# 臨床推理研究治理包 — Failure Shape Transfer

治理文件索引 + 對抗審查對應表 + 病例建構協議 · 2026-05-10

## Paper A Governance Documents — Index

> All governance artifacts for the npj Digital Medicine paper:
> "Failure Shape Transfer Across Cloud, Local, and Agentic Large Language Models in Sequential Clinical Reasoning: A Multi-Stage Evaluation on Taiwan-Specific Vignettes"
>
> Created: 2026-05-10
> Last update: 2026-05-10
> Author: Andrew Huang (M6, NYCU College of Medicine)

---

### Documents

| # | File | Purpose | Action Owner | Status |
|---|---|---|---|---|
| 01 | `01-NYCU-IRB-determination-email.md` | NYCU IRB email draft (中/英) | Andrew | Draft ready, awaiting fill-in |
| 02 | `02-Taipei-VGH-IRB-application-outline.md` | 北榮 Phase 1 + Phase 2 IRB outline | 合作臨床醫師 (with Andrew help) | Draft ready |
| 03 | `03-collaborating clinician-reviewer-charter.md` | Collaborating clinician role boundary contract | Andrew + 合作臨床醫師 sign | Draft ready, awaiting signatures |
| 04 | `04-vignette-construction-protocol.md` | Step-by-step vignette construction rules | Andrew | Locked v1 |
| 05 | `05-vignette-schema.yaml` | JSON Schema for vignette validation | Andrew | Locked v1 |
| 06 | `06-vignette-provenance-ledger-template.md` | Supplementary provenance table | Andrew (one row per vignette) | Template ready |
| 07 | `07-OSF-preregistration-template.md` | OSF pre-registration content | Andrew | Template ready, awaiting OSF submit |

---

### Codex Review Issue Mapping

The 4 critical (🔴) and 11 warning (🟡) issues from Codex review (GPT-5.5, 2026-05-10) are addressed by these documents:

#### 🔴 Critical issues (4)

| Codex finding | Addressed by |
|---|---|
| 1.1 "IRB bypass" wrong frame | Doc 01 (NYCU email) — uses "determination request" language |
| 3.1 PMC license heterogeneity | Doc 04 (protocol) — eliminated PMC case reports as primary source |
| 3.2 Case report consent not transitive | Doc 04 + Doc 06 — provenance ledger flags substantial derivation |
| 4.1 ASUS + real patient data violates Article 4-5 | Doc 04 — ASUS tool excluded entirely |

### 🟡 Warnings (11)

| Codex finding | Addressed by |
|---|---|
| 1.2 "Synthetic" false if traceable | Doc 04 §0 (source restriction) + Doc 06 col 15 |
| 1.3 Collaborating clinician silent case source | Doc 03 (Reviewer Charter) — explicit prohibition |
| 1.4 "Resemblance incidental" weak | Doc 04 — re-identification risk checklist |
| 2.1 Face validity ≠ deployment validity | Doc 07 (OSF) — labeled as construct validity, Phase 1 framing |
| 2.2 Realism rating circularity | Doc 04 §3.7 + Doc 05 + Doc 06 col 19 — blinding required |
| 2.3 Population matching cosmetic | Doc 07 — explicit disclosure of which variables are/aren't matched |
| 2.4 Adversarial overfit | Doc 04 §6 + Doc 07 — pre-register + lock before model run |
| 2.5 RSI inflated naming | Doc 07 — labeled as exploratory, secondary descriptive |
| 3.3 Post-cutoff PMC contamination | Doc 04 — moot now (no PMC case reports), still log model cutoffs |
| 3.4 Domain underpowered | Doc 07 — pre-register no domain-specific claims |
| 3.5 Taiwan registry license | Doc 04 — public aggregate stats only |

### 🔵 Style issues (2)

| Codex finding | Addressed by |
|---|---|
| 2.5 RSI sounds inflated | Doc 07 — secondary descriptive metric |
| 4.4 No audit trail without ASUS | Doc 06 (Provenance Ledger) + git version control |

---

### Workflow

```
W0–1: NYCU IRB email + Reviewer Charter + protocol locked
       ↓
W2–3: Construct 30 vignettes (batched per Doc 04)
       ↓
W4:   Collaborating clinician blinded review + schema validation + lock + OSF pre-register
       ↓
W5–7: Model inference (Paper A main experiment)
       ↓
W8+:  Scoring, analysis, draft
```

---

### Documents NOT Yet Created (P2)

These are needed but later:

- [ ] Scoring Rubric v1 (5-stage rubric for human + LLM judges)
- [ ] LLM Judge Prompt v1 (3-judge consensus protocol)
- [ ] Contamination Detection Script (Python, ~80 lines)
- [ ] Inference Pipeline Code (mlx-lm + cloud API + AgentClinic harness)
- [ ] Statistical Analysis Plan (R / Python notebook)

---

### Cross-References

---

### Version History

| Version | Date | Notes |
|---|---|---|
| v1 | 2026-05-10 | Initial creation after codex-review (GPT-5.5 medium) |

---

### Disclaimers

- These documents reflect best practices as of 2026-05 for medical AI evaluation papers targeting npj Digital Medicine tier
- Final IRB / journal acceptance depends on multiple factors beyond document quality
- Documents should be reviewed by senior PI before submission

---

*來源檔：~/Desktop/local LLM VS Cloud LLM/paper-A-governance/README.md*

# Vignette Construction Protocol v1

> Pre-registered protocol for constructing 30 synthetic clinical vignettes
> Locked before any model inference begins
> Last update: 2026-05-10

---

## 0. Source Restriction (NON-NEGOTIABLE)

Every vignette MUST be derived from at least one of these source types, and ONLY these:

| ✅ Allowed | ❌ Forbidden |
|---|---|
| Standard medical textbooks (Tintinalli's, Harrison's, Williams Obstetrics, Nelson Pediatrics, Greenfield Surgery) | Patient records (any) |
| Public Taiwan CDC disease surveillance reports | Personal clinical memory of any encountered patient |
| Public Taiwan medical society guidelines (TSC, TSEM, TAFM, etc.) | EHR access (any institution) |
| Public WHO / CDC global health pattern reports | Collaborating clinician's personal patient cases |
| Published case series literature (epidemiological patterns only, not derivative case content) | Any data source requiring institutional credentials to access |

**Hard rule**: If during construction you find yourself thinking "this is like that patient I saw" → STOP. Restart from textbook. Never proceed with that thought pattern.

---

## 1. Vignette Domain Distribution

| Domain | Count | Source priority |
|---|---|---|
| Cardiovascular (STEMI, AD, HF) | 5 | Tintinalli's Ch.49-52 + Harrison's |
| ED Critical (sepsis, AKI, DKA, GI bleed) | 8 | Tintinalli's Ch.6-92 |
| OBGYN ED (ectopic, PE, PPH, ovarian torsion, septic abortion) | 5 | Williams Ch.19, 38, 41-42 |
| Neurology (stroke, status epilepticus, meningitis) | 4 | Tintinalli's Ch.165-176 |
| Taiwan Endemic (dengue, scrub typhus, leptospirosis) | 3 | Taiwan CDC + WHO factsheets |
| Pediatrics (Kawasaki, peds sepsis) | 3 | Nelson + Tintinalli's Ch.107-141 |
| Renal/Endocrine (CKD, hyperthyroid storm) | 2 | Harrison's Ch.226-229 |
| **Total** | **30** | |

---

## 2. Realism Tier Stratification

Each domain's vignettes split across 3 realism tiers (proportionally):

| Tier | Description | Source within textbook |
|---|---|---|
| **T-classical** (10) | Classic textbook presentation | "Typical Presentation" sections |
| **T-atypical** (10) | Atypical variants explicitly described in textbooks | "Atypical Presentations" / "Variants" sections |
| **T-messy** (10) | Real-world complexity (comorbidity, masked presentation, social factors) | **"Complications", "Pearls and Pitfalls", "Common Errors", "Missed Diagnoses" sections** — STILL textbook-source, NOT case reports |

**Codex correction**: T-messy was originally planned to use PMC case reports. NOW changed to use textbook complications/pitfalls sections to avoid PMC license + consent transitivity issues.

---

## 3. Construction Steps (each vignette)

### Step 1: Open textbook, navigate to disease chapter

Example: STEMI inferior wall

- Open Tintinalli's 9th ed., Ch. 49 "Acute Coronary Syndromes"
- Note chapter + page

### Step 2: Extract canonical features list

From textbook, extract:

- Demographics (typical age range, sex predominance)
- Onset pattern (sudden / gradual / minutes / hours)
- Symptom: chief complaint + associated symptoms
- Physical findings
- ECG findings
- Lab patterns
- Imaging findings
- Cautions / pearls / pitfalls

Write features list as:

```
disease: "STEMI inferior wall"
source: "Tintinalli's 9th ed. Ch.49 p.342-355"
features_canonical:
  demographics: "older adult (>50), M:F = 2:1"
  onset: "sudden, minutes"
  symptom: "substernal/epigastric pain, may radiate"
  associated: "diaphoresis, nausea"
  signs: "diaphoresis, possible bradycardia"
  ecg: "ST elevation in II, III, aVF; possible RV involvement (V4R)"
  labs: "troponin elevation"
  cautions: "RV infarct = preload-dependent, nitrate-sensitive"
  pitfalls: "may present with isolated nausea/epigastric pain in elderly/diabetic"
```

**Step 3: Pick realism tier and construct accordingly**

**T-classical**

Use canonical features as-is, assemble into vignette:

> "65 男, 突發胸痛 30 分鐘, 冒冷汗, BP 140/85, HR 75, ECG ST↑ II III aVF..."

**T-atypical**

Use textbook-described atypical variants:

> "72 女, 糖尿病, 主訴僅噁心 + 上腹脹 2 小時, 無胸痛, BP 130/80, HR 88, ECG ST↑ II III aVF (病人主訴只覺得"消化不良")..."

(這是 textbook 寫的: "elderly diabetic women may present with only epigastric discomfort or nausea, without chest pain")

**T-messy**

Use textbook-described complications / pitfalls / common errors:

> "68 男, 慢性類固醇 (RA), 突發胸痛 + 暈眩, BP 75/40, HR 110, JVP 抬高, 肺部清, ECG ST↑ II III aVF + V4R ST↑, 預定要開 nitro drip..."

(這是 textbook 寫的 pitfall: "RV infarct + nitrate = catastrophic preload drop; never miss V4R in inferior STEMI")

**Step 4: Inject Taiwan-specific context**

Choose subset:

- 健保轉診 pattern: "從 [地區] 醫院急診轉來"
- 中英台語 code-switching: "病人 says '心臟艱苦' (chest 艱苦, mixing Mandarin + Min-nan)"

- Taiwan endemic context (only for endemic vignettes): "近期 from 高雄 outdoor activity"

**Step 5: Assemble 5-stage cut-points**

Split vignette into 5 stages (see schema YAML for structure):

- Stage 1 (presentation): demographics + chief complaint + vitals
- Stage 2 (initial workup): + first labs + ECG/CXR
- Stage 3 (diagnosis): + key imaging/lab confirming dx
- Stage 4 (management): + confirmed dx, ask next steps
- Stage 5 (reasoning): + full context, ask reasoning explanation

**Step 6: Define ground truth trajectory**

From textbook (or guideline), what is the expected trajectory?

```
ground_truth_trajectory:
  24h_disposition: "ICU"
  24h_events: ["PCI", "IABP_consideration"]
  48h_complications: ["RV_failure", "AV_block"]
  final_disposition: "discharge_day_5-7"
  interventions_done: ["primary_PCI_RCA"]
  forbidden_interventions: ["nitrate_high_dose"]
```

(Trajectory based on textbook expected course, not specific patient)

**Step 7: Collaborating clinician review (clinical realism check, blinded)**

Collaborating clinician rates clinical realism 1-5:

- 5 = This presentation is plausible and I see similar patterns regularly
- 4 = This presentation is plausible
- 3 = Plausible but unusual
- 2 = Implausible — clinical inconsistency
- 1 = Implausible — internal contradiction

**Collaborating clinician blinded to**: tier label, source textbook, intended diagnosis (Collaborating clinician rates the case "cold")

**Threshold**: vignette must score ≥ 3.5 to enter pool; <3.5 → revise or discard

**Step 8: Provenance log**

For every vignette, complete one row in Provenance Ledger (see template):

- vignette_id, source textbook + page, tier, modifications, realism score, lock date

**Step 9: Lock**

Once realism check passed → vignette is **locked**. No modifications after model runs begin.

---

### 4. Modification Rules (for added clinical realism)

When constructing T-atypical and T-messy vignettes, you may modify canonical features within these bounds:

| Modification | Allowed range | Justification |
|---|---|---|
| Age | ±20 years from typical, must remain epidemiologically plausible | "Atypical age" but still plausible |
| Sex | Switch allowed if disease epidemiology permits | (e.g., not for PPH male) |
| Lab values | ±50% from typical, must remain clinically realistic | Add edge cases |
| Comorbidities (T-messy only) | Add 1-2 from textbook common comorbidity list | Add complexity |
| Social factors (T-messy only) | Add from public health data | Real-world context |

**Red lines (NEVER cross)**:

- Don't modify so much that disease epidemiology breaks (e.g., PPH in male)
- Don't add specific features that match a real patient case you've seen
- Don't borrow specific numbers from a remembered case

---

### 5. Lock Date Protocol

| Date | Event |
|---|---|
| W4-end | All 30 vignettes constructed, collaborating clinician review complete, schema-validated |
| W4-end | **VIGNETTE LOCK DATE** registered on OSF |
| W5+ | No vignette modifications under any circumstance |
| W5+ | Model inference begins on locked vignettes |

If a critical issue is discovered post-lock:

- Document the issue in OSF supplementary
- Do NOT silently modify
- Either drop the vignette + report N reduction, or accept as limitation

---

### 6. Anti-Overfit Protocol (for T-messy adversarial cases)

T-messy vignettes designed as "twists" risk overfitting to author's theory of LLM failure. To prevent:

| Risk | Mitigation |
| --- | --- |
| Author designs traps that LLMs are known to fail | Author and collaborating clinician blinded to model output during vignette construction |
| Adversarial subset is hand-tuned during testing | Lock vignettes BEFORE any model inference |
| Reviewer questions cherry-picking | Pre-register adversarial construction rules on OSF |

---

### 7. Documentation Required for Each Vignette

```
vignette_id: "TW-CV-T1-001"  # TW=Taiwan, CV=cardiovascular, T1=T-classical, 001=number
domain: "cardiovascular"
disease: "STEMI inferior wall"
realism_tier: "T-classical"

source:
  primary_textbook: "Tintinalli's Emergency Medicine 9th ed."
  chapter: "49 – Acute Coronary Syndromes"
  pages: "342-355"
  edition_year: 2020

modifications:
  - "Demographics: age 65, M (canonical: older adult, M-predominant)"
  - "Vitals: BP 140/85, HR 75 (canonical range)"
  - "Taiwan context: 健保轉診 from [地區醫院]"

modifications_within_bounds: true  # confirm with checklist

realism_check:
  reviewer: "黃 XX (合作臨床醫師, OBGYN attending)"
  blinded_to: ["tier", "source", "intended_diagnosis"]
  score: null  # filled after review
  decision: null  # accept | revise | reject
  reviewer_signature_date: null

lock_status:
  locked: false
  lock_date: null
  osf_registration_id: null

5_stage_structure:
  stage1_presentation: "..."
  stage2_initial_workup: "..."
  stage3_diagnosis: "..."
  stage4_management: "..."
  stage5_reasoning: "..."

ground_truth_trajectory:
  24h_disposition: "..."
  24h_events: []
```

```
  48h_complications: []
  final_disposition: "..."
  interventions_done: []
  forbidden_interventions: []

contamination_metadata:
  # filled after running detection script per model
  per_model_scores: {}
  tier_contamination: null  # T1 (clean) | T2 (modified) | T3 (suspected)
```

---

### 8. Build Order

| Order | Vignettes | Why |
|---|---|---|
| 1st batch | 5 cardiovascular T-classical | Most familiar, build pipeline |
| 2nd batch | 5 OBGYN (collaborating clinician domain) | Collaborating clinician review depth |
| 3rd batch | 8 ED critical | Bulk |
| 4th batch | 4 neuro | |
| 5th batch | 3 Taiwan endemic | Hardest (rare patterns) |
| 6th batch | 3 peds + 2 renal/endocrine | Final |

---

### 9. Quality Gates

Each vignette must pass ALL before locking:

- [ ] Source textbook + chapter + page recorded
- [ ] Realism tier confirmed (T-classical / T-atypical / T-messy)
- [ ] Modifications within allowed bounds (checked against §4)
- [ ] No personal patient memory used (Andrew self-affirmation)
- [ ] Collaborating clinician realism review ≥ 3.5 (blinded)
- [ ] Schema validator passes (YAML well-formed, all required fields)
- [ ] Provenance ledger row complete
- [ ] OSF preregistration listing includes this vignette_id
- [ ] Locked timestamp recorded

If any gate fails → revise or discard, do not include in 30.

---

### 10. End-of-Construction Checklist

Before vignette lock date:

- [ ] 30 vignettes constructed
- [ ] Domain distribution matches plan (5/8/5/4/3/3/2)
- [ ] Tier distribution matches plan (10/10/10)
- [ ] All collaborating clinician-reviewed
- [ ] All schema-validated
- [ ] Provenance ledger complete (30 rows)
- [ ] OSF preregistration submitted with vignette manifest
- [ ] Lock date recorded
- [ ] Vignettes archived in version control (git commit)

---

*來源檔：~/Desktop/local LLM VS Cloud LLM/paper-A-governance/04-vignette-construction-protocol.md*

# OSF Pre-Registration Template

> Submit at: https://osf.io/registries/osf/new
> Choose template: "Open-Ended Registration" or "Pre-Registration Template"
> Submit BEFORE running any model inference (W4 of timeline)

---

## Study Information

### Title

Failure Shape Transfer Across Cloud, Local, and Agentic Large Language Models in Sequential Clinical Reasoning: A Multi-Stage Evaluation on Taiwan-Specific Vignettes

### Authors

- Andrew Huang (NYCU College of Medicine, M6) — first author, corresponding
- [Collaborating clinician name] (Dept. OBGYN, Taipei VGH) — senior co-author, clinical reviewer
- [1 同學 name] (NYCU / PGY) — co-author, rater
- [Senior PI name + affiliation] — senior author

### Date of registration

[Date you submit OSF]

### Stage of registration

Pre-data collection (no model inference run yet)

---

## Hypotheses

### H1 (primary)

Cloud LLM differential-diagnosis (DDx) universal weakness identified by Rao et al. (JAMA Network Open 2026) **transfers to** local quantized open-weight models when evaluated under matched sequential clinical reasoning protocols.

**Directional**: Local quantized models will show **equal-or-greater** failure rates on DDx breadth, calibration, and harmful omission endpoints compared to cloud frontier models, with the gap **widening** under increased realism tier.

### H2 (secondary)

Agent scaffolding (AgentClinic harness) provides **modest accuracy gains** but introduces **new failure modes** (overconfident wrong recommendations, hallucinated tool outputs) compared to bare LLM inference.

**Directional**: Net change in harmful omission rate under agent scaffolding will be **non-zero** with **shifted error distribution** (decrease in some categories + increase in others), as predicted by Liu et al. (npj Digital Medicine 2026).

#### H3 (exploratory)

Performance degradation across realism tiers (T-classical → T-atypical → T-messy) is **steeper for cloud frontier models** than local quantized models, suggesting cloud models may overfit to canonical textbook presentations.

**Quantification**: Realism Sensitivity (slope of accuracy vs tier ordinal) will differ between cloud and local groups (Wilcoxon test, alpha = 0.05).

---

## Design

### Type

Cross-sectional comparative evaluation, factorial design

### Factors

| Factor | Levels | Type |
|---|---|---|
| Model family | Cloud (GPT-5, Claude 4.7) vs Local (Qwen3-235B-3bit, Llama-3.3-70B-8bit, Med42-70B, MedGemma-27b, gpt-oss-120b) | Between-vignette (within-vignette across models) |
| Scaffold | Bare LLM vs AgentClinic agent harness | Within-vignette |
| Vignette | 30 synthetic | Within-model |
| Stage | 5 (presentation, workup, diagnosis, management, reasoning) | Within-vignette |

### Conditions / Cells

7 models × 2 scaffolds × 30 vignettes × 5 stages = **2,100 condition cells**

With 3 repeated runs per cell (different seeds) = **6,300 model responses**

### Vignettes

- 30 synthetic clinical vignettes
- 100% textbook-derived (no patient data)
- 3 realism tiers × ~10 each
- 7 domains
- See: Vignette Construction Protocol v1
- See: Vignette Schema v1
- Vignette manifest (vignette_id list) attached as supplementary

### Models — exact versions to lock

| Model | Version | Quant | Inference framework |
|---|---|---|---|
| GPT-5 | API as of [date] | — | OpenAI API |
| Claude 4.7 | API as of [date] | — | Anthropic API |
| Qwen3-235B-A22B | mlx-community/Qwen3-235B-A22B-3bit | 3-bit MLX | mlx-lm |
| Llama-3.3-70B-Instruct | mlx-community/Llama-3.3-70B-Instruct-8bit | 8-bit MLX | mlx-lm |
| Med42-70B | sjgdr/Llama3-Med42-70B-mlx-6Bit | 6-bit MLX | mlx-lm |
| MedGemma-27b | mlx-community/MedGemma-27b-it-4bit | 4-bit MLX | mlx-lm |
| gpt-oss-120b | mlx-community/gpt-oss-120b-MXFP4 | MXFP4 | mlx-lm |

#### Inference settings — locked

- Temperature: 0.7 (default)
- Top_p: 0.95
- Max tokens per stage: 500
- System prompt: standardized across all models (see Methods)
- Random seed: 42, 123, 7 (3 runs)

#### Agent scaffold

- AgentClinic harness (Schmidgall et al. 2024, GitHub: agentclinic/agentclinic)
- Tools enabled: literature lookup, calculator, dose calculator
- Max iterations: 10

---

### Endpoints

#### Primary endpoints

1. **DDx breadth at Stage 1** (rubric-scored 0-5)
2. **Calibration ECE at Stages 1-3** (probability assessment)
3. **Harmful omission rate** (binary, see rubric)
4. **Trajectory prediction accuracy** (objective vs ground truth, multi-class + multi-label)

#### Secondary endpoints

1. Sequential reasoning quality at Stage 5 (rubric-scored 0-5)
2. Latency to first token + total latency
3. Token usage
4. Realism Sensitivity Index (slope of accuracy across realism tiers)

#### Endpoint locking

All endpoints + rubrics locked at this preregistration. Post-hoc endpoints will be reported as exploratory only.

---

## Sampling Plan

### Sample size

- 30 vignettes (locked)
- 7 models × 2 scaffolds = 14 conditions per vignette
- 3 repeated runs per condition
- Total responses: 30 × 14 × 3 = 1,260 responses across 5 stages = 6,300 stage-level evaluations

### Power analysis

For primary H1 (cloud vs local DDx breadth difference):

- Anticipated effect size: Cohen's d = 0.5 (medium)
- Alpha: 0.05
- Power: 0.80
- Required N per group: ~64 paired observations
- Available N: 30 vignettes × 5 stages = 150 paired observations per model pair → adequately powered for primary

For H3 (Realism Sensitivity slope difference):

- 30 vignettes / 3 tiers = 10 vignettes per tier
- Underpowered for tier-domain interaction; reported as exploratory

---

## Analysis Plan

### Primary analysis (H1)

- Mixed-effects logistic regression
- Outcome: DDx correctness (binary at Stage 1, top-3 includes ground truth)
- Fixed effects: model_family (cloud vs local), realism_tier, scaffold
- Random effects: vignette_id, run_seed
- Pre-registered contrasts:
- Cloud vs Local: main effect
- Cloud × tier vs Local × tier: interaction (test of H3)
- Multiple comparison correction: Bonferroni for primary endpoints (4 endpoints → α/4)

### Secondary analyses

- Calibration ECE: per model, per tier, per scaffold

- Harmful omission rate: chi-square + 95% CI
- Trajectory prediction accuracy: F1 (multi-label complications) + accuracy (4-class disposition)
- Latency: descriptive + Mann-Whitney for cloud vs local

**Inter-rater reliability**

- LLM-judge consensus vs human consensus: ICC (3,k)
- Threshold: ICC ≥ 0.7 for primary endpoint validation
- If ICC < 0.7: report and discuss; do not change rubric post-hoc

**Realism Sensitivity Index (RSI) — exploratory**

- Per-model slope of accuracy (T-classical → T-atypical → T-messy) using ordinal tier
- Test for difference in slope: Wilcoxon signed-rank
- Reported as descriptive only unless replicated in Phase 2

---

## Inclusion / Exclusion

**Vignettes**

- Include: All 30 vignettes that pass realism check (collaborating clinician score ≥ 3.5) + schema validation
- Exclude: Any vignette failing realism check or showing post-hoc identified clinical inconsistency

**Model responses**

- Include: All responses producing parseable output
- Exclude: Responses with parse failure (record count + reason)
- Tolerance: <5% parse failure per model (else investigate before proceeding)

---

## Pilot Data

A single pilot vignette (STEMI inferior wall) was tested on Med42-70B-6bit prior to preregistration. Result: 9.0 tok/s, 57.5 GB peak RAM, 234 tokens output, response correctness 4/5 stages (full record archived in MODEL_BENCHMARK_RESULTS.md).

This pilot data will be **excluded** from final analysis to prevent overlap with confirmatory study.

---

## Other

**Data sharing**

- All vignettes (after lock) released on OSF + GitHub at submission
- All model responses (raw text) released anonymized

- Code (inference, scoring, analysis) released on GitHub
- License: CC BY 4.0 for vignettes; MIT for code

**Conflicts of interest**

- First author and senior co-author are siblings (disclosed)
- No financial relationships with any LLM developer
- No funding support beyond personal hardware (declared)

**Ethics approval**

- NYCU IRB determination letter: [pending / received]
- Taipei VGH IRB determination letter: [pending / received] (for collaborating clinician's role)
- Phase 2 prospective study planned: separate full IRB review (in preparation)

---

### Anticipated Deviations from Pre-registration

If we deviate from this preregistration during analysis, we will:

1. Pre-specify in this document any contingent analyses (e.g., "If ICC < 0.7, we will conduct sensitivity analysis excluding low-IRR responses")
2. Clearly label all post-hoc analyses as exploratory in the manuscript
3. Maintain a deviation log in the OSF project

---

### Pre-registered Files (attached at submission)

- [ ] This preregistration document
- [ ] Vignette Construction Protocol v1
- [ ] Vignette Schema v1
- [ ] Reviewer Charter (signed by collaborating clinician)
- [ ] Vignette manifest (30 vignette_ids — populated at lock date)
- [ ] Scoring Rubric v1
- [ ] Inference Pipeline code (frozen commit sha)

---

### OSF Submission Checklist

- [ ] Login to OSF
- [ ] Create new project: "Failure Shape Transfer Across Cloud, Local, and Agentic LLMs"
- [ ] Upload 7 governance documents
- [ ] Submit registration (cannot edit after — review carefully)
- [ ] Receive OSF registration ID (osf.io/XXXXX)

- [ ] Update HANDOFF doc with OSF ID
- [ ] Update Vignette Provenance Ledger with OSF ID per vignette

---

*來源檔：~/Desktop/local LLM VS Cloud LLM/paper-A-governance/07-OSF-preregistration-template.md*