REPRODUCIBILITY DOCUMENTATION — EVERY CITED NUMBER, INDEPENDENTLY VERIFIABLE

Dementia Care Decision — Raw Metadata

Every number that can be cited about the run, with how it was recorded and how to verify it. For case-study fact-checking, journalist due diligence, clinical-review verification, and any defensible-claim build.

1. The run

A structured decision-support analysis for a family facing an early-stage dementia diagnosis, produced by Umma in a single autonomous run. The final artifact is a 5-section structured report — status: ready.

2. Time

THING	VALUE
Run started	2026-06-03 12:15:47 UTC
Run completed	2026-06-03 13:49:25 UTC
Total duration	1 hour 33 minutes 38 seconds (5,618.0 s)
Artifact materialized at	2026-06-03 13:49:25 UTC

3. Decomposition

THING	VALUE
Root thoughts	1
Child thoughts	20
Total thoughts in tree	21
Completed children	20 (100%)
Failed children	0
Distinct decomposition roles	5
Tree depth	1 (root + one level of children — flat by design for a parallel fan-out decision-support analysis)
Reactive stages	2 (Stage 1: 12 children; Stage 2: 8 children; a third stage was checked and declined — coverage was already sufficient)

3.1 Children by role

ROLE	COUNT
design	9
research	7
evaluate	2
synthesize	1
compare	1

4. Tool invocations

THING	VALUE
Total invocations	190
Successful	128 (67%)
Failed / exploratory	62 (33% — mostly exploratory REPL checks)
Distinct tool names	10

4.1 All 10 tools by call count

TOOL	CALLS	SUCCESSFUL	AVG DUR (MS)
`execute_code`	89	84	303
`repl_execute`	54	0	305
`search_web`	20	20	5,118
`find_existing_code_project`	14	14	395
`deep_research`	4	4	124,996
`list_code_projects`	3	3	18
`create_code_project`	2	2	480
`batch_write_files`	2	0	70
`create_files_bulk`	1	1	39
`create_code_file`	1	0	62

Caveats: the `repl_execute` success=0 pattern is a tracking convention, not a failure signal — REPL invocations are exploratory by design. The `batch_write_files` failures hit a known tool bug; children routed around it via `create_files_bulk` or individual file-creation calls.

5. Cognitive layer outputs

LAYER	COUNT
L1 interpretations	54
L2 syntheses	4
Provenance links	132
Provenance density per child	6.6

5.1 Interpretations by module

MODULE	INTERPRETATIONS	AVG CONFIDENCE
umma	20	1.00
design	9	1.00
research	7	1.00
evaluate	2	1.00
compare	1	1.00
synthesize	1	1.00
research_followup	14	0.70

Why this matters for credibility: the lower confidence on `research_followup` (0.70) is calibrated, honest reporting. The gap-fill children honestly recorded moderate uncertainty about follow-up evidence (dementia-subtype data, long-term-care trigger specifics, finance-range estimates) rather than uniformly asserting 1.00 across every module.

5.2 Headline synthesis

FIELD	VALUE
Confidence overall	1.000
Convergences	19
Divergences	9
Unresolved disagreements	0

6. Artifact composition

THING	VALUE
Artifact kind	report
Final sections	5
Section types	heading (1) + callout (1) + table (1) + bullets (1) + numbered_list (1)
Total file size on disk	10,378 bytes (~10 KB)
Master table dimensions	3 rows × 5 cols (option comparison)

6.1 Section inventory

ORDINAL	TYPE	TITLE (TRUNCATED)	ROWS × COLS
0	heading	(title block)	0×0
1	callout	“Immediate Month-1 Critical Path: Diagnostic & Legal Blocker…”	0×0
2	table	“Operational and Financial Comparison of Care Options”	3×5
3	bullets	(findings bullets)	0×0
4	numbered_list	(action sequence)	0×0

7. Verifiability

Every figure above is recorded in the run’s structured provenance — the decomposition tree, the per-thought tool log, the typed interpretations with their confidence scores, the synthesis record (with its convergences, divergences, and unresolved-disagreement counts), and the 132 provenance links tracing every claim back to its source. Each count can be independently re-derived from that provenance. Query-level access for fact-checking is available to press, clinical reviewers, and partners on request.

8. What this run did and did not do

For any case-study or clinical-review packet, it’s worth being precise about both sides.

8.1 What the run did

- Performed **190 tool invocations** across 10 distinct tools — web search against published care-cost research, deep-research traversals into dementia-trajectory literature and Medicaid-lookback rules, sandboxed code execution for the financial model (computing the $1.39M vs $2.81–3.46M trajectories across the four options), and project lookups for cross-strand synthesis
- Spawned **20 specialist child thoughts** across 5 roles (design, research, evaluate, synthesize, compare), running in 7 coordinated waves rather than a flat parallel pile
- Produced **54 typed interpretations** with calibrated confidence — including the follow-up research module honestly reporting average confidence at 0.70 rather than asserting 1.00 across all modules
- Reconciled **19 cross-strand convergences and 9 cross-strand divergences** through adversarial synthesis, closing with zero unresolved disagreements at confidence 1.00
- Recorded **132 provenance links**, making every claim in the deliverable traceable back to its source
- Built a unified **four-option scoring framework** (Option A: CCRC / co-located AL→MC; Option B: in-home; Option C: move-in; Option D: Staged Hybrid) with an 8-dimension comparative matrix
- Constructed an **emergent fourth option** (Staged Hybrid) that the user’s original framing did not name, by reconstructing the option space from first principles

- Computed a **10-year financial trajectory** across all four options using placeholder cost ranges from public care-cost research
- Identified a **clinical contradiction** in the original diagnosis worth bringing to a doctor (early-stage dementia with physical decline suggests non-Alzheimer's pathology, including potentially-reversible NPH)
- Flagged the **Medicaid 5-year lookback** as already running, with informal caregiver compensation as penalty-eligible
- Designed **three protocols**: preference elicitation (neutral third-party facilitation, explicit exclusion of both siblings), a facilitated family-meeting structure, and a sibling-conflict diagnostic distinguishing values-conflict from process-conflict
- Produced a **5-section structured report** (heading + critical-path callout + comparison table + key findings + numbered action list) for the family to use as the spine of their own deliberation

These are real cognitive operations, all recorded in the run's provenance.

8.2 What the run did NOT do

Categories of work explicitly out of scope, which the artifact does not claim to substitute for:

1. **Did not give medical advice.** The artifact is decision-support architecture, not a clinical recommendation. Subtype identification, capacity assessment, and treatment selection are referred to qualified geriatric-medicine specialists. The clinical-contradiction insight is a question to bring to a doctor, not a diagnosis.
2. **Did not assess the mother's actual preferences** — only designed the protocol for eliciting them via neutral third-party facilitation, which must be executed by a qualified facilitator before the family-meeting design fires.
3. **Did not model the specific financial situation** — used placeholder ranges from public care-cost research. The model is structural ($1.39M vs $2.81–3.46M over 10 years); actual numbers require real financial data (LTC coverage, asset structure, income, applicable Medicaid spend-down thresholds).
4. **Did not assess the sibling's actual constraints** — only designed the protocol for understanding them. The framework awaits real engagement with the sibling.
5. **Did not perform legal review** — only flagged the legal groundwork as time-sensitive. Instrument execution (POA, advance directives, will review, caregiver agreement) requires an elder-law attorney in the mother's state.
6. **Did not perform a clinical capacity assessment** — only referenced validated instruments (MMSE, MoCA, CDR, FAST). Actual assessment requires a qualified clinician.
7. **Did not perform geographic / market analysis** — availability, waitlists, CCRC financial-pass rates, and agency reputability vary by region. The recommendation is structural; localization was hard-blocked at execution because no geography was selected.
8. **Did not consult social-services, hospice, or palliative providers directly** — only flagged them as resources to engage during the Month-1 critical path.

8.3 What this means for the press claim

The run is what it claims to be: a structured decision-support framework for a family to use as the spine of their own deliberation, with explicit acknowledgment of which steps require professional consultation. It is not a substitute for a geriatric care manager, an elder-law attorney, a geriatrician, or a social worker. It is the upstream structuring that makes those professionals' work more effective — giving the family the right shape of conversation, the right questions to bring, and the right operational arithmetic before they sit down with each specialist. The combination of explicit "what was done" and explicit "what was not done" — both rooted in the recorded trail — is part of what distinguishes this case study from "we asked a chatbot what mom should do" content. Honesty about scope is load-bearing here, not optional.
