises those sizes to 13.05 and 8.31 and raises Jaccard by 5.81 and 5.33 points. The other non-sequential adapted models gain substantially less under that plug-in, with changes of -0.38 to $+0.77$ points.

Table 18: **Relation-semantics illustration: six training-split herb associations on TCM-Onsite.** Held-out same-base alternatives have the negative sign consistent with exclusion; one classical incompatibility audit pair is positively associated. These selected pairs illustrate that association signs need not encode pharmacological meaning; they are not a population-level validation. Pair Jaccard is the intersection-over-union of herb occurrence sets.

Pair	Relation	Pearson	Pair Jaccard
zhi gancao / *sheng gancao*	same-base	-0.804	0.2%
kuqin / *huangqin*	same-base	-0.730	0.0%
qing banxia / *sheng banxia*	same-base	-0.671	0.0%
sheng baishao / *chao baishao*	same-base	-0.291	0.0%
hongshen / *lilu*	**antagonism audit**	$\mathbf{+0.209}$	**62.5%**
sheng gancao / *haizao*	antagonism audit	-0.005	7.6%

Table 19: **Supplementary language-model baselines on MIMIC-III/IV.** Prompted and fine-tuned rows use the same fixed 1,000-visit sample per corpus, so each fine-tuned row is paired with its prompted row; PRAUC is unavailable because generation does not yield a complete marginal vector. MedAlign and DAPS are re-scored on the same visits from their final sets, so PRAUC is likewise not reported.

	MIMIC-III					MIMIC-IV				
Method	JA	F_1	PRAUC	DDI	# Med.	JA	F_1	PRAUC	DDI	# Med.
Prompted zero-shot										
Qwen3-8B	18.85	30.58	–	5.73	38.33	17.33	27.46	–	5.28	30.42
Qwen3-32B	28.52	42.59	–	6.89	50.90	28.11	41.20	–	7.88	31.56
DeepSeek-V4-Flash	34.57	49.42	–	6.31	36.24	35.35	49.82	–	6.52	19.17
GLM-4.7-Flash	21.61	33.76	–	5.71	55.35	22.38	34.30	–	**5.08**	34.36
Qwen3.5-27B	27.73	41.67	–	**5.66**	39.57	28.99	42.41	–	5.51	20.96
LLaMA3.1	19.71	31.70	–	8.83	46.55	14.13	23.39	–	7.91	48.75
Aloe-Beta	22.33	35.65	–	8.53	40.15	17.34	28.46	–	8.26	29.64
LoRA fine-tuned on the training split										
Qwen3-8B	43.91	60.23	–	7.48	11.90	45.53	60.91	–	8.52	9.49
Qwen3-32B	44.15	60.45	–	7.03	11.63	45.65	61.01	–	8.68	9.46
DeepSeek-V4-Flash	36.42	52.53	–	7.19	14.06	33.99	49.26	–	8.16	9.50
GLM-4.7-Flash	47.11	63.26	–	7.56	13.79	47.92	63.18	–	8.84	11.09
Qwen3.5-27B	44.10	60.33	–	7.78	11.36	45.48	60.82	–	8.71	9.20
Frozen backbone and DAPS *on the same sample*										
MedAlign	53.97	69.29	–	7.06	20.14	50.81	66.00	–	6.58	15.82
+ DAPS	**56.91**	**71.73**	–	6.96	20.86	**56.13**	**70.45**	–	6.35	14.93

This pattern is consistent with cardinality contributing to the observed backbone dependence, although these comparisons do not isolate cardinality from changes in selected membership. DAPS explicitly selects cardinality after calibration and refinement; the PRAUC improvements in Table 17 additionally show ranking gains. Within Table 17, the Jaccard gains are much larger on the TCM pairs ($+11.50$ and $+10.59$) than on MIMIC ($+3.02$ and $+4.04$), whereas the size-independent PRAUC gains are of similar magnitude across corpora ($+5.86$ and $+2.88$ versus $+3.23$ and $+4.54$). On MIMIC the Jaccard gain thus tracks the ranking gain, while on TCM it exceeds it by 5.6–7.7 points, consistent with cardinality correction of MedAlign's under-sized TCM sets.

Correlation Analysis of Initial Set-Size Error. Figure 3(b) uses the 36 main-table pairs. For each pair, let $e_b = |\bar{k}_b - \bar{k}_{\mathrm{ref}}|$ be the absolute difference between mean predicted and reference set sizes, and let g_b be the Jaccard gain. The plotted reference means are 19.79, 13.93, 14.66, and 12.43 for MIMIC-III, MIMIC-IV, TCM-Onsite, and TCM-Internet, respectively; these evaluation references differ from the full-corpus means in Table 1. The MIMIC-III figure uses the shared plotted reference even for backbones with a different multi-visit cohort, so this is a descriptive comparison rather than a fully cohort-adjusted analysis.

Table 20: **Supplementary language-model baselines on TCM.** Language-model rows use the fixed 1,000-case samples described in Appendix E.3; API models are prompted only, and PRAUC is unavailable for direct generations. PresRecST and DAPS are re-scored on the same cases from their final sets, so PRAUC is likewise not reported.

	TCM-Onsite					TCM-Internet				
Method	JA	F_1	PRAUC	Inc.	# H.	JA	F_1	PRAUC	Inc.	# H.
Prompted zero-shot, API models										
GPT-5.6-sol	7.74	13.73	–	**0.01**	12.99	12.61	20.87	–	**0.00**	12.62
Claude Opus 5	11.15	19.45	–	0.06	16.72	17.75	28.01	–	0.01	14.56
Prompted zero-shot										
DeepSeek-V4-Flash	9.93	17.46	–	**0.01**	15.21	16.07	25.87	–	**0.00**	10.99
GLM-4.7-Flash	7.72	13.86	–	0.19	21.21	11.04	17.28	–	0.01	11.98
Qwen3.5-27B	8.65	15.39	–	0.50	22.02	8.16	13.86	–	0.11	10.75
Qwen3-32B	6.13	10.93	–	0.03	10.78	9.21	15.65	–	**0.00**	9.04
Qwen3-8B	6.13	10.95	–	0.03	11.95	10.39	17.57	–	0.01	10.80
ShenNong-TCM	3.66	6.65	–	0.05	7.57	11.11	18.90	–	**0.00**	9.81
HuatuoGPT2-7B	0.42	0.80	–	0.04	6.70	1.07	2.00	–	0.12	17.89
LoRA/QLoRA fine-tuned on the training split (Zhou et al., 2024)										
ShenNong-TCM	43.90	58.64	–	1.54	14.15	25.28	35.96	–	**0.00**	7.50
HuatuoGPT2-7B	43.12	57.94	–	1.49	14.06	25.80	36.51	–	**0.00**	7.67
Qwen3-8B	43.97	58.73	–	1.40	13.33	25.39	36.12	–	**0.00**	7.26
Qwen3-32B	47.12	61.56	–	1.35	13.36	26.54	37.48	–	0.01	7.18
DeepSeek-V4-Flash	34.54	49.59	–	1.82	14.71	22.79	33.63	–	**0.00**	8.16
GLM-4.7-Flash	46.17	60.78	–	1.34	13.98	26.47	37.47	–	0.01	7.87
Qwen3.5-27B	50.08	64.14	–	1.36	13.83	27.42	38.53	–	0.01	7.50
Frozen backbone and DAPS *on the same sample*										
PresRecST	49.09	63.73	–	1.31	15.00	30.09	43.99	–	**0.00**	12.00
+ DAPS	**52.10**	**66.39**	–	1.33	15.87	**32.58**	**46.57**	–	**0.00**	15.34

Pearson correlation between e_b and g_b is 0.76. To control for initial backbone Jaccard z_b, we separately regress e_b and g_b on $(1, z_b)$ and correlate their residuals, giving 0.74. This control removes a linear association with the baseline score; it does not control for every architectural or cohort difference, or establish a causal effect of set size. The plotted pairs and correlations use the displayed table entries; panel (a)'s mean markers and labels use the main-table group summaries computed before row-level rounding.

Semantic Analysis with Held-Out Relation Labels. This is the experiment behind the semantic analysis in Section 4. Table 18 recomputes drug–drug structure on TCM-Onsite from the training split alone, and the audit labels it is read against never enter relation learning. The four same-base pairs, raw and processed forms of one herb, have Pearson correlations from -0.804 to -0.291 and pair Jaccard of at most 0.2%: prescriptions select one form or the other, as the exclusion channel expects. The two classical antagonism pairs behave differently. *Sheng gancao* and *haizao* are essentially uncorrelated (-0.005), whereas *hongshen* and *lilu* are positively associated ($+0.209$) with pair Jaccard 62.5%. A negative sign thus matches substitution, but a listed incompatibility can carry either sign, so co-selection statistics cannot recover pharmacological conflict. This is why DAPS keeps external conflict knowledge in a separate channel rather than inferring it from observed co-selection.

Additional Language-Model Baselines. Tables 19 and 20 collect the sampled language-model evaluations, kept separate from the matched comparisons in the main tables. Each model is evaluated on the same fixed 1,000-example sample within a corpus, so zero-shot/fine-tuned comparisons are paired within this protocol. Both tables also re-score the frozen backbone and its DAPS decoding on these identical samples, making those rows directly comparable with the language-model rows.

Discussion of the MIMIC Comparison. On MIMIC-III/IV (Table 19), prompted open-weight models over-predict: they return 36–55 and 19–49 medications per visit against evaluation means of 19.79 and 13.93, and the best, DeepSeek-V4-Flash, reaches 34.57/35.35 Jaccard. Their low DDI rates, including the column minima of 5.66 and 5.08, accompany this weak overlap and do not indicate better prescriptions. Fine-tuning raises Jaccard to 36.42–47.11 and 33.99–47.92 but

Table 21: **Plug-in comparison: GiantMed on Western backbones adapted to TCM.** Every plug-in row is paired with the identical frozen backbone; ¶ marks our local rerun. VITA is omitted because GiantMed's boundary refinement does not fit its visit-aggregated probabilities (Appendix E.8). Adaptation details are in Appendix E for all these backbones.

	TCM-Onsite					TCM-Internet				
Method	JA ↑	F_1 ↑	PRAUC ↑	Inc. ↓	# H.	JA ↑	F_1 ↑	PRAUC ↑	Inc. ↓	# H.
MedAlign	41.74	56.88	**68.78**	1.81	10.15	23.25	34.53	**49.03**	0.01	5.26
+ GiantMed¶	**47.55**	**62.54**	68.74	1.46	13.05	**28.58**	**41.55**	47.58	0.01	8.31
DEPOT	42.44	58.38	67.56	1.15	24.22	23.69	37.19	45.85	0.14	32.85
+ GiantMed¶	42.66	58.66	67.72	1.08	25.14	23.31	36.77	45.81	**0.00**	34.19
MoleRec	38.26	54.45	63.45	0.90	28.43	21.30	34.29	44.30	0.21	39.80
+ GiantMed¶	39.03	55.27	63.65	**0.88**	28.54	21.14	34.10	44.27	**0.00**	40.53
SafeDrug	41.60	57.69	68.07	1.09	26.18	22.32	35.62	46.35	0.09	37.65
+ GiantMed¶	41.92	58.05	68.28	1.05	26.99	22.04	35.27	46.30	**0.00**	38.85
Mean change over the four matched backbones										
Δ + GiantMed¶	+1.78	+1.78	+0.13	−0.12	+1.19	+1.13	+1.52	−0.39	−0.11	+1.58

reverses the size error: fine-tuned models predict 11.36–14.06 and 9.20–11.09 medications, below both evaluation means. The strongest, GLM-4.7-Flash, reaches 47.11/47.92, still below the frozen MedAlign backbone (53.97/50.81). DAPS raises MedAlign to 56.91/56.13, 9.80/8.21 points above the best fine-tuned model, while its DDI rate (6.96/6.35) is lower than that of every fine-tuned model.

Discussion of the TCM Comparison. Table 20 separates three regimes. Prompted models remain far below every trained method: the strongest zero-shot result is Claude Opus 5 with 11.15/17.75 Jaccard on TCM-Onsite/TCM-Internet, the best open-weight model reaches 9.93/16.07, and HuatuoGPT2-7B stays near zero. Their near-zero incompatibility rates therefore reflect little overlap with recorded herbs rather than safer prescriptions. Fine-tuning closes most of this gap, raising Jaccard to 34.54–50.08 on TCM-Onsite and 22.79–27.42 on TCM-Internet; the lowest values come from DeepSeek-V4-Flash, trained on a 1,000-case subset. The strongest fine-tuned model, Qwen3.5-27B, reaches 50.08 on TCM-Onsite, above the PresRecST backbone (49.09) but below DAPS (52.10). On TCM-Internet, every fine-tuned model predicts only 7.18–8.16 herbs against an evaluation mean of 12.43, and the best reaches 27.42, below both PresRecST (30.09) and DAPS (32.58). DAPS therefore attains the highest Jaccard and F_1 on both corpora, with the largest margin where the language models under-size prescriptions, and on TCM-Onsite its incompatibility rate (1.33) is below that of every fine-tuned model.

Additional GiantMed Comparisons on TCM. Appendix Table 21 reports GiantMed on the adapted backbones whose DAPS results appear in Table 3. Over the four matched backbones, its mean Jaccard change is +1.78 on TCM-Onsite and +1.13 on TCM-Internet, compared with +12.11 and +11.18 for DAPS on the same backbones in Table 3. The GiantMed gain is concentrated on MedAlign, whose under-sized sets it enlarges (+5.81 and +5.33 points); DEPOT, MoleRec, and SafeDrug change by at most +0.77 points and decrease slightly on TCM-Internet. DAPS attains higher Jaccard than GiantMed on all eight backbone–corpus pairs, and GiantMed leaves PRAUC nearly unchanged (+0.13 and −0.39), consistent with an adjustment of set membership near the decision boundary rather than of the underlying ranking. VITA is not included: GiantMed revises candidates whose single backbone probability lies in a boundary band, whereas VITA aggregates probabilities over selected visits, so its published scores do not form such boundary sets and the pairing would require a different score.

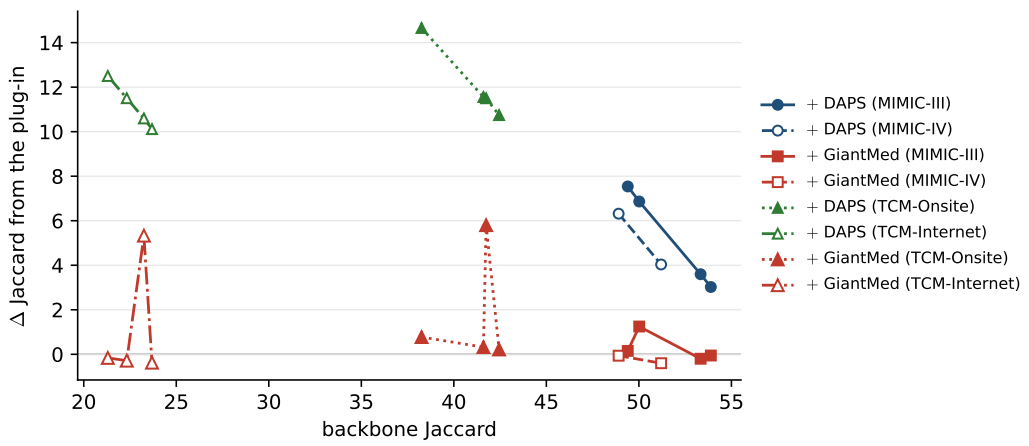


Figure 7: **Plug-in comparison: gains versus frozen-backbone strength.** Across Western and TCM corpora, DAPS is positive on every plotted pair, whereas GiantMed adds 5.8 and 5.3 Jaccard points only on MedAlign in both TCM corpora and at most 1.25 points on every other plotted backbone. Lines connect runs only within a corpus; the TCM curves are split by corpus and plug-in.