

Minimax Optimization in Deep Learning for Drug Recommendation: A Structured Review of Safety Objectives and Evaluation Practice

Suvarna¹, Dr. Mohammed Abdul Waheed²

¹Research Scholar, Department of Computer Science and Engineering, Visvesvaraya Technological University Centre for PG Studies Kalaburagi and Visvesvaraya Technological University, Belagavi - 590018. Email id: suvarnangangbo@gmail.com

²Associate Professor, Department of Computer Science and Engineering, Visvesvaraya Technological University Centre for PG Studies Kalaburagi and Visvesvaraya Technological University, Belagavi - 590018. Email id: mawaheed@vtu.ac.in

Abstract

Medication recommendation is now one of the most active applications of deep learning to electronic health records, and the published numbers suggest steady progress: Jaccard scores on MIMIC-III have climbed from 0.4509 for GAMENet in 2019 to roughly 0.53 for the strongest recent models. Read side by side rather than one at a time, that progress is harder to credit. This review organises the literature into two groups, one predicting a medication set from longitudinal records and one drawing on other signals entirely, and compares them on methodology, evaluation data, safety objective, and what their reported gains rest on. The convergence on a Jaccard of about 0.53 turns out to be largely an artefact of evaluation practice. Baseline rows are copied between papers rather than re-run; the same model appears with a 31 percent spread in its own headline safety metric across sources; and the label space is a therapeutic class rather than a prescribable drug, so that scoring the same data at finer granularity drops Jaccard to between 0.16 and 0.26. The field's only safety measure derives from a single spontaneous-reporting resource with no severity grading, and no reviewed model has been validated prospectively or at an external institution. We set out the gap under six headings and propose testable next steps: a controlled re-evaluation of the canon under one pipeline, prescribable-granularity labels, severity-weighted and deviation-based safety metrics, and hybrid designs pairing a structured sequence model with retrieval-grounded reasoning while reporting medication burden alongside accuracy.

Keywords: Drug recommendation, medication recommendation, deep learning, electronic health records, drug-drug interaction, graph neural networks, large language models, MIMIC, benchmark validity, clinical decision support

1. Introduction

Prescribing for a patient with several coexisting conditions is a combinatorial problem that clinicians solve under time pressure and with incomplete information. Multimorbidity and polypharmacy are common enough in older adults to be the ordinary case rather than the exception, and the clinical guidelines that inform each decision are written for single diseases, so a physician treating three conditions is left to reconcile three documents that were never designed to be read together. The appeal of an automated recommender is obvious. Given a patient's diagnoses, procedures and prescribing history, a model that proposes a medication set would compress a large search space into a shortlist, and one that also screened the shortlist for known interactions would do something a busy clinician cannot reliably do unaided.

Deep learning has supplied the machinery. Over roughly eight years the problem has been formalised as multi-label prediction over a fixed medication vocabulary, conditioned on a sequence of visits drawn from a de-identified intensive care database [1], [2]. The metric quartet has barely changed since 2019: Jaccard similarity against the physician's actual prescription, macro F1, the area under the precision-recall curve, and a drug-drug interaction rate computed against a pairwise resource mined from spontaneous adverse-event reports [3]. The medication vocabulary is drawn from the Anatomical Therapeutic Chemical classification, and molecular structures, where used, come from DrugBank. Architectures have moved through the sequence one would expect: recurrent encoders over visit histories, then graph-augmented memory, then models reading the molecular structure of each candidate drug, then generative decoders that copy forward drugs the patient already takes, and most recently causal formulations and large language models used either as the recommender or as one agent in a simulated care team.

The reported results describe a field in good health. Each paper improves on the last, the improvements are reported with confidence intervals and significance tests, and the safety metric usually falls at the same time as the accuracy

metric rises. A reader arriving at this literature would reasonably conclude that medication recommendation is close to being solved on the benchmark and that the remaining work is deployment.

That conclusion does not survive a careful reading of the papers side by side. GAMENet, the model whose architecture almost every later system inherits, reports a Jaccard of 0.4509 in its own paper on a 145-medication vocabulary [7]; the value quoted for GAMENet in later work ranges from 0.4848 to 0.5159 on a differently preprocessed cohort, which means the gap created by re-preprocessing a single baseline is of the same order as the gap the field has closed in six years of architectural work. SafeDrug's drug-drug interaction rate, the number that gives the model its name, is reported as 0.0589 by its own authors [8] and by COGNet [10], as 0.0649 by 4SDrug [11], and as 0.0773 by MoleRec [12], a relative spread of 31 percent with no acknowledged protocol change. COGNet's baseline rows are digit-identical to SafeDrug's, which is to say they were copied rather than re-run, and at least one recent paper states plainly that it took its baseline numbers from a third paper rather than reproducing them [16]. Where the reported gains are small, as they usually now are, this matters a great deal.

There is a second and more consequential problem, which concerns what is being predicted rather than how it is scored. Almost all of this work operates on drugs collapsed to the third level of the Anatomical Therapeutic Chemical classification, a therapeutic class rather than a prescribable medicine. Metformin and glipizide are the same label at that granularity, although they are different drugs with different contraindications, and a model that recommends the class has not made the decision the clinician actually faces [13]. When the same MIMIC data are scored at a finer level, the numbers change beyond recognition. A recent agentic system evaluating at the fourth ATC level reports a Jaccard of 0.2631 on MIMIC-IV and 0.1620 on MIMIC-III [28], against the roughly 0.53 the same datasets yield at class level. A prescription-level benchmark requiring a drug, dose and route triple finds the best of sixteen large language models reaching 46.10 percent exact-match accuracy, while those same models exceed 89 percent on medical licensing questions [29]. The plateau at 0.53 is not where prescribing competence saturates. It is where the label space stops resolving.

1.1 Motivation and Contribution

A researcher entering this area today faces a fragmented picture. The phrase drug recommendation is used by at least three communities that share almost no references: one predicting medication sets from structured hospital records, one predicting drug-disease associations for repositioning, and one ranking drugs by sentiment in review text. Within the largest of these, the evaluation protocol is inherited rather than specified, the safety objective rests on a single resource whose limitations are rarely discussed, and the only dedicated survey of deep learning for medication recommendation [4] is cited by one of the ten recent papers examined here. The broader review literature covers health recommender systems, interaction extraction and neural recommendation in general, but does not touch the evaluation questions raised below.

The contributions of this review are six.

1. The reviewed methods are organised into two groups along the input-signal axis: models that predict a medication set from longitudinal structured electronic health records, which now form a recognisable canon descending from a common patient encoder, and models that draw on alternative signals, including dialogue, review text, molecular graphs, and retrieval-grounded or agentic language models.
2. Methods are compared on four axes beyond headline accuracy: the cohort they were evaluated against, the safety objective they optimise, the medication burden their recommendations impose, and whether their baseline numbers were re-run or copied.
3. The field's own self-diagnosis is collected and attributed, so that the criticisms recent papers make of their predecessors can be read as a single body of evidence rather than as isolated remarks in unrelated related-work sections.
4. The benchmark is subjected to a validity analysis. We document specific, citable instances in which the same model is reported with materially different numbers in different papers, and we quantify how much of the field's apparent performance is attributable to label granularity.
5. The research gap is discussed under six headings: uncontrolled evaluation, coarse label granularity, a severity-flat and single-sourced safety objective, absent external and prospective validation, unpriced medication burden, and the persistent cold-start case.
6. Future directions are framed as incremental and testable rather than as a single integrated solution: a controlled re-evaluation of the existing canon under one pipeline, prescribable-granularity labelling, deviation-based and severity-weighted safety metrics, and hybrid architectures that combine a structured sequence model with retrieval-grounded reasoning.

2. Related Work

The literature divides cleanly into two groups, and the division is not one of technique but of what the model is allowed to read. The first group predicts a medication set from structured longitudinal records, uses MIMIC almost exclusively, and shares a dense common citation core. The second is more heterogeneous, covering systems that read consultation

transcripts or patient reviews, systems that reason over guidelines with a language model, and work on dosage, explanation and deployment that touches recommendation without being benchmarked against it. They are presented separately because scoring them on one metric would mislead.

It is worth beginning with the structure that the first group inherited, because almost everything since has been a modification of it. Early work split along a line that later papers still use: instance-based methods, which read only the current visit, and longitudinal methods, which read the history. LEAP [6] belongs to the first kind and framed prescription as a sequential decision, emitting drugs one at a time from a recurrent decoder with attention over diagnoses and fine-tuning against an interaction reward. Its weakness, as later work observed, is that a plan built from the present visit alone cannot see a patient's trajectory, which is what matters most in chronic disease. Memory-augmented variants followed, giving the two asynchronous views of diagnoses and procedures separate storage. RETAIN [5] contributed the two-level attention scheme that made the patient encoder inspectable; its own paper reported no medication metrics at all, yet it has been re-implemented as a baseline in almost every subsequent study, and the values different groups obtain for it do not agree. Alongside these sat conventional multi-label learners, chiefly logistic regression and classifier chains, which still occupy the top rows of most baseline tables.

GAMENet [7] is the pivot. It combined a dual recurrent patient encoder with a memory bank whose keys were graph-convolutional embeddings of an electronic health record co-prescription graph and an interaction graph, and it fixed both the metric quartet and the architectural template that the field still uses. One recent paper puts the situation more bluntly than most: a considerable portion of current longitudinal work still follows GAMENet's visit representation module [18]. That observation is worth holding onto, because it means the accumulated gains since 2019 have been made largely at the periphery of a patient encoder nobody has revisited. The one clear exception is the pretraining line, which uses masked-code objectives to exploit the large single-visit population that every sequential model discards; it remains a minority position, and the cold-start problem it was designed to solve is still open. Contemporaries extended the graph framing in other directions, adding reinforcement learning to make the prediction order-free, or pairing dual attention over visits with graph attention over drug co-occurrence, but none displaced the underlying encoder.

The next move was to open up the drug side. SafeDrug [8] paired a global substructure encoder with a message-passing encoder over molecular graphs, and introduced a controllable interaction loss that projects predictions back below a chosen threshold, turning safety from a soft penalty into a constraint the user sets. MoleRec [12] refined the granularity from whole molecules to substructures, weighting substructure relevance against the patient's condition, and reached near-COGNet accuracy at a lower interaction rate with roughly seven fewer drugs per visit. DEPOT [14] went further, arguing that substructure work captures only the inclusion relation between a drug and its fragments while missing the relationships among the fragments themselves. It builds a motif tree by BRICS fragmentation, encodes it with a structure-aware graph transformer carrying degree and recency embeddings, and splits prescribing into repeat, explore and discriminate branches so that drug turnover is modelled directly. Its most transferable idea is smaller than its architecture: weighting the interaction loss by a running average of recent batches rather than the current one, which stabilises training and lets SafeDrug and MoleRec converge to a lower loss bound with visibly smaller variance. Around this core sits a long tail of further variations on drug representation: sample-adaptive prediction depth, additional knowledge types, sampling corrections for missing records, patient retrieval, adaptive interaction between the two clinical views, multi-graph coordination. Molecular structure helps, consistently. The assumption underneath it, noted sharply by one recent paper, is that medications with similar structures behave similarly, and that is not generally true [18].

Two other modifications deserve separate mention because they change the prediction target rather than the encoder. MICRON [9] predicts the residual between consecutive visits, that is which drugs are added and removed, and then reconstructs the set. This mirrors how regimens are actually adjusted, and does so with about 275,000 parameters. COGNet [10] goes the other way, generating the set with a copy-or-predict decoder that either carries a drug forward or emits a new one. It held the accuracy record for some time, and the manner in which it did so is instructive: its Jaccard of 0.5336 comes with an interaction rate of 0.0852, which exceeds the 0.0808 rate that the same paper reports for the physicians' own prescriptions, and an average of 28.09 drugs per visit against SafeDrug's 19.92. Ranking by Jaccard alone rewards recommending more drugs, and COGNet is the clearest demonstration of it. The ground-truth rate is not itself a stable quantity here: 0.0808 comes from the 131-drug pipeline SafeDrug established, while a recent paper preprocessing the same database reports 0.0904 [18]. Two papers disagreeing about what the physicians themselves prescribed is a compact illustration of the problem examined below.

A causal strand emerged in response to the observation that all of this learns correlation. The first entrant applied front-door adjustment to deconfound recommendation bias and encoded interaction avoidance as a satisfiability problem rather than a differentiable penalty, though it required symptom data that standard MIMIC pipelines do not contain, which is why later work reports it as not directly comparable. CRec [15] extends the idea to the molecule itself, building a structural causal model in which a patient-conditioned Bernoulli mask, sampled with Gumbel-Softmax, splits each molecular graph into a causal and a spurious part; the spurious branch is then destroyed by a loss pushing its predictions towards uniform. The motivation is sound and the qualitative results are convincing: the highlighted substructures correspond to recognisable chemistry. What is asserted rather than shown is the causal claim itself. Backdoor adjustment is implemented by shuffling the spurious representation within a training batch, a heuristic

never proved equivalent to the stated adjustment, and the model is evaluated only in distribution, so nothing distinguishes a causal model from an accurate one. The rest of the causal branch shares this limitation.

The most recent structural work addresses how visits are selected rather than how drugs are encoded. RaVSNet [18] makes the sharpest version of the argument: aggregating all of a patient's history without filtering introduces noise when past events are unrelated to the present admission, and modelling only longitudinal similarity leaves first-visit patients unserved. It therefore selects relevant past visits differentially with Gumbel-Softmax and retrieves transversally similar visits from other patients, which is a direct attack on the cold-start case that COGNet structurally cannot handle. It also proposes a better safety target than the field has been using: since real prescriptions in MIMIC-III and MIMIC-IV carry interaction rates of 0.0904 and 0.0883, driving a model's rate towards zero is not obviously desirable, and the absolute deviation from the ground-truth rate is the more defensible quantity. CD-MDRec [16] attacks the same sparsity problem from the representation side, denoising visit embeddings with a diffusion module and aligning two attention pipelines with a contrastive objective. A parallel group addresses the long tail, rare-disease patients and visit weighting, all of which are versions of the same complaint: the benchmark's averages are carried by common cases, and the patients a recommender would most help are the ones it serves worst.

REFINE [13] is the one paper in this group that questions the task definition rather than the architecture, and it is the more valuable for it. Its two claims are that class-level recommendation cannot distinguish patients needing one drug from a class from those needing several, and that treating all interactions as equally severe makes models over-conservative, since clinicians routinely co-prescribe moderately interacting pairs when the benefit warrants it. It therefore recommends individual drug names, encodes interaction severity as graded weights from an external severity-annotated resource, and subtracts a co-prescription benefit term from the interaction penalty. The effect on behaviour is measurable and in the expected direction: SafeDrug misses 4.29 drugs per visit against a ground truth of 8.72 on MIMIC-IV, while REFINE misses 2.94. Table 1 brings this group together.

Table 1. Deep learning models for medication recommendation from longitudinal electronic health records.

Reference	Methodology	Advantages	Shortcomings
[5] Choi et al., 2016 (RETAIN)	Two-level attention over visits and codes with a reverse-time recurrent encoder.	Made the patient encoder inspectable; still the standard interpretable baseline.	Reports no medication metrics in its own paper; every Jaccard attributed to it is a third-party re-implementation.
[6] Zhang et al., 2017 (LEAP)	Recurrent decoder with content-based attention over diagnoses; set size determined automatically; RL fine-tuning with an interaction-aware reward.	First to frame prescription as set generation under an explicit safety reward.	Instance-based: cannot see the patient trajectory; absolute metrics from the original paper are not independently recoverable.
[7] Shang et al., 2019 (GAMENet)	Dual RNN encoder querying a memory bank keyed by GCN embeddings of an NAïVE co-prescription graph and an interaction graph.	Fixed the metric quartet and the architectural template; first to combine history with a DDI graph.	Its own Jaccard (0.4509, 145 drugs) is not the value later papers quote (0.4848-0.5159); its encoder is still reused largely unexamined.
[8] Yang et al., 2021 (SafeDrug)	Dual molecular encoders (global substructure and message-passing) with a controllable DDI loss projecting predictions below a set threshold.	Made safety a tunable constraint rather than a soft penalty; the most copied experimental protocol in the field.	Over-conservative: misses 4.29 drugs per visit on MIMIC-IV; its own DDI rate is reported as 0.0589, 0.0649 and 0.0773 across three papers.
[9] Yang et al., 2021 (MICRON)	Recurrent residual network predicting the medication change between consecutive visits, then reconstructing the set.	Clinically faithful framing; roughly 275k parameters and about 1.5x faster training than comparable models.	Reports no PRAUC or average drug count of its own; single-dataset evaluation.
[10] Wu et al., 2022 (COGNet)	Encoder-decoder generation with a copy-or-predict module choosing between reusing a historical drug and emitting a new one.	Highest accuracy of its generation (Jaccard 0.5336) and an explicit model of drug reuse.	DDI rate 0.0852 exceeds the ground-truth 0.0808 and it recommends 28.09 drugs per visit; baseline rows are digit-identical to SafeDrug's, i.e. copied.

[11] Tan et al., 2022 (4Sdrug)	Set-to-set matching from a symptom set to a drug set with importance-based aggregation and explicit set-size and intersection penalties.	Serves the realistic case where only a symptom list exists; smallest recommended sets (17.50 drugs) at DDI 0.0600.	Different cohort construction (27,869 symptom-set visits) makes its table numerically incomparable to the longitudinal lineage despite matching metric names.
[12] Yang et al., 2023 (MoleRec)	Hierarchical substructure-aware encoding weighting molecular substructures against the patient's condition.	Near-COGNet accuracy (0.5305) at a materially lower DDI rate (0.0724) and about seven fewer drugs per visit.	Omits MICRON and COGNet from its main table; rule-based fragmentation is patient-independent; no MIMIC-IV results.
[13] Bhoi et al., 2023 (REFINE)	Transformer encoders over dosage and lab-response trends; severity-weighted DDI graph with GATv2; balanced interaction loss offsetting co-prescription benefit.	Recommends individual drugs rather than ATC classes; only system to grade interaction severity; validated on a real outpatient cohort.	Second dataset is proprietary and irreproducible; Jaccard is absent from the main table; the safety loss and the safety metric are computed from the same source, so the gain is partly circular.
[14] Zhao et al., 2024 (DEPOT)	BRICS motif trees encoded by a structure-aware graph transformer; repeat/explore/discriminate decomposition; interaction loss weighted by a running history of recent batches.	Models relations among motifs rather than mere inclusion; evaluated on four datasets including Eicu and OMOP; the history-weighted loss transfers to other models.	No runtime or parameter cost reported for a markedly heavier model; the large Eicu and OMOP gains are measured against a baseline the authors concede is handicapped on those datasets.
[15] Zhang et al., 2025 (Crec)	Structural causal model with a patient-conditioned Gumbel-Softmax mask separating causal from spurious molecular substructure; spurious branch suppressed towards uniform prediction.	Learned, patient-specific segmentation replaces rigid rule-based fragmentation; chemically plausible qualitative attributions.	Backdoor adjustment is implemented as batch shuffling without proof of equivalence; only four baselines; evaluated in distribution only, so causal validity is asserted rather than tested.
[16] Ma et al., 2025 (CD-MDRec)	Diffusion-based representation denoising with dual self-attention and pair-wise attention pipelines aligned by an InfoNCE contrastive loss.	First use of diffusion for denoising in this task; helps most in the sparse regime; evaluated on MIMIC-II, III and IV.	Baseline results are copied from another paper rather than re-run; loses on DDI rate on MIMIC-III and MIMIC-IV and on PRAUC on MIMIC-IV, while the abstract reports only the two metrics it wins.
[18] He et al., 2026 (RaVSNet)	Gumbel-Softmax filtering of longitudinal visits plus retrieval of transversally similar visits from other patients; pretraining by masked reconstruction and causal-effect estimation.	Directly targets the first-visit cold-start case; the pretraining module transfers to five other models; proposes deviation from the true DDI rate as the safety target.	Assumes symptom codes MIMIC does not natively contain, and the extraction procedure is cited to prior work rather than described; high computational cost; transversal retrieval over the visit matrix carries a leakage risk that is never addressed.

A sceptical reading of Table 1 is necessary, because the column of headline numbers is less solid than it looks. Four observations are worth making explicit. First, the baselines are frequently not re-run. COGNet's rows for logistic regression, RETAIN, LEAP, GAMENet and SafeDrug are digit-identical to SafeDrug's, and CD-MDRec states that it took its baseline results from another paper, and MedFusion-Agent draws its MIMIC-IV comparison from prior benchmark-aligned studies rather than reproducing them [19]; any preprocessing error in an early paper is therefore

inherited unexamined by everything downstream. Second, where a full re-run does exist it is always produced by authors with a model in the comparison. DEPOT retunes the hyper-parameters of every baseline it reports, and RaVSNet places ten prior models in a single table on both MIMIC releases, which is exactly the right practice; but a comparison conducted by a competitor is not an independent one, and the two do not agree with each other or with the originating papers. Between them the field has no neutral reference table. Third, the metric sets do not line up: MoleRec omits MICRON and COGNet, MICRON reports no PRAUC of its own, and one widely used memory-network baseline never reported Jaccard at all, so every value quoted for it is a re-implementation, and the two published re-implementations differ by three points. Fourth, the randomness protocols are not comparable: SafeDrug and COGNet report standard deviations over ten bootstrap rounds sampling eighty percent of the test set, which shrinks the interval relative to independent retraining, so overlapping error bars across papers say nothing. Fifth, and most simply, the reported gains are now smaller than the documented disagreement between sources about the same system. Recent papers claim improvements of one to three percent, CD-MDRec's headline being an average of 2.12 percent in Jaccard; against that, SafeDrug's own interaction rate varies from 0.0589 to 0.0773 depending on which paper reports it, a spread wider than the gap separating the safest from the least safe model in most published comparisons. When the noise floor of the benchmark exceeds the effect size, a significance test computed inside a single pipeline is not evidence of much.

The second group of work takes a different input. Its internal diversity is real, and it is also where the field's terminology becomes unreliable. Three separate research communities use the phrase drug recommendation for tasks that share no data, no metrics and no citations. DS-GAT [17] represents the dialogue branch, and its motivating argument is a fair criticism of the first group: a structured record describes what has been coded about a patient, not what the patient is complaining of now, so a consultation transcript carries information the electronic record does not. It encodes utterances with RoBERTa, treats them as nodes in a graph attention network whose attention scores are reweighted by grey relational analysis to compensate for sparse neighbourhoods, and fuses the result with entities retrieved from a medical knowledge graph. The idea is sound; the evaluation is not comparable to anything else, since the label space is 70 drugs and 16 diseases against the 131 to 735 of the MIMIC benchmarks, no interaction or safety metric appears anywhere in a paper about recommending drugs, and the ground truth is whichever drug an online physician happened to mention in a chat window.

HyBLRecs [21] represents the review-based branch, which is the branch closest to the sentiment-analysis framing that motivates much applied work in this area. It projects a frozen BioBERT representation of each review down to 128 dimensions, concatenates it with a VADER sentiment embedding and four normalised statistical features, and classifies efficacy with LightGBM, reporting an F1 of 0.9691 on the Drugs.com corpus. Two things about it are useful. The compressed representation beats the raw 768-dimensional one while training about seven times faster, and the ranking layer fixes a real failure mode by penalising drugs whose perfect average rests on a single review, demoting eight such drugs by around sixty-nine positions in the authors' acne case study. Two are not. The target is a patient star rating rather than clinical efficacy, which the authors concede in a footnote, so a high F1 measures how well the model reconstructs a subjective score from the text that produced it. And the claim of outperforming existing methods does not hold on the authors' own table, where a RoBERTa baseline records 0.9709 against HyBLRecs's 0.9691. The safety result is weaker than it sounds too: after re-ranking, 90.67 percent of top-ten lists still contain a drug the authors classify as high risk, down from 95.27.

Large language models entered the field along two paths. The first treats the model as a knowledge-grounded reasoner. DrugGPT [22] is the most substantial example, coordinating three language models over a purpose-built disease-symptom-drug graph with prompting strategies for knowledge consistency and evidence traceability. It handles the leakage problem more carefully than most work does, building three datasets from drugs and interactions released after the baselines' training cutoff; where GPT-4 falls below 65 percent on those, DrugGPT reaches 93.5 percent on the vaccine interaction set. Its own numbers also contain the most honest result in this part of the literature. On open-ended recommendation, as opposed to multiple choice, it records a precision of 48.3 and a recall of 35.3, missing roughly two thirds of the drugs a physician would list. The second path fine-tunes or distills, either compressing a prompted model into a deployable one or adapting it directly. LAMO [27] fine-tunes on clinical notes with an explicit penalty on set size, and documents the failure mode that makes naïve language models unusable here: GPT-4 recommends more than eighty medications per patient and LLaMA around 150, against a physician's twenty-odd. LAMO is also, as far as we can establish, the only work here reporting temporal validation across the ICD-9 to ICD-10 boundary together with external multi-centre results. A third and more conservative path keeps the structured model in charge and invokes the language model only on boundary cases near the decision margin, which caps both the inference cost and the exposure to unconstrained generation. The clinical evidence base behind all three paths remains thin.

The agentic branch is newer and its results are mixed in an interesting way. MedFusion-Agent [19] is the strongest instance: rather than letting several language-model agents converge by discussion, it has role-specialised guideline agents emit candidate actions into a fixed schema, then resolves them with an explicit conflict graph and a binary constrained optimisation, then screens the survivors against a patient-specific interaction matrix. Repair then proceeds

one drug at a time until the residual risk reaches zero or an iteration cap is hit, at which point the lowest-risk compromise is kept with its residual risks annotated rather than silently accepted. On the authors' stress cohort the overall hazard rate falls to 1.8 percent against 29.8 percent for a single language model, and it holds across two different backbones, which suggests the gain is architectural rather than a property of one vendor's model. The caveats are equally clear. The primary benchmark is synthetic, built with Synthea, adversarially perturbed by the authors and adjudicated against a standard they wrote themselves, so the hazard rate measures compliance with their own rule set. A single language model failing 45 of 45 metformin cases suggests the baseline may be under-prompted. On MIMIC-IV the system reports an F1 of 0.7121 and a Jaccard of 0.6153 against AMGNet's 0.6491 and 0.5063, but those baseline figures are drawn from prior studies rather than reproduced, which puts this paper in the same category as the copied comparisons above. Set against it, a careful evaluation study found single agents performing comparably to multi-agent systems, with the best models still introducing unnecessary medications and drug-condition conflicts [30]. Prescribing remains the least developed of the agentic clinical tasks, and the least equipped with the monitoring a deployed system would need.

Two further strands touch the problem without being part of it, and both are worth separating explicitly so that they are not mistaken for baselines. FusionMVSA [20] fuses four drug similarity views and two disease views with shared-parameter self-attention and reports strong results on the Fdataset and Cdataset benchmarks, but the task is drug-disease association prediction for repositioning: 593 drugs, 313 diseases, no patients, no dosages, and interaction data entering as one similarity view rather than a safety objective. It cites none of the recommendation canon, and the word recommendation in its title does entirely different work; the same applies to multi-omics models of cancer drug response, which share the vocabulary and none of the task. Work on dosage and explanation, by contrast, is converging on the recommendation task: the framework of [23] couples retrieval-augmented evidence, a patient digital twin, a reinforcement learning dosing policy and SHAP attributions, reporting dosage errors about 18 percent below a guideline baseline and trust ratings from 40 clinicians. Its evidential status is weaker than its numbers suggest. The language model is never named, the guideline baseline never operationalised, a quarter of the training data is synthesised by the model under evaluation, and the authors describe their own results as feasibility illustrations.

One recent paper is worth citing as a cautionary case rather than as a method. A transformer-titled framework for multi-label recommendation from clinical notes [24] reports an architecture containing no attention mechanism, and constructs its labels by string-matching drug names in the note text and then assigning each patient one to three additional drugs at random. Random labels are unlearnable by construction, which the reported micro F1 of 0.1600 and micro Jaccard of 0.0870 duly reflect, and which cannot be reconciled with the same paper's claims of roughly 90 percent sensitivity and average precision near 0.97. The authors are candid about the synthetic labelling in their limitations section. The case matters because the paper appeared in a reputable venue, which says something about how thinly this area is currently being reviewed.

Finally, a small body of work addresses what happens after a model is deployed, and it deserves more attention from modellers than it gets. An assessment of ten clinically validated mobile therapeutic applications found five storing personally identifiable or protected health information in cleartext, eight without certificate pinning, and two permitting unauthenticated write access to the data driving therapeutic output [25]. On the model side, a quantum-safe authentication scheme for digital-twin pipelines shows a clinical classifier at roughly 99 percent accuracy collapsing to zero under a modest projected-gradient perturbation [26]. Neither paper recommends medications. Both describe the conditions under which a recommender's accuracy would stop meaning anything. Table 2 summarises this group.

Table 2. Alternative-signal, generative, agentic and adjacent approaches relevant to drug recommendation.

Reference	Methodology	Advantages	Shortcomings
[17] Nemade et al., 2025 (DS-GAT)	RoBERTa utterance encoding over a dialogue graph with grey-relational reweighted attention, fused with TransR knowledge-graph entities.	Captures the patient's presenting complaint, which structured records omit; reported as faster than its strongest baseline, though the margin is given only in a figure.	Label space of 70 drugs and 16 diseases; no DDI or safety metric of any kind; single Chinese-language platform; reported gains are single-run point estimates; in-text citations do not resolve to the printed reference list.
[21] Pham and Phan, 2026 (HyBLRecs)	Frozen BioBERT projected to 128 dimensions, concatenated with VADER sentiment and four statistical features, classified by LightGBM; confidence re-ranking penalising low-evidence drugs.	ID-free design addresses cold start; the compressed representation beats the raw one while training about seven times faster; the re-ranker fixes a real single-review failure mode.	Optimises a patient star rating, not clinical efficacy; does not actually beat its strongest baseline (F1 0.9691 vs 0.9709); 90.67 percent of top-ten lists still contain a self-defined high-

			risk drug; no interaction or contraindication check.
[22] Zhou et al., 2026 (DrugGPT)	Three collaborating language models over a purpose-built disease-symptom-drug graph with knowledge-consistency and evidence-traceable prompting.	Grounded, citable outputs; constructs post-cutoff datasets to defeat leakage, reaching 93.5 percent where GPT-4 falls below 65 percent.	On open-ended recommendation, precision 48.3 and recall 35.3, i.e. two thirds of a physician's drugs are missed; depends on a closed 175B API; human evaluation uses two experts on 100 samples with no agreement statistic.
[27] Zhao et al., 2025 (LAMO)	Parameter-efficient fine-tuning on clinical notes with an explicit penalty on recommendation set size.	Quantifies naive-LLM overprescription (GPT-4 above 80 drugs per patient, LLaMA around 150); the only reviewed work reporting both temporal and external multi-centre validation.	Preprint without a peer-reviewed venue at the time of writing; transfer is reported as limited for dissimilar medication categories.
[19] Ou et al., 2026 (MedFusion-Agent)	Role-specialised guideline agents emitting schema-constrained actions, resolved by an explicit conflict graph and constrained optimisation, then screened and repaired against a patient-specific DDI matrix.	Replaces language-level consensus with auditable symbolic arbitration; hazard rate 1.8 percent against 29.8 for a single LLM, and the gain holds across two model backbones.	Primary benchmark is synthetic and adjudicated against the authors' own silver standard; the single-LLM baseline failing 45 of 45 cases suggests under-prompting; the zero-shot MIMIC-IV win over trained models invites leakage scrutiny; no clinician adjudication.
[28] Wang et al., 2026 (SafeRx-Agent)	Specialty-aware agents with clinical-knowledge grounding and an explicit safety-verification stage, generating traceable fourth-level ATC medication sets.	Operates at a prescribable granularity and produces fewer, more defensible predictions (6.18 against 9.61 for its baseline).	Jaccard of 0.2631 on MIMIC-IV, which is not a weakness of the method so much as the clearest evidence that class-level scores are inflated by the label space.
[29] Chen et al., 2026 (RxEval)	Prescription-level benchmark of 1,547 questions over 584 patients requiring a medication, dose and route triple, with distractors from perturbed reasoning chains.	Scores at the granularity a clinician actually prescribes at; shows medical fine-tuning yields at most 2.20 F1.	Multiple-choice rather than open generation; best exact-match accuracy of 46.10 percent establishes a floor, not a diagnosis of why.
[30] Wu et al., 2025	Comparative evaluation of single-agent and multi-agent LLM systems for multimorbidity therapy, with metrics for clinical goals met and medication burden.	A rare published negative result; introduces burden-aware metrics rather than precision and recall alone.	Qualitative; no single headline metric; small evaluation scope.
[23] Daglarli, 2026	Retrieval-augmented evidence ingestion, a patient digital twin with GAN augmentation, a transformer RL dosing policy, and SHAP attributions with calibration gating.	Produces dose, calibrated risk and an inspectable rationale in one artefact; ablations identify retrieval and the digital twin as the load-bearing components.	The LLM is never named and the guideline baseline never operationalised, so nothing is reproducible; a quarter of the data is synthesised by the model under evaluation; explanations are validated against an internal self-

			consistency score, not clinician judgement.
[24] Naveed et al., 2026	Embedding, global average pooling and dense layers over MIMIC-IV clinical notes for multi-label prediction.	Honest reporting of the degenerate behaviour at a fixed threshold.	Labels are partly assigned at random, making the task unlearnable by construction; contains no transformer or attention despite the title; micro F1 of 0.1600 cannot be reconciled with the claimed 90 percent sensitivity.
[20] Meng et al., 2026 (FusionMVSA)	Shared-parameter multi-view self-attention over four drug and two disease similarity views with similarity-guided attention bias.	Replaces stacked graph convolution with single-pass global attention; supports predictions with database evidence and molecular docking.	Not medication recommendation: drug-disease association prediction on 593 drugs and 313 diseases with no patients, dosages or safety objective; cites none of the recommendation canon.
[25] Balossini et al., 2026	STRIDE threat model for digital therapeutics plus static, dynamic and API assessment of ten clinically validated mobile applications.	First threat model treating patient health rather than data as the primary asset; every recommendation traces to an observed finding.	Ten Android applications only, with prescription-gated products excluded; findings are binary with no severity scoring; the highest-impact model-level threats were not empirically tested.
[26] Pandey et al., 2026	Ring-LWE post-quantum authenticated key agreement for the patient-to-digital-twin path, with an FGSM and PGD robustness study on the downstream classifier.	Lowest communication cost among schemes meeting the full security feature set; quantifies the collapse of a 99 percent classifier to zero accuracy under perturbation.	White-box attacks assume the adversary already holds the weights, which channel authentication does not prevent; security analysis is informal; no connection to medication data.

Taken across both tables, several patterns stand out. The first group has converged on a narrow set of architectural choices, a single institution's intensive care data, and a metric quartet unrevised since 2019, and within those constraints the returns have visibly diminished. The second group is more varied and more willing to question the task definition, but weaker on evaluation discipline, with synthetic benchmarks, proprietary datasets and unnamed language models recurring. Where the two meet, the result is unflattering to both. A structured model scoring 0.53 at class level and an agentic model scoring 0.26 at prescribable level are not disagreeing about prescribing; they are measuring different things, and only one is measuring what a clinician needs. The most useful recent contributions, on this reading, are not architectures but reframings: interaction severity as graded rather than binary [13], deviation from the observed prescribing rate rather than minimisation as the safety target [18], medication burden reported alongside accuracy [27], [30], and the label space treated as an experimental variable [28], [29].

3. Research Gap and Future Directions

Reading the two tables side by side surfaces gaps that individual papers, each measured against baselines of their own choosing, tend to conceal.

7. Evaluation is uncontrolled. Baseline rows are copied rather than re-run; the tables that do compare the canon under one pipeline come from authors with a model in the contest and disagree with each other; randomness protocols differ, so error bars are not comparable across papers. Recent gains of one to three percent sit below the documented spread for the same system, which makes a significance test computed inside a single pipeline close to meaningless.

8. Labels are too coarse. ATC level three is a therapeutic class, not a prescribable medicine, and the distinction it erases is the one the prescriber makes. Scored at the fourth level the same MIMIC data yield a Jaccard of 0.16 to 0.26 [28]; scored as a drug, dose and route triple, the best of sixteen language models manages 46.10 percent exact match [29]. A 0.53 at class level is compatible with a system that could not safely write a prescription.

9. The safety objective is single-sourced and severity-flat. Interaction rate inherits three properties from its origin: one spontaneous-reporting resource with known reporting and indication bias [3], no severity grading, and zero assumed risk for any pair the resource omits. Only REFINE grades severity, and it draws loss and metric from the

same numbers, so that gain is partly circular. Minimisation is the wrong target anyway: physicians prescribe at 0.0904 and 0.0883 on MIMIC-III and MIMIC-IV [18].

10. Validation stops at the training distribution. Almost everything here is retrospective on one or two overlapping cohorts from the same intensive care unit. DEPOT reaches eICU and OMOP; LAMO reports temporal and external multi-centre results. Neither is prospective, and no model has been tested with a clinician in the loop on live cases. For systems justified entirely by patient safety, that is a wide gap between claim and evidence.

11. Medication burden is not priced. Jaccard rewards recommending more drugs: the highest-accuracy model of its generation recommends 28.09 per visit and outranks a model recommending 19.92. Set size, false-miss rate on safety-critical classes and deviation from observed prescribing all carry information a single overlap coefficient does not. Nor does anything in the metric set say whether a model's probabilities are trustworthy at the point of prescribing.

12. Cold start is unsolved, and structurally so. A model that copies drugs forward has nothing to copy on a first admission, and the patients who would benefit most are exactly those without a history in the system. RaVSNet's transversal retrieval [18] and masked-code pretraining on single-visit records are the two credible responses; neither has been widely adopted.

The directions that follow are incremental and testable rather than a single integrated proposal. The most immediately useful contribution anyone could make here is not another architecture but a controlled re-evaluation: one preprocessing pipeline, one split protocol, independent retraining rather than bootstrap resampling of a fixed test set, and the full metric set including average set size for every model. A second is to move the label space down at least one ATC level and report the cost, which would reset expectations and make the remaining headroom visible; the early agentic results at that granularity suggest it is large. A third is to replace the binary interaction matrix with a severity-graded one drawn from a source independent of the evaluation metric. A fourth is external and eventually prospective validation, for which the multi-centre databases now available make the retrospective half straightforward. On architecture, the evidence assembled here points towards hybrid designs rather than further refinement of any single paradigm, and the evidence points to a particular combination. Structured sequence models do well what they were built for, exploiting a patient's coded trajectory, and badly at everything the code space omits: the presenting complaint, the free-text narrative, the current guideline, the reason a combination is contraindicated. Retrieval-grounded and agentic models are the reverse, reading guidelines and narrative fluently while overprescribing badly when unconstrained [27] and failing to settle on a defensible answer when arbitration is left to discussion [30]. The failure modes are complementary, which is the argument for combining them: a structured encoder to propose, a retrieval-grounded reasoner to justify against current evidence, and a symbolic layer to arbitrate conflicts and report violations in a form a pharmacist can check. MedFusion-Agent's conflict graph and repair loop are the clearest existing sketch of that third component, and the obvious next experiment is to put a trained sequence model at the proposal stage in place of a zero-shot language model, then test the composite on real rather than synthetic cases. The boundary-aware designs already noted are the nearest existing approximation, and suggest such a system need not be expensive at inference time. Patient-reported signal, developed in isolation by the review-based literature, is a plausible fourth input where tolerability drives adherence, provided its selection bias is modelled rather than ignored. None of this is worth much without the evaluation work above: a hybrid design cannot currently be shown to help because the benchmark is not sensitive enough to tell.

4. Conclusion

Read as one body of work rather than one paper at a time, this literature looks different. The architectural progression from recurrent encoders through graph-augmented memory, molecular substructure, generative decoding and visit similarity is real, and several ideas along the way will last: controllable safety constraints, change prediction in place of set prediction, motif-level structure, transversal retrieval for patients with no history. What is not established is that these ideas produced the improvements attributed to them. The benchmark carrying the claims is uncontrolled, its baselines are copied rather than reproduced, its safety metric comes from one severity-flat resource, and its labels are therapeutic classes rather than prescribable medicines. Relax that last constraint and performance halves, which locates most of the apparent progress in the measurement rather than the models. The more useful recent work is the work that changes the question: grading interaction severity, targeting deviation from observed prescribing instead of minimisation, reporting medication burden beside overlap, scoring at a granularity a clinician could act on. A hybrid pairing a structured proposal stage with retrieval-grounded justification and symbolic arbitration is the natural next step, and this review's recommendation. Whether it beats what exists is unanswerable until the field agrees on a protocol able to detect the difference.

References

- [1] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, “MIMIC-III, a freely accessible critical care database,” *Scientific Data*, vol. 3, art. no. 160035, 2016, doi: 10.1038/sdata.2016.35.
- [2] A. E. W. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, L.-w. H. Lehman, L. A. Celi, and R. G. Mark, “MIMIC-IV, a freely accessible electronic health record dataset,” *Scientific Data*, vol. 10, art. no. 1, 2023, doi: 10.1038/s41597-022-01899-x.
- [3] N. P. Tatonetti, P. P. Ye, R. Daneshjou, and R. B. Altman, “Data-driven prediction of drug effects and interactions,” *Science Translational Medicine*, vol. 4, no. 125, art. no. 125ra31, Mar. 2012, doi: 10.1126/scitranslmed.3003377.
- [4] Z. Ali, Y. Huang, I. Ullah, J. Feng, C. Deng, N. Thierry, A. Khan, A. U. Jan, X. Shen, W. Rui, and G. Qi, “Deep learning for medication recommendation: A systematic survey,” *Data Intelligence*, vol. 5, no. 2, pp. 303–354, 2023, doi: 10.1162/dint_a_00197.
- [5] E. Choi, M. T. Bahadori, J. A. Kulas, A. Schuetz, W. F. Stewart, and J. Sun, “RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism,” in *Advances in Neural Information Processing Systems* 29 (NIPS 2016), 2016.
- [6] Y. Zhang, R. Chen, J. Tang, W. F. Stewart, and J. Sun, “LEAP: Learning to prescribe effective and safe treatment combinations for multimorbidity,” in *Proc. 23rd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD '17)*, 2017, pp. 1315–1324, doi: 10.1145/3097983.3098109.
- [7] J. Shang, C. Xiao, T. Ma, H. Li, and J. Sun, “GAMENet: Graph augmented memory networks for recommending medication combination,” in *Proc. AAAI Conf. Artificial Intelligence*, vol. 33, no. 1, 2019, pp. 1126–1133, doi: 10.1609/aaai.v33i01.33011126.
- [8] C. Yang, C. Xiao, F. Ma, L. Glass, and J. Sun, “SafeDrug: Dual molecular graph encoders for recommending effective and safe drug combinations,” in *Proc. 30th Int. Joint Conf. Artificial Intelligence (IJCAI-21)*, 2021, pp. 3735–3741, doi: 10.24963/ijcai.2021/514.
- [9] C. Yang, C. Xiao, L. Glass, and J. Sun, “Change matters: Medication change prediction with recurrent residual networks,” in *Proc. 30th Int. Joint Conf. Artificial Intelligence (IJCAI-21)*, 2021, pp. 3728–3734, doi: 10.24963/ijcai.2021/513.
- [10] R. Wu, Z. Qiu, J. Jiang, G. Qi, and X. Wu, “Conditional generation net for medication recommendation,” in *Proc. ACM Web Conf. 2022 (WWW '22)*, 2022, pp. 935–945, doi: 10.1145/3485447.3511936.
- [11] Y. Tan, C. Kong, L. Yu, P. Li, C. Chen, X. Zheng, V. S. Hertzberg, and C. Yang, “4SDrug: Symptom-based set-to-set small and safe drug recommendation,” in *Proc. 28th ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD '22)*, 2022, pp. 3970–3980, doi: 10.1145/3534678.3539089.
- [12] N. Yang, K. Zeng, Q. Wu, and J. Yan, “MoleRec: Combinatorial drug recommendation with substructure-aware molecular representation learning,” in *Proc. ACM Web Conf. 2023 (WWW '23)*, 2023, pp. 4075–4085, doi: 10.1145/3543507.3583872.
- [13] S. Bhoi, M. L. Lee, W. Hsu, and N. C. Tan, “REFINE: A fine-grained medication recommendation system using deep learning and personalized drug interaction modeling,” in *Advances in Neural Information Processing Systems* 36 (NeurIPS 2023), 2023.
- [14] C. Zhao, H. Zhao, X. Zhou, and X. Li, “Enhancing precision drug recommendations via in-depth exploration of motif relationships,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 12, pp. 8164–8178, Dec. 2024, doi: 10.1109/TKDE.2024.3437775.
- [15] J. Zhang, X. Zang, H. Chen, X. Yan, and B. Tang, “Revisiting drug recommendation from a causal perspective,” *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 2, pp. 1525–1533, Feb. 2025, doi: 10.1109/JBHI.2024.3471637.
- [16] G.-F. Ma, X.-H. Yang, X. Wen, Y. Zhou, H. Long, and W. Huang, “Contrastive deep fusion-based diffusion multiview drug recommendation,” *IEEE Transactions on Computational Social Systems*, early access, 2025, doi: 10.1109/TCSS.2025.3641219.

- [17] B. Nemade, S. Alegavi, and V. Bharadi, "Graph attention dialog network-based drug recommendation model for next-gen healthcare and consumer-centric devices," *IEEE Transactions on Consumer Electronics*, vol. 71, no. 2, pp. 5908–5916, May 2025, doi: 10.1109/TCE.2025.3540448.
- [18] Y. He, S. Dong, Y. Lin, X. Zheng, and J. Hu, "Pretraining-based relevance-aware visit similarity network for drug recommendation," *IEEE Journal of Biomedical and Health Informatics*, vol. 30, no. 7, pp. 5542–5555, Jul. 2026, doi: 10.1109/JBHI.2025.3590391.
- [19] Y.-Y. Ou, C.-Y. Chen, and J.-F. Wang, "MedFusion-Agent: A multi-agent framework for conflict-aware clinical pharmacotherapy planning in multimorbidity," *IEEE Access*, vol. 14, pp. 93913–93934, 2026, doi: 10.1109/ACCESS.2026.3704838.
- [20] Y. Meng, Z. Zhang, X. Shang, X. Tang, J. Li, Z. Zhang, F. Cui, S. Jin, J. Xu, and P. Wang, "FusionMVSA: Multi-view fusion strategy with self-attention for enhancing drug recommendation," *IEEE Journal of Biomedical and Health Informatics*, vol. 30, no. 2, pp. 1746–1758, Feb. 2026, doi: 10.1109/JBHI.2025.3586758.
- [21] D. T. Pham and H. T. Phan, "HyBLRecs: A hybrid drug recommendation system integrating content analysis and reliability enhancement," *Journal of Information and Telecommunication*, published online 3 Jul. 2026, doi: 10.1080/24751839.2026.2690713.
- [22] H. Zhou, F. Liu, J. Wu, W. Zhang, G. Huang, L. Clifton, D. Eyre, H. Luo, F. Liu, K. Branson, P. Schwab, X. Wu, Y. Zheng, A. Thakur, and D. A. Clifton, "A collaborative large language model for drug analysis," *Nature Biomedical Engineering*, vol. 10, pp. 870–881, May 2026, doi: 10.1038/s41551-025-01471-z.
- [23] E. Daglarli, "Drug and dosage recommendation based on explainable generative AI using patient-specific modeling," *IEEE Access*, vol. 14, pp. 14965–14984, 2026, doi: 10.1109/ACCESS.2026.3657398.
- [24] S. Naveed, M. Husnain, A. Ikram, S. Ikram, M. J. Babar, and N. Alsubaie, "Transformer-inspired deep learning framework for multi-label drug recommendation using clinical notes: A step toward personalized medicine," *IEEE Access*, vol. 14, pp. 46590–46606, 2026, doi: 10.1109/ACCESS.2026.3672010.
- [25] M. Balossini, D. A. Gervasio, E. G. Caiani, S. Zanero, M. Carminati, and S. Longari, "Evaluating threats and proposing evidence-based recommendations for mobile digital therapeutics applications," *IEEE Access*, vol. 14, pp. 75749–75762, 2026, doi: 10.1109/ACCESS.2026.3693997.
- [26] A. K. Pandey, B. Bera, A. Vangala, A. K. Das, Y. Park, and B. Sikdar, "Quantum-safe digital twin authentication for ML-driven early disease detection in healthcare systems," *IEEE Journal of Biomedical and Health Informatics*, early access, 2026, doi: 10.1109/JBHI.2026.3667781.
- [27] Z. Zhao, C. Fan, C. Gao, F. Feng, and X. He, "Addressing overprescribing challenges: Fine-tuning large language models for medication recommendation tasks," *arXiv:2503.03687*, Mar. 2025.
- [28] X. Wang, H. Wu, Z. Tai, S. Lyu, Q. Lu, Z. Zhao, J. Chi, J. Tian, X.-W. Chang, and Z. Song, "SafeRx-Agent: A knowledge-grounded multi-agent framework for safe and explainable medication recommendation," *arXiv:2605.29146*, May 2026.
- [29] S. Chen, W. Jiang, C. Wang, X. Wu, X. Shi, Y. Zhang, and J. T. Kwok, "RxEval: A prescription-level benchmark for evaluating LLM medication recommendation," *arXiv:2605.14543*, May 2026.
- [30] Y. Wu, T. Chen, I. Hochberg, Z. Sun, R. Edry, Z. Huang, and M. Peleg, "Lessons learned from evaluation of LLM based multi-agents in safer therapy recommendation," *arXiv:2507.10911*, Jul. 2025.