

Supplementary Materials

WikiPlanner: wiki-derived temporal monitors for safe LLM agent planning - a medical-domain proof of concept

Bailing Zhang

Institute of Data Science and Intelligent Technology, NingboTech University, Ningbo 315336, Zhejiang, China.

Correspondence to: Dr. Bailing Zhang, Institute of Data Science and Intelligent Technology, NingboTech University, Ningbo 315336, Zhejiang, China. E-mail: bailing.zhang@nit.zju.edu.cn

This file contains the three items referred to as "Supplementary Material" in the main text: the wiki corpus file list with per-document chunk counts (Supplementary Table 1), the verbatim TC-IR extraction system prompt (Supplementary Text 1), and the complete 40-symbol action vocabulary used for plan generation (Supplementary Table 2). All three are also available in the WikiPlanner repository at <https://github.com/aussie-bzhang/wikiplanner>.

Supplementary Table 1. Wiki corpus: file list and per-document chunk counts

The source corpus comprises 20 de-identified YAML case documents from two educational medical wikis: a general practice (GP) clinic wiki (10 files, GP-001 to GP-010) and a medical school wiki (10 files, MS-001 to MS-010), covering chronic disease management, psychiatry, emergency medicine, pharmacology, and clinical examination skills. Each document was flattened to plain text and split into chunks of at most 1,800 characters at newline boundaries, yielding 85 chunks in total. Each chunk was submitted to the extractor exactly once.

Document	Chunks	Document	Chunks
GP-001.yaml	5	MS-001.yaml	7
GP-002.yaml	3	MS-002.yaml	3
GP-003.yaml	4	MS-003.yaml	3
GP-004.yaml	4	MS-004.yaml	5
GP-005.yaml	2	MS-005.yaml	3
GP-006.yaml	7	MS-006.yaml	3
GP-007.yaml	3	MS-007.yaml	4
GP-008.yaml	8	MS-008.yaml	2
GP-009.yaml	7	MS-009.yaml	2
GP-010.yaml	5	MS-010.yaml	5
GP subtotal	48	MS subtotal	37

Total: 20 documents, 85 chunks. GP: general practice; MS: medical school.

Supplementary Text 1. Verbatim TC-IR extraction system prompt

The following system prompt was used unchanged for every chunk of both the source corpus and the held-out nursing corpus, with GLM-4-flash at temperature 0.1, max_tokens 1,024, up to three retries with exponential back-off on transport errors, and a 1-second inter-call interval. The user message consisted of the chunk text only. The JSON structure specified below is the TC-IR schema of Definition 1 in the main text.

You are a temporal constraint extractor for medical wikis.

Your task: read a medical text and extract ALL temporal ordering constraints between clinical actions.

Return a JSON array. Each element must have EXACTLY these fields:

```
{
  "constraint_type": one of [PREREQUISITE, MANDATORY_FOLLOWUP,
BOUNDED_FOLLOWUP, MUTUAL_EXCLUSION, FORBIDDEN_SEQUENCE, ORDERING,
RESPONSE],
  "trigger_action": "snake_case_action_id",
  "related_actions": {depends on type - see below},
  "natural_language": "exact quote or close paraphrase from text",
  "severity": "hard" or "soft",
  "confidence": float 0.0-1.0,
  "time_bound_steps": null or integer (if BOUNDED_FOLLOWUP),
  "time_bound_hours": null or integer (if time window mentioned),
  "notes": "brief explanation"
}
```

related_actions by type:

- PREREQUISITE: {"prior_action": "action_id"}
- MANDATORY_FOLLOWUP: {"followup_action": "action_id"}
- BOUNDED_FOLLOWUP: {"followup_action": "action_id"} [set time_bound_steps: 1=same visit, 2=next visit, 3=within week]
- MUTUAL_EXCLUSION: {"excluded_action": "action_id"}

- FORBIDDEN_SEQUENCE: {"forbidden_action": "action_id", "guard_action": null or "action_id"}
- ORDERING: {"must_come_after": "action_id"}
- RESPONSE: {"response_action": "action_id"}

Action id rules: lowercase snake_case, max 5 words. Examples:

prescribe_acei, check_egfr, order_blood_test, refer_to_specialist,
monitor_blood_pressure, adjust_dose, schedule_followup, assess_symptoms,
review_medication, stop_medication, start_medication, perform_ecg

If no temporal constraints found, return empty array [].

Return ONLY the JSON array, no explanation, no markdown fences.

Supplementary Table 2. Complete 40-symbol action vocabulary for plan generation

Plan generation used the fixed 40-symbol vocabulary listed below. Generated steps were checked against this list and out-of-vocabulary generations triggered a constrained retry. Of the 61 distinct action symbols referenced by the 45-constraint library, only 15 also occur in this vocabulary; the remaining 46 are wiki-idiosyncratic identifiers. Symbols are grouped by clinical function; the grouping is descriptive only and carries no semantics in the pipeline.

Group	<i>n</i>	Action symbols
Assessment	7	assess_symptoms, take_history, perform_physical_exam, check_blood_pressure, monitor_vitals, assess_fall_risk, check_oxygen_saturation
Investigation	8	order_blood_test, check_egfr, check_renal_function, monitor_inr, check_hba1c, perform_ecg, check_thyroid_function, perform_chest_xray
Prescribing	12	prescribe_acei, prescribe_arb, prescribe_beta_blocker, prescribe_statin, prescribe_diuretic, prescribe_metformin, prescribe_insulin, prescribe_maoi, prescribe_ssri, prescribe_antibiotic, prescribe_steroid, prescribe_warfarin
Specific procedures	7	start_anticoagulation_warfarin,

		initiate_lithium_therapy, identify_sepsis, start_iv_antibiotics, detect_anaphylaxis, administer_epinephrine, administer_iv_contrast adjust_dose, stop_medication, review_medication, monitor_blood_pressure, schedule_followup, refer_to_specialist
Management	6	
Total	40	