

Supplementary Material: Leveraging Large Language Models for Personalized Parkinson’s Disease Treatment

You can also visit this url <https://github.com/360ZMEM/PD-LLM> for obtaining the code implementation and this supplementary material.

Contents

A	Additional Statistical Information of the PPMI Dataset	2
B	Details for MDP and Reinforcement Learning Baseline	2
B.1	Action Space	2
B.2	RL Implementation Details	4
C	Hyperparameter Choosing and MCTS details	4
D	Cross-LLM Analysis	5
E	Subgroup Analysis	5
F	Details on LLM workflow	6
F.1	Full Prompt	6
F.2	Similar Case Retrieval	10
G	More Information of Case study	11
G.1	PATNO 3500 - Early Term (2nd Clinical Visit)	13
G.1.1	Relative Cases	13
G.1.2	Answer from LLM	14
G.2	PATNO 3500 - Medium Term (6th Clinical Visit)	14
G.2.1	Relative Cases	14
G.2.2	Answer from LLM	15
G.3	PATNO 3500 - Late Term (9th Clinical Visit)	16
G.3.1	Relative Cases	16
G.3.2	Answer from LLM	17
G.4	PATNO 3507 - Early Term (1st Clinical Visit)	18
G.4.1	Relative Cases	18
G.4.2	Answer from LLM	18
G.5	PATNO 3507 - Late Term (6th Clinical Visit)	19
G.5.1	Relative Cases	19
G.5.2	Answer from LLM	21
H	VSS Robustness Analysis	22
H.1	Differing Patient Robustness	22
H.2	Outlier Robustness Analysis	22

A Additional Statistical Information of the PPMI Dataset

In this section, we summarize some statistical details regarding clinical visits from the PPMI dataset.

When filtering the dataset, to ensure that the RL Baseline has sufficient historical reference data and to allow the LLM to perform a certain depth of simulation to complete the MCTS process, we selected patients with at least 7 valid clinical visit records, totaling 399 patients. Figure 1(a) illustrates the distribution of clinical visit counts among these 399 patients, with the maximum number of valid clinical visits recorded being 17, and the average and standard deviation being 12.0 and 2.9, respectively.

Additionally, we analyze the time intervals between clinical visits, with the distribution shown in Figure 1(b). The average time interval was 9.2 months, with a standard deviation of 5.3 months.

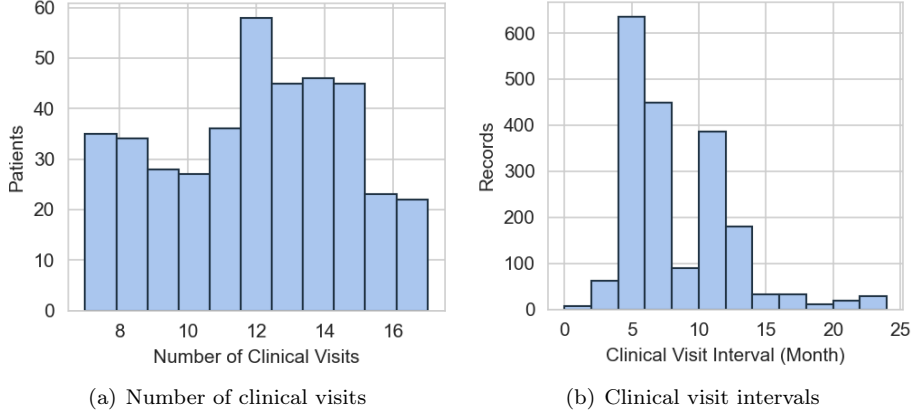


Figure 1: Statistical metrics of clinical visits in PPMI dataset.

B Details for MDP and Reinforcement Learning Baseline

B.1 Action Space

The action space of RL and our MCTS is discrete. It may not be too small, as some advantageous prescription combinations might be excluded; however, an excessively large action space will lead to a sharp increase in MCTS search costs. To achieve a balance of action space scale, we specify several basic single-dose strategies, which are then combined by a certain dose frequency. In this study, we define each action as a combination of the following six types:

- **L-dopa IR**: 100 mg, 250 mg.
- **L-dopa CR**: 100 mg, 200 mg.
- **Rytary**: 95 mg, 195 mg.

Table 1 Available prescriptions for RL/MCTS (73)

<p> [] (No medication) [[100,100,250,250],[],[]] [[100,0],[0,100],[]] [[100,100,0,0],[0,0,100,100],[]] [[100,0],[0,200],[]] [[100,100,0,0],[0,0,200,200],[]] [[100,0],[],[0,95]] [[100,100,0,0],[],[0,0,95,95]] [[100,0],[],[0,195]] [[100,100,0,0],[],[0,0,195,195]] [[250,250,100,100],[],[]] [[250,0],[0,100],[]] [[250,250,0,0],[0,0,100,100],[]] [[250,0],[0,200],[]] [[250,250,0,0],[0,0,200,200],[]] [[250,0],[],[0,95]] [[250,250,0,0],[],[0,0,95,95]] [[250,0],[],[0,195]] [[250,250,0,0],[],[0,0,195,195]] [[0,100],[100,0],[]] [[0,0,100,100],[100,100,0,0],[]] [[0,250],[100,0],[]] [[0,0,250,250],[100,100,0,0],[]] [[],[100,100,200,200],[]] [[],[100,0],[0,95]] [[],[100,100,0,0],[0,0,95,95]] [[],[100,0],[0,195]] [[],[100,100,0,0],[0,0,195,195]] [[0,100],[200,0],[]] [[0,0,100,100],[200,200,0,0],[]] [[0,250],[200,0],[]] [[0,0,250,250],[200,200,0,0],[]] [[],[200,200,100,100],[]] [[],[200,0],[0,95]] [[],[200,200,0,0],[0,0,95,95]] [[],[200,0],[0,195]] [[],[200,200,0,0],[0,0,195,195]] </p>	<p> [[0,100],[],[95,0]] [[0,0,100,100],[],[95,95,0,0]] [[0,250],[],[95,0]] [[0,0,250,250],[],[95,95,0,0]] [[],[0,100],[95,0]] [[],[0,0,100,100],[95,95,0,0]] [[],[0,200],[95,0]] [[],[0,0,200,200],[95,95,0,0]] [[],[],[95,95,195,195]] [[0,100],[],[195,0]] [[0,0,100,100],[],[195,195,0,0]] [[0,250],[],[195,0]] [[0,0,250,250],[],[195,195,0,0]] [[],[0,100],[195,0]] [[],[0,0,100,100],[195,195,0,0]] [[],[0,200],[195,0]] [[],[0,0,200,200],[195,195,0,0]] [[],[],[195,195,95,95]] [[100],[],[[]]] [[100,100],[],[[]]] [[100,100,100,100],[],[[]]] [[250],[],[[]]] [[250,250],[],[[]]] [[250,250,250,250],[],[[]]] [[],[100],[[]]] [[],[100,100],[[]]] [[],[100,100,100,100],[[]]] [[],[200],[[]]] [[],[200,200],[[]]] [[],[200,200,200,200],[[]]] [[],[],[95]] [[],[],[95,95]] [[],[],[95,95,95,95]] [[],[],[195]] [[],[],[195,195]] [[],[],[195,195,195,195]] </p>
---	--

Given that patients may benefit from two-drug combinations, we define 73 available actions in this study. These actions are categorized based on the frequency of daily dosing, as follows:

- **No medication** (1 action).
- **Once daily** (6 actions).
- **Twice daily**, with equal doses if the same drug is administered twice in one day (30 actions).
- **Four times daily**, with identical types and dosages for the first two and last two doses (36 actions).

We also provide a list of all available actions in Table 1. We express them in a format as lists of three sublists in Python style, which include L-dopa IR, L-dopa CR, and Rytary medication data. The sublist shows the dosage and frequency of each drug in milligrams. For example, the treatment plan

[[200, 0], [], [0, 95]] indicates that for L-dopa IR and Rytary, two equally spaced doses are administered during the 17 waking hours each day, with the first dose being 200 mg of L-dopa IR. The last dose of Rytary is 95 mg.

Notice that LLM can directly output prescription plans and will not be restricted by action space listed above.

B.2 RL Implementation Details

This section provides some implementation details of the RL baseline. The observation space of the RL model we employed is a subset of the observation space of the predictive model for MDS-UPDRS-III scores. Specifically, it includes the patient’s vector of covariates (gender, age, years since PD diagnosis) and the patient’s MDS-UPDRS-III scores over the past five clinical visits. Additionally, we added cluster index obtained through K-means clustering (with $K = 4$) into the state representation to enhance the personalization of medication strategies for patients.

We use the reward function utilized by Baucum et al. The reward is determined based on the patient’s current bradykinesia score b_t (L-dopa medication can reduce the slowing of movement) and dyskinesia score d_t (L-dopa medication may induce involuntary movements). The reward function is defined as follows:

$$r(s_t) = \mathbb{I}_{\tanh(b_t) \leq \bar{b}} - \alpha \cdot \mathbb{I}_{d_t > \bar{d}}, \quad (1)$$

where \mathbb{I} denotes the indicator function. In other words, a reward is applied if $\tanh(b_t)$ is less than or equal to $\bar{b} = -0.279$ (corresponding to a unnormalized score of 25), and a penalty is applied if d_t exceeds $\bar{d} = -0.775$ (corresponding to a unnormalized score of 9). The coefficient α adjusts the weight of the dyskinesia penalty, and we set $\alpha = 0.5$ in this study.

Additionally, the RL method we used aligns with that of Baucum et al., namely a general approach of actor-critic implementation.

C Hyperparameter Choosing and MCTS details

In this section, we select several representative parameters to investigate the impact of their variations, and evaluate the robustness of our LLM workflow.

Number of Similar Cases. In our paper, we default to providing two similar cases. We integrate varying numbers of similar cases into the prompt, and the results are shown in Table 2. As the number of similar cases increases, the performance of the LLM improves to some extent. However, when the number of similar cases exceeds four, the performance begins to decline. Additionally, a higher number of cases leads to an increase in the number of input tokens.

Number of MCTS Simulations. The number of simulations in MCTS affects the accuracy of the final evaluation of different medication decisions. However, setting this number too high can lead to a significant increase in computational costs. In our paper, we defaulted to 2500 simulation trials. Table 3 shows the performance as the number of trials is reduced (with the number of similar cases remaining at two). It can be seen that reducing the number of trials to one-fourth of the original does not significantly degrade performance, but further reduction leads to noticeable performance degradation.

Table 2: Key metrics of MDS-UPDRS-III Change (ΔU) with different number of similar cases.

Number of cases	Average↓	Median↓
1	1.96	2.26
2	1.83	1.86
3	1.82	1.91
4	1.75	1.79
5	1.80	1.80

Additional MCTS details. Here we elaborate on additional technical details of MCTS not explicitly mentioned in the main text. First, in the UCT formula, we set $C_p = 2$, as we put more emphasis on

Table 3: Key metrics of MDS-UPDRS-III Change (ΔU) with different number of MCTS simulations per records.

Number of simulations	Average↓	Median↓
2500	1.83	1.86
1200	1.89	2.02
600	2.12	2.35
300	2.40	2.55
150	2.66	2.80

Table 4: Key metrics of MDS-UPDRS-III change (ΔU) for various LLMs.

Metric	Average↓	Median↓	25th percentile↓	75th percentile↓
DeepSeek-V2.5	1.83	1.86	-3.83	8.14
DeepSeek-V3	1.74	1.69	-4.01	8.20
Qwen2.5-72B-Instruct	1.89	2.00	-3.59	7.93
Qwen3-235B-A22B	1.66	1.77	-3.72	7.68

prescription outcome rather than exploration. After completing a fixed number of simulations via MCTS, a similar case dataset is generated (see Section F.2 for detailed format specifications). By default, this dataset includes four recommended prescriptions and four unrecommended prescriptions derived from first-level child state nodes. Each medication prescription is followed by the averaged MDS-UPDRS-III score changes recorded during three follow-up assessments from corresponding medical records.

D Cross-LLM Analysis

In this section, we conducted additional evaluations using the following open-source LLMs under the same experimental setting:

- DeepSeek-v3 (DeepSeek, 2024)
- Qwen2.5-72B-Instruct (Alibaba, 2024)
- Qwen3-235B-A22B (Alibaba, 2025)

The comparative results are summarized in Table 4.

While DeepSeek-v3 and Qwen3 show slightly better performance in average/median scores, the overall performance differences are marginal, and all models exhibit highly similar interquartile ranges. These results suggest that general-purpose LLMs—without domain-specific medical grounding—achieve comparable performance, with limited differentiation in aggregate metrics.

In our framework, the role of the LLM is not to dominate the prediction task but to enhance interpretability and decision support, especially in boundary conditions where multiple sources of clinical information need to be integrated and interpreted coherently. In these scenarios, the LLM serves as a reasoning layer that provides explainable justifications to assist physicians. Given this role, DeepSeek-v2.5 offers a favorable trade-off between effectiveness and computational efficiency, and remains the default model in our deployment.

E Subgroup Analysis

In this section, we evaluate the model’s performance across two clinically relevant subgroup dimensions: gender and age. The gender distribution includes 252 males and 147 females. For age, we group patients based on age at diagnosis into four bins:

- (0, 55] years (71 patients)

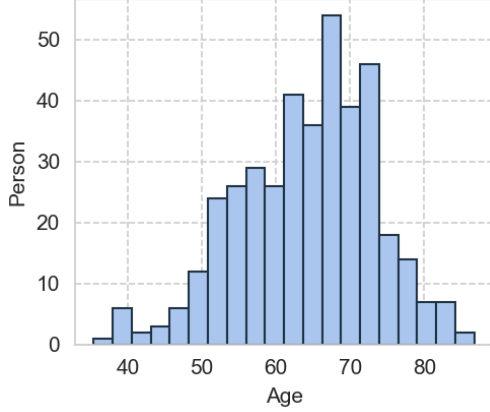


Figure 2: Age (at first clinical visit) distribution

Table 5: Key metrics of MDS-UPDRS-III change (ΔU) for male and female subgroups.

Gender	Metrics	Average ↓	Median ↓
Male	Real	6.44	6.00
	Physician	4.74	5.61
	LLM	3.18	2.73
	RL	3.77	3.68
Female	Real	2.99	1.00
	Physician	0.72	1.03
	LLM	-0.49	-0.26
	RL	1.28	1.96

- (55, 65] years (127 patients)
- (65, 75] years (163 patients)
- (75, ∞) years (38 patients)

To evaluate subgroup robustness and potential fairness concerns, we analyzed the model’s performance across these cohorts. Figure 2 shows the age distribution of patients, while Tables 5 and 6 summarize the average and median MDS-UPDRS-III score changes under different treatment strategies within each subgroup.

These findings demonstrate that our LLM-based system maintains reasonable robustness across diverse demographic subgroups, and exhibits lower performance variance than the RL baseline, supporting its potential for more equitable clinical decision support.

F Details on LLM workflow

F.1 Full Prompt

In this section, we present the complete Prompt integrated with medical knowledge used in our experiments. Specifically, the complete Prompt includes a comprehensive explanation of the structured text, along with corresponding empirical descriptions, and provides a detailed introduction to the steps that the output should follow, ultimately generating structured medication information.

Thanks to the highly interpretable text, this part of the Prompt can be modified according to user needs. In our experiments, this part of the Prompt will be input as the system instruction, while the user input will consist of the patient’s structured text information, similar case text information, and a recommended

Table 6: Key metrics of MDS-UPDRS-III change (ΔU) for different age subgroups.

Age	Metrics	Average ↓	Median ↓
[0,55]	Real	0.59	0.00
	Physician	0.67	1.82
	LLM	-1.00	-0.45
	RL	1.24	-0.21
(55,65]	Real	3.20	2.00
	Physician	1.70	2.47
	LLM	0.94	1.05
	RL	2.06	2.98
(65,75]	Real	7.64	6.00
	Physician	4.41	5.12
	LLM	2.91	2.44
	RL	3.25	3.38
(75,∞)	Real	9.70	10.00
	Physician	8.34	10.06
	LLM	5.45	6.11
	RL	5.88	5.98

medication list. This means that the Prompt shown below does not need to change with variations in patient information.

```

1
2 You are a doctor specializing in Parkinson's disease treatment. We present
  → information about a given patient, along with some machine-learning
  → (ML)-generated treatment plans for other similar cases. You need to analyze the
  → patient information and similar cases, and subsequently provide the formatted
  → prescription following format specified below.
3
4 ## Explanations of related terms
5
6 1. MDS-UPDRS-III Score: The score of revised unified Parkinson's disease rating
  → scale part III motor examination (MDS-UPDRS-III) ranges from 0 to 53
  → (99th-percentile), with higher scores indicating more severe symptoms. Each
  → adjacent score is evaluated approximately a year apart.
7
8 2. Nonlev: Nonlev is the sum of the daily doses of other non-levodopa drugs (e.g.,
  → dopamine and MAO-B inhibitors) prescribed by a physician and is measured as the
  → daily equivalent of L-dopa in milligrams.
9
10 3. The medication history followed the same format as the given treatment plans for
  → Levodopa and equivalent Levodopa doses for Nonlev and was also prescribed
  → approximately a year apart.
11
12 4. The treatment plan is a list of three sublists, which include L-dopa IR
  → ("immediate release"), L-dopa CR ("controlled release"), and Rytary (which is an
  → extended-release form of L-dopa) medication data. These medications differ in
  → how quickly they release L-dopa into patients' bloodstreams. The sublist shows
  → the dosage and frequency of each drug in milligrams. For example, the treatment
  → plan `[[200, 0, 0], [], [0, 95, 95]]` indicates that for L-dopa IR and Rytary,
  → three equally spaced doses are administered during the 17 waking hours each day,
  → with the first dose being 200 mg of L-dopa IR. The last two doses of Rytary are
  → 95 mg each.

```

5. We use ML methods to obtain prescriptions that provide four alternatives
 → classified as recommended and not recommended, respectively. Specifically, each
 → alternative is accompanied by the predicted MDS-UPDRS-III score after 3 years as
 → predicted by the neural network.

6. Patients who have not been diagnosed with PD will be listed as `N/A` in the
 → `Years_sincePD` metric. These patients may be prescribed medication for
 → prevention or may exhibit some motor abnormalities and disorders, the extent of
 → which might be reflected in the MDS-UPDRS-III score.

Factors to consider

We present some of the main considerations when deciding which medication plan to
 → prescribe to a given patient, with the intention of illustrating how these
 → factors guide prescription decisions. Specifically, this includes several
 → aspects:

The Patient's Information: Age, gender, number of years since the diagnosis of
 → PD, medication history, and history of disease progression obtained from the
 → MDS-UPDRS-III score may significantly influence prescription preferences,
 → including the suitability and incompatibility of certain drug combinations.

Similar Cases: We also provide several similar cases and the corresponding
 → recommended treatment plans, specifically:

- For each similar case, the patient's information is presented in the same format
 → as the given patient. The information is generally similar to the patient being
 → prescribed, but there are still some differences.
- Several recommended and unrecommended treatment plans are given for each similar
 → case, and MDS-UPDRS-III scores predicted by machine learning are also provided.
 → These treatment plans should also be appropriate to a certain extent for the
 → given patient. Specifically, these plans are judged not only by predicted
 → MDS-UPDRS-III scores, where differences greater than ± 3 are generally considered
 → significant, but also by considering the potential differences in subsequent
 → disease development these medication recommendations may cause and whether they
 → align with the actual condition of the patient. After conducting a comprehensive
 → analysis, the recommended plans for certain similar cases with higher priority
 → may be selected as the basis for modification.
- It should be noted that due to the differences that still exist in similar cases,
 → there will be variations in the outcome of treatment plan candidates, which
 → should also be taken into account.

Final Output: The prescription for the final output should strictly follow the
 → format described below.

Output Guide and Format

Your final output should contain only a simple analysis of about 200 words, and a
 → Python code block containing the final recommended prescription after an
 → additional explanation.

31 The brief analysis includes the following: (a) A brief analysis of Important
 ↳ information about the patient that may affect the decision of the treatment; (b)
 ↳ A brief analysis of similar cases and an overall analysis of recommended
 ↳ treatment plans. (c) Based on the comprehensive analysis, indicate the final
 ↳ prescription strategy and whether it is modified based on a certain treatment
 ↳ plan from the similar cases. (d) What adjustments (alter the frequency/dosage,
 ↳ change the medicine type...) and medication strategies have been made to further
 ↳ improve the potential effect to lower the MDS-UPDRS-III score.

32

33 Then, based on the above analysis, the prescription which could maximize the
 ↳ decrease of MDS-UPDRS-III score should be obtained from similar cases or
 ↳ information from the given patient. The prescription contains the frequency and
 ↳ dosage of L-dopa IR, L-dopa CR, and Rytary intake on a daily basis. You should
 ↳ output the additional textual summary to specifically specify the dose and
 ↳ frequency of each medicine, then output the corresponding formatted prescription
 ↳ in python code block, as we explained in the 4) of the section "Explanations of
 ↳ related terms".

34

35 For example, for following ****textual**** explanation:

36

37 - ****Frequency:**** 3 doses (2 valid doses)
 38 - ****Medicine and Dosage:**** (1) Morning: 200mg L-dopa IR (2) Noon: N/A (3) Evening:
 ↳ 95mg Rytary
 39 - ****Additional comment:**** ...

40

41 The corresponding ****formatted**** output is:

42

43 ````python`
 44 `[[200, 0, 0], [0, 0, 0], [0, 0, 95]]`
 45 `````

46

47 Note that treatment plans for the given patient include a range of possible
 ↳ prescription. However, you should analyze the patient's information, evaluating
 ↳ possible differences and choosing the treatment that is most promising to reduce
 ↳ the patient's MDS-UPDRS-III in a long-term perspective.

48

49 **## Notes**

50

51 - 1 dose / 2 doses (Morning, Evening) / 3 doses (Morning, Noon, Evening) / 4 doses
 ↳ (Morning, Noon, Afternoon, Evening)
 52 - For non-levodopa (Nonlev) drugs, no advice is required, as this part is directly
 ↳ prescribed by professional physicians.
 53 - To facilitate pattern matching, you should only output one python block in the
 ↳ ``### Formatted Treatment Plan`` section.
 54 - For ML-predicted prescriptions, as the decision space is discrete and limited,
 ↳ there may be other doses or frequencies that work better but are not included in
 ↳ the recommended prescription list.
 55 - The sample prescriptions given in the above (for example, ``[[200, 0, 0], [], [0,`
 ↳ `95, 95]]`` and ``[[[], [100, 100, 0, 0], [0, 0, 245, 245]]``) are only used to
 ↳ illustrate the output format and meaning and ****DO NOT**** constitute any treatment
 ↳ recommendations. You should prescribe based on the actual situation.

56

Also, we use a simple user input template as follows:

```

2  ## Patient Information to be Analyzed
3
4  The information of patient you need to analyze is:
5
6  <patient_information>
7
8  ## Relative cases
9
10 Information of several similar patients given below, and we adopt ML methods to
    ↳ predict the recommended treatment plans, and obtain explanations and suggestions
    ↳ output by LLMs.
11
12 <similar_cases>
13

```

F.2 Similar Case Retrieval

In this section, we provide key implementation details regarding VSS.

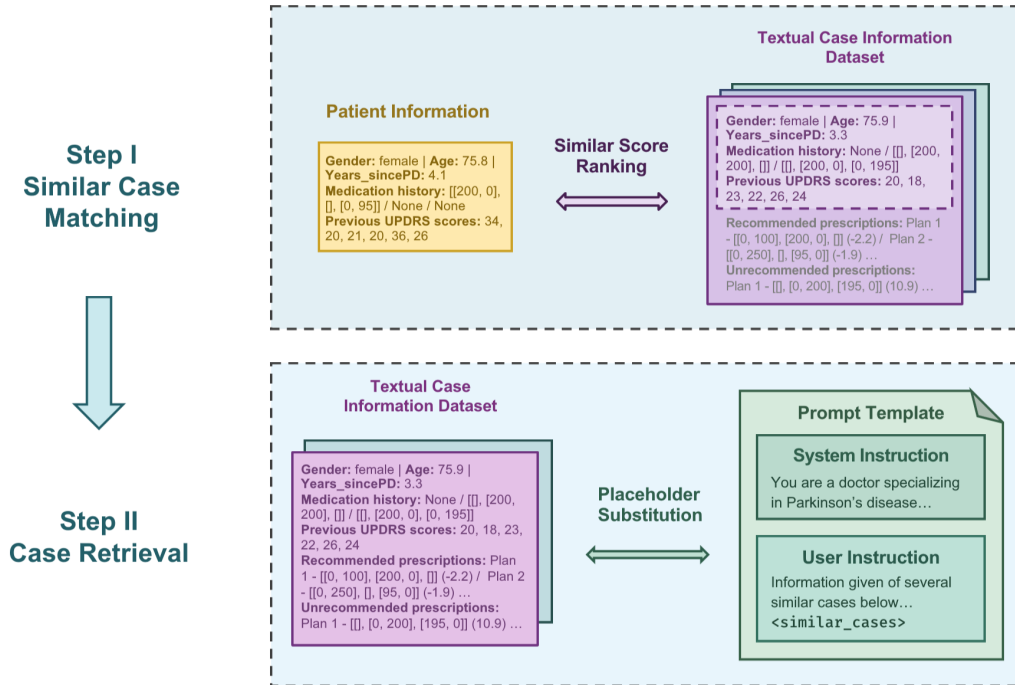


Figure 3: A schematic illustration of similar case matching and case retrieval.

During the MCTS phase, we generate a dataset of similar cases that includes patient information and recommended prescription details.

```

1
2 Gender: female | Age: 67.6 | Years_sincePD: N/A
3 Medication history (from old to new): None / None / [[100, 100, 100], [], []] /
    ↳ [[200, 200], [], []] / [[200, 200], [], []]
4 Nonlevodopa dose (from old to new): 0.0, 0.0, 0.0, 0.0, 8.0
5 Six previous cycles of UPDRS scores (from old to new): 17.0, 22.0, 21.0, 23.0, 20.0,
    ↳ 22.0

```

```

6 Recommended prescriptions generated by ML method: Plan 1 - [[], [100, 100, 0, 0],
  ↳ [0, 0, 95, 95]] (-4.218) / Plan 2 - [[], [200, 0], [0, 195]] (-4.134) / Plan 3
  ↳ - [[0, 0, 250, 250], [], [195, 195, 0, 0]] (-3.758) / Plan 4 - [[], [200, 200,
  ↳ 0, 0], [0, 0, 195, 195]] (-3.652) /
7 Example prescriptions not recommended: Plan 1 - [[100], [], []] (7.523) / Plan 2 -
  ↳ [[], [], [95, 95, 95, 95]] (5.921) / Plan 3 - [[], [200], []] (5.675) / Plan 4
  ↳ - [[100, 0], [0, 200], []] (5.348) /
8
9 ----
10 Gender: female | Age: 69.6 | Years_sincePD: N/A
11 Medication history (from old to new): None / None / None / None / None
12 Nonlevodopa dose (from old to new): 0.0, 0.0, 0.0, 0.0, 0.0
13 Six previous cycles of UPDRS scores (from old to new): 29.0, 37.0, 37.0, 33.0, 44.0,
  ↳ 40.0
14 Recommended prescriptions generated by ML method: Plan 1 - [[], [], [195, 195, 95,
  ↳ 95]] (-3.746) / Plan 2 - [[100, 100, 0, 0], [], [0, 0, 195, 195]] (-3.639) /
  ↳ Plan 3 - [[], [200, 200], []] (-3.054) / Plan 4 - [[100, 100], [], []] (-2.92)
  ↳ /
15 Example prescriptions not recommended: Plan 1 - [[], [200, 200, 200, 200], []]
  ↳ (5.503) / Plan 2 - [[0, 100], [], [95, 0]] (5.329) / Plan 3 - [[100, 100, 0,
  ↳ 0], [0, 0, 200, 200], []] (5.119) / Plan 4 - [[100, 100, 0, 0], [0, 0, 100,
  ↳ 100], []] (4.057) /
16
17 ----
18
19 ...(to be continued)

```

For similarity matching of cases, we discard the prescription information to maintain text format consistency. Each similar case record is then encoded separately (utilize ---- as delimitator) using an embedding model. Subsequently, the patient’s textual information is encoded using the same model, and several records with the highest cosine similarity scores are selected (records originating from the same PATNO are excluded). After obtaining these similar cases, we select records containing prescription information based on the aforementioned ranking to replace the placeholder `<similar_cases>` in the user instruction. This workflow is summarized in Figure 3. Specific examples will also be presented in Section G.

G More Information of Case study

In this section, we provide more detailed information from Section VI.B of the paper. The medications prescribed by RL and LLM during all clinical visits are shown in Table 7 and Table 8, and Figure 4 illustrates the daily dosage changes of different types of L-dopa over time.

To further elaborate on the source of the medication recommendations provided by LLM, we will present the specific inputs and LLM responses (with system instructions and patient information omitted) for PATNO 3500 during the early, middle, and late phases, and for PATNO 3507 during the early and late phases.

Table 7 RL and LLM’s prescription for PATNO 3500 during the whole process

RL	LLM
[[], [], [95,95]]	[[200,0,0], [], [0,195,195]]
[[100,100,0,0], [0,0,100,100], []]	[[250,0,0], [], [0,195,95]]
[[100,100,0,0], [0,0,100,100], []]	[[200,0], [], [0,195,]]
[[100,100,0,0], [0,0,100,100], []]	[[250,250,0,0], [], [0,0,95,95]]
[[100,0], [0,100], []]	[], [200,200,0], []]
[[100,100,0,0], [0,0,100,100], []]	[[250,250,0], [100,100,0], [0,0,195]]
[[100,0], [0,100], []]	[[250,250,0,0], [0,0,200,200], []]
[[100,100,0,0], [0,0,100,100], []]	[[250,0,0,0], [0,200,200,200], []]
[[100,100,0,0], [0,0,100,100], []]	[[100,0,0,0], [0,200,0,0], [0,0,0,100]]

Table 8 RL and LLM’s prescription for PATNO 3507 during the whole process

RL	LLM
[[0,100], [], [195,0]]	[], [100,100,100], []]
[[0,100], [], [195,0]]	[], [], [0,0,195]]
[], [0,200], [95,0]]	[[100,0,0], [0,200,0], [0,95,95]]
[], [0,200], [95,0]]	[[100,0,0], [0,200,0], [0,0,95]]
[], [0,200], [95,0]]	[[250,250,0,0], [0,0,200,200], []]
[[0,100], [200,0], []]	[[100,100,0,0], [0,0,200,0], [0,0,0,95]]
[[0,100], [200,0], []]	[[200,0,0], [0,100,100], []]
[], [100,0], [0,95]]	[[100,100,0,0], [], [0,0,195,195]]
[[0,100], [200,0], []]	[[100,0,0], [], [0,0,95]]
[], [100,0], [0,95]]	[[100,0,0], [0,200,0], [0,0,95]]

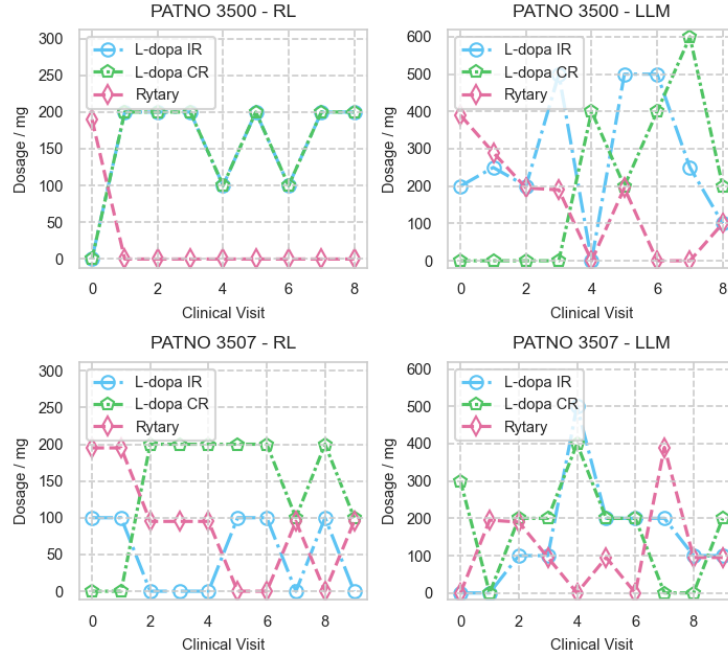


Figure 4: Trends of prescription over time for two representative patients under different experimental settings.

Here we also briefly summarize the key information from these inputs and responses. For patient

3500, since the matched similar cases during the initial treatment phase recommended a combination of IR and Rytary, LLM also adopted this medication strategy. However, to achieve faster therapeutic effects in the early stage, the dosage of IR in the morning was increased, and to maintain sustained disease control, the dosage and frequency of Rytary were appropriately increased. In the middle stage, the first two recommended medication plans from the first similar case were integrated, and the medication frequency was balanced, with a low-dose Rytary supplement added in the evening. In the late stage, the frequency of the recommended medication plans from two similar cases was balanced, and a low-dose Rytary supplement was added in the evening.

For patient 3507, the medication recommendations provided by LLM in the early stage were entirely based on the first recommended medication plan from the first similar case, taking into account that the patient's condition was less severe compared to the similar case, thus reducing the medication frequency and omitting Rytary. In the middle and late stages, to stabilize the disease, the medication strategy was not directly copied from the similar cases but instead adopted a dispersed low-dose medication strategy. In summary, this comprehensive and flexible medication selection strategy significantly enhanced the reliability and effectiveness of our framework.

G.1 PATNO 3500 - Early Term (2nd Clinical Visit)

G.1.1 Relative Cases

```

1
2  ## Relative cases
3
4  Several similar patient information given below, we adopt ML methods to predict the
   → recommended treatment plans, and obtain explanations and suggestions output by
   → LLMs.
5
6  ### Similar Case 1
7  Gender: female | Age: 68.1 | Years_sincePD: N/A
8  Medication history (from old to new): [[100, 100, 100], [], []] / [[200, 200], [],
   → []] / [[200, 200], [], []] / [[[], [100, 100, 0, 0], [0, 0, 95, 95]] / [[[], [100,
   → 100, 0, 0], [0, 0, 195, 195]]
9  Nonlevodopa dose (from old to new): 0.0, 0.0, 8.0, 9.0, 9.0
10 Six previous cycles of MDS-UPDRS-III scores (from old to new): 17.0, 22.0, 21.0,
   → 23.0, 20.0, 22.0
11 Recommended prescriptions generated by ML method: Plan 1 - [[0, 250], [], [195, 0]]
   → (-6.181) / Plan 2 - [[[], [0, 0, 200, 200], [95, 95, 0, 0]] (-5.894) / Plan 3 -
   → [[[], [200, 0], [0, 195]] (-5.616) / Plan 4 - [[250, 250], [], []] (-5.02) /
12 Example prescriptions not recommended: Plan 1 - [[0, 100], [100, 0], []] (5.803) /
   → Plan 2 - [[250, 250, 250, 250], [], []] (3.778) / Plan 3 - [[0, 100], [], [195,
   → 0]] (2.888) / Plan 4 - [[[], [], [95]]] (2.692) /
13 ### Similar Case 2
14 Gender: female | Age: 68.4 | Years_sincePD: N/A
15 Medication history (from old to new): None / None / [[[], [200], []] / [[250, 250, 0,
   → 0], [], [0, 0, 195, 195]] / [[0, 250], [200, 0], []]
16 Nonlevodopa dose (from old to new): 0.0, 0.0, 0.0, 0.0, 0.0
17 Six previous cycles of MDS-UPDRS-III scores (from old to new): 14.0, 17.0, 15.0,
   → 17.0, 19.0, 19.0
18 Recommended prescriptions generated by ML method: Plan 1 - [[0, 250], [], [95, 0]]
   → (-6.315) / Plan 2 - [[100, 0], [0, 200], []] (-5.136) / Plan 3 - [[[], [],
   → [95]]] (-4.863) / Plan 4 - [[250, 250, 100, 100], [], []] (-4.646) /
19 Example prescriptions not recommended: Plan 1 - [[[], [200, 200, 0, 0], [0, 0, 95,
   → 95]]] (3.307) / Plan 2 - [[[], [], [195, 195]]] (2.266) / Plan 3 - [[100, 100, 0,
   → 0], [0, 0, 200, 200], []] (2.243) / Plan 4 - [[0, 0, 250, 250], [], [195, 195,
   → 0, 0]] (2.121) /

```

G.1.2 Answer from LLM

```
1  **(a) Important Information About the Patient:**
2  - The patient is a 68.9-year-old female with an unknown number of years since
   ↳ Parkinson's disease (PD) diagnosis (indicated as N/A).
3  - The patient has a history of stable medication use, primarily involving L-dopa IR
   ↳ and Rytary, with no non-levodopa drugs.
4  - The MDS-UPDRS-III scores have shown a decreasing trend over the past six cycles,
   ↳ indicating some improvement in symptoms.
5
6  **(b) Analysis of Similar Cases and Recommended Treatment Plans:**
7  - **Similar Case 1:** A 68.1-year-old female with similar medication history and
   ↳ stable MDS-UPDRS-III scores. The recommended plans include a mix of L-dopa IR
   ↳ and Rytary, with a focus on higher doses in the morning and evening.
8  - **Similar Case 2:** A 68.4-year-old female with a similar medication history and
   ↳ slightly lower MDS-UPDRS-III scores. The recommended plans also involve a mix of
   ↳ L-dopa IR and Rytary, with a preference for higher morning doses.
9
10 **(c) Final Prescription Strategy:**
11 - Based on the similar cases, the final prescription strategy will focus on higher
   ↳ morning doses of L-dopa IR and Rytary to manage symptoms effectively throughout
   ↳ the day.
12 - The strategy is modified from Similar Case 1's Plan 1, which showed a significant
   ↳ decrease in predicted MDS-UPDRS-III score.
13
14 **(d) Adjustments and Medication Strategies:**
15 - The frequency will be set to 3 doses (Morning, Noon, Evening).
16 - The morning dose will be increased to 250 mg of L-dopa IR to provide a strong
   ↳ initial response.
17 - The noon dose will be adjusted to 195 mg of Rytary to maintain symptom control
   ↳ during the day.
18 - The evening dose will be adjusted to 95 mg of Rytary to ensure continued symptom
   ↳ management through the night.
19
20 ### Formatted Treatment Plan
21
22 ```python
23 [[250, 0, 0], [0, 0, 0], [0, 195, 95]]
24 ```
```

G.2 PATNO 3500 - Medium Term (6th Clinical Visit)

G.2.1 Relative Cases

```
1
2 ## Relative cases
3
4 Several similar patient information given below, we adopt ML methods to predict the
   ↳ recommended treatment plans, and obtain explanations and suggestions output by
   ↳ LLMs.
5
6 ### Similar Case 1
7 Gender: female | Age: 73.9 | Years_sincePD: 5.3
8 Medication history (from old to new): [[100, 100, 100], [], [] / [[100, 100, 100],
   ↳ [], [] / [[250, 250, 0, 0], [0, 0, 200, 200], [] / [[[], [], [95, 95]] / [[250,
   ↳ 250, 0, 0], [0, 0, 100, 100], []]
```

9 Nonlevodopa dose (from old to new): 25.0, 25.0, 17.0, 17.0, 22.0
 10 Six previous cycles of MDS-UPDRS-III scores (from old to new): 35.0, 44.0, 37.0,
 ↳ 52.0, 34.0, 41.0
 11 Recommended prescriptions generated by ML method: Plan 1 - [[0, 0, 100, 100], [100,
 ↳ 100, 0, 0], []] (-7.277) / Plan 2 - [[250, 250, 0, 0], [0, 0, 200, 200], []]
 ↳ (-6.658) / Plan 3 - [[0, 250], [100, 0], []] (-6.384) / Plan 4 - [[], [0,
 ↳ 200], [195, 0]] (-5.944) /
 12 Example prescriptions not recommended: Plan 1 - [[], [100, 100, 100, 100], []]
 ↳ (4.212) / Plan 2 - [[], [0, 100], [95, 0]] (3.481) / Plan 3 - [[250, 250, 0,
 ↳ 0], [0, 0, 100, 100], []] (3.451) / Plan 4 - [[], [200, 200], []] (2.629) /
 13 **### Similar Case 2**
 14 Gender: female | Age: 73.4 | Years_sincePD: 4.8
 15 Medication history (from old to new): [[100, 100, 100], [], []] / [[100, 100, 100],
 ↳ [], []] / [[100, 100, 100], [], []] / [[250, 250, 0, 0], [0, 0, 200, 200], []] /
 ↳ [[], [], [95, 95]]
 16 Nonlevodopa dose (from old to new): 0.0, 25.0, 25.0, 17.0, 17.0
 17 Six previous cycles of MDS-UPDRS-III scores (from old to new): 35.0, 44.0, 37.0,
 ↳ 52.0, 34.0, 41.0
 18 Recommended prescriptions generated by ML method: Plan 1 - [[250, 250, 0, 0], [0, 0,
 ↳ 100, 100], []] (-5.764) / Plan 2 - [[100, 0], [0, 200], []] (-5.615) / Plan 3
 ↳ - [[], [100], []] (-4.717) / Plan 4 - [[0, 250], [], [95, 0]] (-4.7) /
 19 Example prescriptions not recommended: Plan 1 - [[], [200], []] (4.395) / Plan 2 -
 ↳ [[], [0, 100], [195, 0]] (3.496) / Plan 3 - [[], [], [95]] (3.088) / Plan 4 -
 ↳ [[250, 250, 0, 0], [], [0, 0, 95, 95]] (2.904) /

G.2.2 Answer from LLM

1
 2 **### Brief Analysis**
 3
 4 ****(a) Important Information About the Patient:****
 5 - The patient is a 73.8-year-old female diagnosed with Parkinson's disease (PD) 2.2
 ↳ years ago.
 6 - The medication history shows a mix of L-dopa IR, L-dopa CR, and Rytary usage, with
 ↳ varying dosages and frequencies.
 7 - The Nonlevodopa dose has been minimal, suggesting reliance primarily on L-dopa
 ↳ medications.
 8
 9 ****(b) Analysis of Similar Cases and Recommended Treatment Plans:****
 10 - ****Similar Case 1:**** A 73.9-year-old female with 5.3 years since PD diagnosis. The
 ↳ recommended plans include a mix of L-dopa IR and CR, with a focus on higher
 ↳ dosages and more frequent administration.
 11 - ****Similar Case 2:**** A 73.4-year-old female with 4.8 years since PD diagnosis. The
 ↳ recommended plans also involve higher dosages of L-dopa IR and CR, with some use
 ↳ of Rytary.
 12 - Both similar cases show a trend towards higher dosages and more frequent
 ↳ administration to manage more advanced PD symptoms.
 13
 14 ****(c) Final Prescription Strategy:****
 15 - Based on the patient's recent low MDS-UPDRS-III scores and the need to prevent
 ↳ future escalations, a balanced approach is warranted. The strategy will be to
 ↳ maintain a moderate dosage of L-dopa IR and CR, ensuring steady symptom
 ↳ management without overmedicating.

```

16 - The final prescription will be modified from Similar Case 1's Plan 1, which
    ↳ includes a balanced use of L-dopa CR and IR. The dosage and frequency will be
    ↳ adjusted to align with the patient's current needs and prevent potential
    ↳ overmedication.
17
18 **(d) Adjustments and Medication Strategies:**
19 - Increase the dosage and frequency of L-dopa IR and CR to manage symptoms more
    ↳ effectively.
20 - Introduce Rytary in the evening to provide extended relief.
21 - The final prescription will include three doses of L-dopa IR in the morning, noon,
    ↳ and evening, with additional L-dopa CR in the morning and noon, and Rytary in
    ↳ the evening.
22
23
24 - **Frequency:** 3 doses
25 - **Medicine and Dosage:**
26   - (1) Morning: 250mg L-dopa IR, 100mg L-dopa CR
27   - (2) Noon: 250mg L-dopa IR, 100mg L-dopa CR
28   - (3) Evening: 195mg Rytary
29 - **Additional comment:** This plan aims to provide consistent and effective symptom
    ↳ management throughout the day, with higher dosages in the morning and noon to
    ↳ address early symptoms and Rytary in the evening for extended relief.
30
31 ```python
32 [[250, 250, 0], [100, 100, 0], [0, 0, 195]]
33 ```
34

```

G.3 PATNO 3500 - Late Term (9th Clinical Visit)

G.3.1 Relative Cases

```

1
2 ## Relative cases
3
4 Several similar patient information given below, we adopt ML methods to predict the
    ↳ recommended treatment plans, and obtain explanations and suggestions output by
    ↳ LLMs.
5
6 ### Similar Case 1
7 Gender: male | Age: 75.9 | Years_sincePD: 4.3
8 Medication history (from old to new): None / None / [[], [200, 200, 200, 200], []] /
    ↳ [[], [200, 200, 0, 0], [0, 0, 195, 195]] / [[], [100, 100], []]
9 Nonlevodopa dose (from old to new): 0.0, 0.0, 0.0, 0.0, 0.0
10 Six previous cycles of MDS-UPDRS-III scores (from old to new): 20.0, 18.0, 23.0,
    ↳ 22.0, 26.0, 24.0
11 Recommended prescriptions generated by ML method: Plan 1 - [[0, 0, 100, 100], [200,
    ↳ 200, 0, 0], []] (-10.289) / Plan 2 - [[0, 250], [], [95, 0]] (-1.852) / Plan 3
    ↳ - [[], [0, 100], [95, 0]] (-1.608) / Plan 4 - [[100, 100, 100, 100], [], []]
    ↳ (-1.123) /
12 Example prescriptions not recommended: Plan 1 - [[], [0, 0, 200, 200], [195, 195, 0,
    ↳ 0]] (10.921) / Plan 2 - [[], [100, 100, 0, 0], [0, 0, 95, 95]] (10.655) / Plan
    ↳ 3 - [[0, 250], [200, 0], []] (10.46) / Plan 4 - [[250, 0], [], [0, 195]] (9.35)
    ↳ /
13 ### Similar Case 2

```


14 Gender: female | Age: 73.4 | Years_sincePD: 4.8
 15 Medication history (from old to new): [[100, 100, 100], [], []] / [[100, 100, 100],
 ↪ [], []] / [[100, 100, 100], [], []] / [[250, 250, 0, 0], [0, 0, 200, 200], []] /
 ↪ [[], [], [95, 95]]
 16 Nonlevodopa dose (from old to new): 0.0, 25.0, 25.0, 17.0, 17.0
 17 Six previous cycles of MDS-UPDRS-III scores (from old to new): 35.0, 44.0, 37.0,
 ↪ 52.0, 34.0, 41.0
 18 Recommended prescriptions generated by ML method: Plan 1 - [[250, 250, 0, 0], [0, 0,
 ↪ 100, 100], []] (-5.764) / Plan 2 - [[100, 0], [0, 200], []] (-5.615) / Plan 3
 ↪ - [[], [100], []] (-4.717) / Plan 4 - [[0, 250], [], [95, 0]] (-4.7) /
 19 Example prescriptions not recommended: Plan 1 - [[], [200], []] (4.395) / Plan 2 -
 ↪ [[], [0, 100], [195, 0]] (3.496) / Plan 3 - [[], [], [95]] (3.088) / Plan 4 -
 ↪ [[250, 250, 0, 0], [], [0, 0, 95, 95]] (2.904) /

G.3.2 Answer from LLM

```

1  ### Brief Analysis
2
3  **(a) Important Information About the Patient:**
4  - The patient is a 75.8-year-old female diagnosed with Parkinson's disease (PD) 4.1
      ↪ years ago.
5  - The medication history shows a progression from lower to higher doses of L-dopa,
      ↪ with a recent trend towards more frequent dosing.
6  - Nonlevodopa doses have been consistently low, suggesting minimal use of
      ↪ non-levodopa medications.
7
8  **(b) Analysis of Similar Cases and Recommended Treatment Plans:**
9  - **Similar Case 1 (male, 75.9 years, 4.3 years since PD):** The recommended plans
      ↪ focus on higher doses of L-dopa CR and Rytary, with significant predicted
      ↪ MDS-UPDRS-III score reductions. The best plan (Plan 1) suggests a combination of
      ↪ L-dopa CR and Rytary, with a predicted MDS-UPDRS-III score reduction of -10.289.
10 - **Similar Case 2 (female, 73.4 years, 4.8 years since PD):** The recommended plans
      ↪ also emphasize higher doses of L-dopa IR and CR, with moderate predicted
      ↪ MDS-UPDRS-III score reductions. The best plan (Plan 1) suggests a combination of
      ↪ L-dopa IR and CR, with a predicted MDS-UPDRS-III score reduction of -5.764.
11
12 **(c) Final Prescription Strategy:**
13 - Based on the analysis, the final prescription strategy will be modified from
      ↪ Similar Case 1's Plan 1, which showed the most significant predicted
      ↪ MDS-UPDRS-III score reduction.
14
15 **(d) Adjustments and Medication Strategies:**
16 - The frequency will be adjusted to 4 doses per day to better manage symptoms
      ↪ throughout the day.
17 - The dosage of L-dopa CR will be increased to match the recommended plan, and
      ↪ Rytary will be introduced to provide extended relief.
18 - This strategy aims to stabilize and potentially reduce the MDS-UPDRS-III score by
      ↪ ensuring consistent L-dopa levels throughout the day.
19
20 ### Formatted Treatment Plan
21
22 - **Frequency:** 4 doses
23 - **Medicine and Dosage:**
24    - (1) Morning: 100mg L-dopa IR
25    - (2) Noon: 200mg L-dopa CR
  
```

```

26 - (3) Afternoon: 100mg L-dopa IR
27 - (4) Evening: 100mg Rytary
28 - **Additional comment:** This plan is designed to provide balanced and extended
    ↳ coverage of L-dopa throughout the day, aiming to manage symptoms effectively and
    ↳ potentially reduce the MDS-UPDRS-III score.
29
30 ```python
31 [[100, 0, 100, 0], [0, 200, 0, 0], [0, 0, 0, 100]]
32 ```

```

G.4 PATNO 3507 - Early Term (1st Clinical Visit)

G.4.1 Relative Cases

```

1
2 ## Relative cases
3
4 Several similar patient information given below, we adopt ML methods to predict the
    ↳ recommended treatment plans, and obtain explanations and suggestions output by
    ↳ LLMs.
5
6 ### Similar Case 1
7 Gender: male | Age: 73.2 | Years_sincePD: N/A
8 Medication history (from old to new): None / None / None / None / None
9 Nonlevodopa dose (from old to new): 0.0, 0.0, 0.0, 0.0, 0.0
10 Six previous cycles of MDS-UPDRS-III scores (from old to new): 7.0, 6.0, 12.0, 11.0,
    ↳ 7.0, 16.0
11 Recommended prescriptions generated by ML method: Plan 1 - [[], [100, 100, 100,
    ↳ 100], []] (-2.204) / Plan 2 - [[], [100, 100, 200, 200], []] (-1.328) / Plan 3
    ↳ - [[250, 250, 0, 0], [], [0, 0, 195, 195]] (-1.306) / Plan 4 - [[], [0, 100],
    ↳ [95, 0]] (-1.064) /
12 Example prescriptions not recommended: Plan 1 - [[], [0, 0, 200, 200], [195, 195, 0,
    ↳ 0]] (6.462) / Plan 2 - [[], [0, 200], [95, 0]] (5.546) / Plan 3 - [[], [0,
    ↳ 100], [195, 0]] (5.421) / Plan 4 - [[0, 100], [200, 0], []] (5.074) /
13 ### Similar Case 2
14 Gender: male | Age: 72.2 | Years_sincePD: N/A
15 Medication history (from old to new): None / None / None / None / None
16 Nonlevodopa dose (from old to new): 0.0, 0.0, 0.0, 0.0, 0.0
17 Six previous cycles of MDS-UPDRS-III scores (from old to new): 23.0, 22.0, 34.0,
    ↳ 29.0, 43.0, 31.0
18 Recommended prescriptions generated by ML method: Plan 1 - [[0, 0, 100, 100], [200,
    ↳ 200, 0, 0], []] (2.895) / Plan 2 - [[100], [], []] (3.667) / Plan 3 - [[100,
    ↳ 100], [], []] (4.972) / Plan 4 - [[100, 100, 0, 0], [0, 0, 100, 100], []]
    ↳ (5.249) /
19 Example prescriptions not recommended: Plan 1 - [[250, 250, 0, 0], [0, 0, 200, 200],
    ↳ []] (14.811) / Plan 2 - [[], [], [195, 195, 195, 195]] (13.828) / Plan 3 -
    ↳ [[0, 250], [200, 0], []] (13.698) / Plan 4 - [[0, 250], [100, 0], []] (13.682)
    ↳ /

```

G.4.2 Answer from LLM

```

1 ### Brief Analysis
2
3 **(a) Important Information About the Patient:**
4 - The patient is a 70.2-year-old male with no history of Parkinson's disease (PD)
    ↳ diagnosis (Years_sincePD: N/A).

```

```

5 - The patient has no prior medication history for PD, and the Nonlevodopa dose has
  ↳ been consistently 0.0 mg.
6
7 **(b) Analysis of Similar Cases and Recommended Treatment Plans:**
8 - **Similar Case 1:** A 73.2-year-old male with similar non-diagnosis status and a
  ↳ MDS-UPDRS-III trend from 7.0 to 16.0. The recommended plans include L-dopa CR
  ↳ with four doses of 100 mg each (Plan 1) and a combination of L-dopa IR and CR
  ↳ (Plan 2 and Plan 3). These plans are predicted to reduce the MDS-UPDRS-III score
  ↳ significantly.
9 - **Similar Case 2:** A 72.2-year-old male with a more severe MDS-UPDRS-III
  ↳ progression from 23.0 to 31.0. The recommended plans are less effective, with
  ↳ predicted MDS-UPDRS-III scores showing minimal reduction or even slight
  ↳ increases.
10
11 **(c) Final Prescription Strategy:**
12 - Based on the analysis, the patient's condition resembles Similar Case 1 more
  ↳ closely, especially in terms of the MDS-UPDRS-III progression and non-diagnosis
  ↳ status. Therefore, the treatment plan from Similar Case 1 (Plan 1) is chosen as
  ↳ the basis for modification.
13
14 **(d) Adjustments and Medication Strategies:**
15 - The chosen plan from Similar Case 1 involves four doses of L-dopa CR, which is
  ↳ effective but may not be necessary for the current patient due to the less
  ↳ severe progression. Therefore, the frequency is reduced to three doses,
  ↳ maintaining the dosage to ensure efficacy.
16
17 ### Formatted Treatment Plan
18
19 - **Frequency:** 3 doses
20 - **Medicine and Dosage:** (1) Morning: 100mg L-dopa CR (2) Noon: 100mg L-dopa CR
  ↳ (3) Evening: 100mg L-dopa CR
21 - **Additional comment:** This plan balances the need for regular medication with
  ↳ the patient's current condition, aiming to stabilize and potentially reduce the
  ↳ MDS-UPDRS-III score.
22
23 ```python
24 [[0, 0, 0], [100, 100, 100], [0, 0, 0]]
25 ```

```

G.5 PATNO 3507 - Late Term (6th Clinical Visit)

G.5.1 Relative Cases

```

1 ## Relative cases
2
3
4 Several similar patient information given below, we adopt ML methods to predict the
  ↳ recommended treatment plans, and obtain explanations and suggestions output by
  ↳ LLMs.
5
6 ### Similar Case 1
7 Gender: male | Age: 70.7 | Years_sincePD: N/A
8 Medication history (from old to new): [[100, 100, 100], [], []] / [[100, 100, 100],
  ↳ [], []] / [[250, 250, 0, 0], [0, 0, 100, 100], []] / [[], [0, 0, 200, 200],
  ↳ [195, 195, 0, 0]] / [[0, 0, 250, 250], [], [195, 195, 0, 0]]

```

```

9 Nonlevodopa dose (from old to new): 50.0, 50.0, 37.0, 37.0, 37.0
10 Six previous cycles of MDS-UPDRS-III scores (from old to new): 25.0, 41.0, 38.0,
   ↳ 34.0, 35.0, 35.0
11 Recommended prescriptions generated by ML method: Plan 1 - [[100, 0], [0, 200], []]
   ↳ (-3.732) / Plan 2 - [[], [100, 0], [0, 195]] (-3.73) / Plan 3 - [[], [100,
   ↳ 100, 0, 0], [0, 0, 95, 95]] (-3.396) / Plan 4 - [[], [0, 200], [95, 0]]
   ↳ (-2.749) /
12 Example prescriptions not recommended: Plan 1 - [[], [200], []] (9.451) / Plan 2 -
   ↳ [[], [100], []] (7.47) / Plan 3 - [[], [0, 100], [95, 0]] (7.309) / Plan 4 -
   ↳ [[], [], [95, 95]] (7.244) /
13 ### Similar Case 2
14 Gender: male | Age: 62.5 | Years_sincePD: N/A
15 Medication history (from old to new): [[100, 100, 100], [], []] / [[100, 100, 100],
   ↳ [], []] / [[0, 0, 100, 100], [100, 100, 0, 0], []] / [[], [100, 0], [0, 195]] /
   ↳ [[], [100, 100, 0, 0], [0, 0, 195, 195]]
16 Nonlevodopa dose (from old to new): 0.0, 0.0, 0.0, 0.0, 0.0
17 Six previous cycles of MDS-UPDRS-III scores (from old to new): 20.0, 24.0, 15.0,
   ↳ 27.0, 20.0, 25.0
18 Recommended prescriptions generated by ML method: Plan 1 - [[100, 100, 0, 0], [],
   ↳ [0, 0, 195, 195]] (-0.082) / Plan 2 - [[0, 0, 250, 250], [100, 100, 0, 0], []]
   ↳ (0.766) / Plan 3 - [[0, 0, 250, 250], [200, 200, 0, 0], []] (0.915) / Plan 4 -
   ↳ [[250, 250, 0, 0], [], [0, 0, 95, 95]] (1.75) /
19 Example prescriptions not recommended: Plan 1 - [[0, 100], [], [195, 0]] (12.019) /
   ↳ Plan 2 - [[100, 100, 0, 0], [], [0, 0, 95, 95]] (11.517) / Plan 3 - [[], [0,
   ↳ 200], [95, 0]] (10.767) / Plan 4 - [[0, 0, 100, 100], [], [195, 195, 0, 0]]
   ↳ (10.659) /
20
21 The information of patient you need to analyze is:
22
23 Gender: male | Age: 72.7 | Years_sincePD: N/A
24 Medication history (from old to new): [[200, 0, 0], [0, 0, 200], [0, 95, 0]] /
   ↳ [[200, 0, 0], [0, 0, 0], [0, 0, 195]] / [[100, 100, 0], [0, 0, 0], [0, 0, 95]] /
   ↳ [[150, 0, 0], [100, 100, 0, 0], [0, 0, 0, 195]] / [[100, 100, 0, 0], [0, 0, 0,
   ↳ 0], [0, 0, 195, 195]]
25 Nonlevodopa dose (from old to new): 0.0, 0.0, 0.0, 0.0, 0.0
26 Six previous cycles of MDS-UPDRS-III scores (from old to new): 21.0, 0.03, 0.03,
   ↳ 0.03, 48.68, 7.1
27
28
29 ## Relative cases
30
31 Several similar patient information given below, we adopt ML methods to predict the
   ↳ recommended treatment plans, and obtain explanations and suggestions output by
   ↳ LLMs.
32
33 ### Similar Case 1
34 Gender: male | Age: 70.7 | Years_sincePD: N/A
35 Medication history (from old to new): [[100, 100, 100], [], []] / [[100, 100, 100],
   ↳ [], []] / [[250, 250, 0, 0], [0, 0, 100, 100], []] / [[], [0, 0, 200, 200],
   ↳ [195, 195, 0, 0]] / [[0, 0, 250, 250], [], [195, 195, 0, 0]]
36 Nonlevodopa dose (from old to new): 50.0, 50.0, 37.0, 37.0, 37.0
37 Six previous cycles of MDS-UPDRS-III scores (from old to new): 25.0, 41.0, 38.0,
   ↳ 34.0, 35.0, 35.0

```

```

38 Recommended prescriptions generated by ML method: Plan 1 - [[100, 0], [0, 200], []]
   ↳ (-3.732) / Plan 2 - [[], [100, 0], [0, 195]] (-3.73) / Plan 3 - [[], [100,
   ↳ 100, 0, 0], [0, 0, 95, 95]] (-3.396) / Plan 4 - [[], [0, 200], [95, 0]]
   ↳ (-2.749) /
39 Example prescriptions not recommended: Plan 1 - [[], [200], []] (9.451) / Plan 2 -
   ↳ [[], [100], []] (7.47) / Plan 3 - [[], [0, 100], [95, 0]] (7.309) / Plan 4 -
   ↳ [[], [], [95, 95]] (7.244) /
40 ### Similar Case 2
41 Gender: male | Age: 62.5 | Years_sincePD: N/A
42 Medication history (from old to new): [[100, 100, 100], [], []] / [[100, 100, 100],
   ↳ [], []] / [[0, 0, 100, 100], [100, 100, 0, 0], []] / [[], [100, 0], [0, 195]] /
   ↳ [[], [100, 100, 0, 0], [0, 0, 195, 195]]
43 Nonlevodopa dose (from old to new): 0.0, 0.0, 0.0, 0.0, 0.0
44 Six previous cycles of MDS-UPDRS-III scores (from old to new): 20.0, 24.0, 15.0,
   ↳ 27.0, 20.0, 25.0
45 Recommended prescriptions generated by ML method: Plan 1 - [[100, 100, 0, 0], [],
   ↳ [0, 0, 195, 195]] (-0.082) / Plan 2 - [[0, 0, 250, 250], [100, 100, 0, 0], []]
   ↳ (0.766) / Plan 3 - [[0, 0, 250, 250], [200, 200, 0, 0], []] (0.915) / Plan 4 -
   ↳ [[250, 250, 0, 0], [], [0, 0, 95, 95]] (1.75) /
46 Example prescriptions not recommended: Plan 1 - [[0, 100], [], [195, 0]] (12.019) /
   ↳ Plan 2 - [[100, 100, 0, 0], [], [0, 0, 95, 95]] (11.517) / Plan 3 - [[], [0,
   ↳ 200], [95, 0]] (10.767) / Plan 4 - [[0, 0, 100, 100], [], [195, 195, 0, 0]]
   ↳ (10.659) /

```

G.5.2 Answer from LLM

```

1  ### Brief Analysis
2
3  **Important Information about the Patient:**
4  - The patient is a 72.7-year-old male with an unknown number of years since
   ↳ Parkinson's Disease (PD) diagnosis (indicated as N/A).
5  - The patient has been on various L-dopa regimens, including immediate-release (IR)
   ↳ and controlled-release (CR) forms, with no non-levodopa medications prescribed.
6
7  **Analysis of Similar Cases:**
8  - **Similar Case 1:** A 70.7-year-old male with a similar unknown PD duration. The
   ↳ patient has a history of stable MDS-UPDRS-III scores around 35. The recommended
   ↳ plans focus on a balanced use of L-dopa IR and CR, with a preference for lower
   ↳ doses spread throughout the day.
9  - **Similar Case 2:** A 62.5-year-old male with a similar unknown PD duration. The
   ↳ patient has a history of fluctuating MDS-UPDRS-III scores, similar to the given
   ↳ patient. The recommended plans suggest higher doses of L-dopa IR and CR, with a
   ↳ focus on maintaining steady levels throughout the day.
10
11 **Final Prescription Strategy:**
12 - Based on the patient's rapid progression and the need to stabilize symptoms, the
   ↳ strategy will be to use a combination of L-dopa IR and CR to maintain steady
   ↳ dopamine levels.
13 - The prescription will be modified based on Similar Case 1's recommended plan,
   ↳ which suggests a balanced approach with lower doses spread throughout the day.
14
15 **Adjustments and Medication Strategies:**
16 - **Frequency:** 4 doses (Morning, Noon, Afternoon, Evening) to ensure steady
   ↳ dopamine levels.
17 - **Medicine and Dosage:**

```

```

18 - Morning: 100mg L-dopa IR
19 - Noon: 100mg L-dopa IR
20 - Afternoon: 200mg L-dopa CR
21 - Evening: 95mg Rytary
22 - **Additional Comment:** This regimen aims to prevent peak-dose dyskinesia and
    ↪ off-periods by maintaining a steady dopamine level throughout the day.
23
24 ### Formatted Treatment Plan
25
26 ```python
27 [[100, 100, 0, 0], [0, 0, 200, 0], [0, 0, 0, 95]]
28 ```

```

H VSS Robustness Analysis

H.1 Differing Patient Robustness

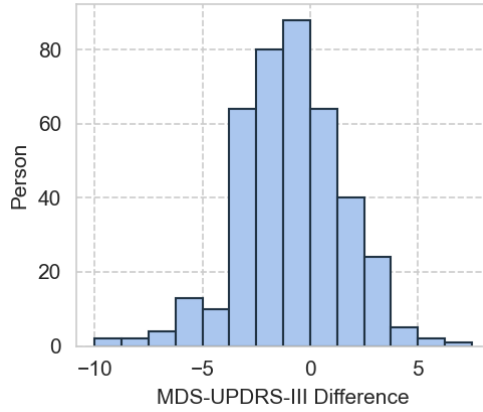


Figure 5: Score difference (PD-LLM vs. w/o VSS) distribution

To assess the credibility of the VSS component, we analyze the performance of VSS across the full evaluation cohort, which includes patients with diverse progression patterns. Specifically, we computed the per-patient difference in MDS-UPDRS-III score changes between the full “PD-LLM workflow” and the “w/o VSS” setting. Negative values indicate improved performance when using VSS. The distribution of these differences is shown in the histogram (Figure 5). As illustrated, the majority of patients benefit from the inclusion of VSS:

- The average improvement of “PD-LLM” over the “w/o VSS” baseline is -0.57 (1.83-2.40), consistent with Table IV of the paper.
- 264 out of 399 patients (66.2%) experienced an improvement (i.e., MDS-UPDRS-III difference < 0), indicating that the benefit of VSS is not limited to a small subset of cases.

These findings demonstrate that the effect of VSS is stable and consistent across a heterogeneous patient population, supporting the credibility of its contribution to the overall workflow.

H.2 Outlier Robustness Analysis

In this section, we conduct a quantitative analysis of similarity-based retrieval coverage across the 399 patients involved in our experiments. Specifically, we computed the maximum cosine similarity between each patient and the retrieved cases in the historical case base, reflecting the semantic closeness of the

Table 7: Key metrics of MDS-UPDRS-III change (ΔU) for different cosine similarity subgroups.

Max Cosine Similarity	Group	Average ↓	Median ↓
(0.955,0.965]	Real	-0.14	2.00
	Physician	-2.68	-1.84
	LLM	-3.5	-1.47
	RL	2.42	3.76
(0.965,0.975]	Real	3.69	2.00
	Physician	2.36	2.15
	LLM	1.06	1.45
	RL	2.08	2.49
(0.975,0.985]	Real	6.10	2.00
	Physician	4.03	4.42
	LLM	2.50	2.27
	RL	3.14	3.22

retrieved support set. The histogram (Figure 6) shows the distribution of these maximum cosine similarity scores, grouped as follows:

- (0.955,0.965]: 21 patients (highly out-of-distribution)
- (0.965, 0.975]: 100 patients (moderate similarity)
- (0.975, 0.985]: 278 patients (high similarity)

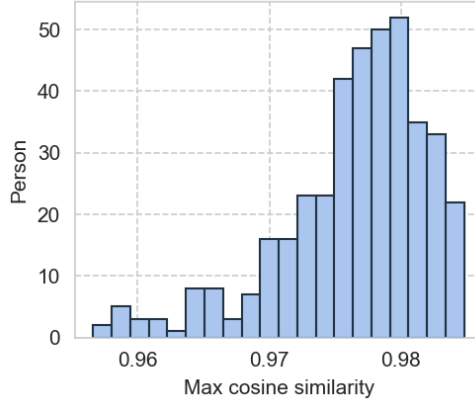


Figure 6: Max cosine similarity distribution

We further examined treatment performance across these three groups using four metrics: real MDS-UPDRS-III score change, physician-prescribed regimens, our LLM-generated treatment plans, and RL-based baselines. The table summarizes both the average and median of MDS-UPDRS-III score changes. The results (Table 7) reveal a clear trend:

Across all similarity groups, the LLM maintains stable and clinically meaningful performance. In contrast, the RL baseline shows substantial degradation in low-similarity groups, likely due to the limited coverage of the training set and its resulting tendency to underfit in out-of-distribution settings. This comparison highlights the greater robustness and generalization ability of the LLM workflow, particularly under data-sparse or atypical patient conditions.