Appendix A: pr4xis HCBS/EVV Compliance Navigator — Track 2 Supporting Material

This appendix sits outside the 15-page Project Narrative limit (ACL Application Outline, Appendices section). It supports the narrative rather than extending it: where each sub-criterion is addressed (A.1); this solution's corpus and current behavior (A.2); the strata it works across and what each requires (A.3); how it is measured and planned (A.4–A.7); and the instruments specified for Phase 2 execution (A.8–A.9).

What follows describes software that is built and running: a browser-resident demonstrator needing no server, no account and no per-query cost; answers carrying inline federal statutory and regulatory citations; abstention rather than guessing, and no generative stage that could fabricate a citation in the first place; the same question returning the same answer on every run, auditably; and a standing regression suite enforced in CI, with failing tests written before fixes. Two conventions govern the figures below. A measured figure is stated in the present tense with the file or command that reproduces it; an expected one is labelled a projection with its assumptions written out. Each measurement carries the exact criterion it tests, stated *at* the measurement. Every present-tense number names its instrument, and *architectural* is reserved for a property that follows from a structural choice.

A.1 Judging-criteria crosswalk

"§" numbers are Project Narrative sections; "A.n" refers to this appendix.

Criterion / sub-criterion	Where	What the reviewer finds
1 Understanding of the Need	§1; A.3	The Cures Act mandate and the 2024 Ensuring Access rule (89 Fed. Reg. 40542) cited provision by provision; a stratification naming what each class of question requires
1 Responsiveness (viability)	§1; A.3	Track 2′s use cases met literally, scope boundaries named; a falsifiable, publicly gated per-stratum commitment in a compliance staffer’s terms
1 Impact	§4 P3; A.6	Net-Time-Saved projection, arithmetic and assumptions visible, its dependence on A.3 stated
2 Caregiver Input (workers, compliance staff)	§1; A.2, A.8	Institutionally-mediated input, sourced: 131 sources across this track’s slice, state EVV FAQs, a vendor knowledge base, Tseng (2020); recruited sessions open Phase 2, with the protocol that obtains them
2 Co-Implementation	§2; A.8	Recruited compliance-staff and DSP sessions as the *first* Phase 2 milestone — protocol attached, executable as written
3 Deployment Readiness · Timeline	§2; A.7	Zero-server WebAssembly architecture, channel integration named as pilot work; a quarter-dated plan gated on named tests, with a deliberately short first half
3 Metrics (credibility)	§2; A.2, A.4, A.7	CI-enforced infrastructure with a record of catching its own operators — including a figure of mine replaced by the bound its instrument reproduces (A.4) — and a Phase 2 table stating evaluation *design*, not only instruments
3 Evaluation and Adaptation	§2; A.7	Ratchet plus per-source floors, enforced today; a decision rule stated in advance for a falsified premise
4 Error prevention	§3; A.2, A.5	Typed outcomes decided in code; every `Answered` stands on a loaded, cited edge, at corpus scale (`tests/chat_capability.rs`); that
[illegible]	[illegible]	[illegible]

A.2 What this track measures against, and where it stands today

Every figure here is a property of committed files, re-derivable without running the reasoner. The corpus of record is **4,219 US-sourced questions**, of which **2,147 are this track's audience slice** (`track2_workforce` plus questions tagged for both tracks) and **1,663 are track-exclusive**. That slice spans **131 distinct sources**; the 1,663 track-exclusive questions are drawn from 78 of them. The purpose-built **HCBS Workforce and Compliance Lexicon carries 279 synsets**, and **every one of the 279 definitions names its source inline**: 185 cite the U.S. Code, C.F.R., Federal Register, Statutes at Large or a Public Law by number, and the other 94 name the federal agency guidance, state code, state Medicaid FAQ, vendor documentation, uniform act, clinical literature or policy-research report they were authored from. The split moves with the matching rule, so the rule ships as the command below — the same rule §1 quotes — rather than as a characterization.

```
F=crates/praxis-corpus-tests/tests/fixtures/caregiver_question_corpus.json
L=crates/domains/data/care/hcbs_compliance_lexicon.xml
T='select(.track=="track2_workforce" or .track=="both")'
jq 'length' $F ; jq "[.[]|$T]|length" $F ; jq "[.[]|$T|.source]|unique|
length" $F
                                        # 4,219 ; 2,147 ; 131 (78 sources
without `or …"both"`)
grep -c '<Synset ' $L                   # 279
grep -o '<Definition>[^<]*</Definition>' $L | \
  grep -cE '[0-9]+ (USC|U\.S\.C\.|CFR|C\.F\.R\.|Stat\.)|Fed\. Reg\.|[0-9]+ FR
[0-9]|Pub\. L\.'   # 185
```

The corpus is US-jurisdiction only, and that was a change. 398 rows were removed on 2026-07-24 — 395 from twelve Alzheimer's Society (UK) subforums, 3 from Dementia Australia — because the authorities this engine answers US questions from are US federal — statute, regulation and federal program guidance — so a question decided under another jurisdiction's law has a correct answer this loaded set does not hold. The lexicons carry a small number of comparative non-US entries, each cited to its own source, which is exactly why the corpus was scoped to US rows rather than left to measure a jurisdiction mismatch as a capability gap. The effect is recorded on both axes: in aggregate it moved the corpus-wide picture barely at all, and the two Alzheimer's Society source slices that had been standing at a committed floor of 0 left the floors table entirely along with their rows, with their prior classifications published so the change can be audited for self-interest. What remains in that table is fifteen slices, every one of them a US source measuring a real capability.

A reviewer running `jq` will also meet a residual field: rows still carry a `praxisCapability: "out_of_scope_abstain"` string whose *scoring* use was abolished (A.4). The scoring path ignores it; it remains as inert provenance, and every row is now scored as owed a grounded answer.

Current behavior is measured corpus-wide on every commit, by `cargo test --manifest-path crates/praxis-corpus-tests/Cargo.toml --release --test caregiver_capability_ratchet`. Every question lands in exactly one typed outcome class: `Green` — answered from a cited edge, naming the asked-about term; `MissingTerm` — the key term never becomes an addressable concept; `UnparsedKnownTerm` — the term is loaded, but the question's structure never reaches it; `PossibleMisroute` — answered, but not about the term that was asked. Each class sits under a committed ceiling CI may only let fall (A.4). Track-scoped behavior comes from a dedicated probe (`probe_corpus_pass_rate_by_track`, `crates/praxis-corpus-tests/tests/scratch_probe.rs`); every Track 2-specific result quoted here is re-measured since the rescoping.

What "Green" counts, exactly. A question is Green when the turn produced a `Conditional`/`RuleResolved` outcome — a cited rule plus the fact it still needs, which for a rule-governed compliance

question is the correct answer *shape* — or an `Answered` outcome whose response contains the recorded key term as a case-folded substring (`classify_case`, `crates/praxis-corpus-tests/src/caregiver.rs`). Substring containment is a **routing** check: with the property below it establishes that a loaded edge stood behind the answer and that the answer is about the term asked about; whether that edge carried its own citation is the separate fact the next paragraph states. Sense selection and governing provision are what a blind two-rater audit of the whole Green set grades, the first Phase 2 measurement (A.7); every claim here rests on the routing check alone.

The governing property. The engine answers where a loaded authority governs and declines elsewhere, and the property behind every answer it gives is architectural rather than tuned. *Holding across all 4,219:* every `Answered` outcome stands on a loaded `Definition` or `Subsumption` edge the engine already holds — the never-false-affirm property, exercised at corpus scale over the full loaded graph in `chat_capability.rs`. Citation is a property of the edge's source: a definition from either purpose-built lexicon declares its cited authority, while a surface outside both answers from a general-vocabulary sense, uncited and tracked as a named defect. That property is proved at corpus scale and registered as a witness-scale sentinel axiom, which the ontology gate sweeps every build. Its scope is affirmations; A.5 maps that boundary. The class that carries a confidently wrong compliance answer is `PossibleMisroute`, held under a committed ceiling that may only move down. For a compliance officer the trade is explicit: a decline rather than a guess wherever the authority is not loaded — naming the unresolved surface where the parse identifies one — and cited authority behind every answer that does come.

A.3 The strata, and what each one requires

Naming each class precisely is the substance of my claim to understand the need, because each implies a different kind of work: `MissingTerm` implies term construction and then authority acquisition (below); `UnparsedKnownTerm` implies grammar and pragmatics coverage — interrogative shapes, ellipsis, embedded questions; `PossibleMisroute` is the wrong-answer class and carries the tightest ceiling.

One structural cause dominates, and it is compound concept identity rather than vocabulary content. Two passes support that, each describing its own population.

Freshly derived from the corpus of record: **3,751 of 4,219 questions (88.9%) have a multi-word key term**, and **1,969 of this track's 2,147 (91.7%)** — HCBS and EVV vocabulary is compound-heavy even by caregiving standards. Re-derive with `jq '[.[]|.keyTerm]|map(select(test(" ")))|length'`, and again under the track filter above.

Measured on the unresolved-term population, dated to its own snapshot: a pass over those rows **as they stood before the 398 non-US rows were removed** found that almost all of their key terms are multi-word compounds, that in the large majority *every constituent word is already loaded*, and that the remaining tail is small — vendor product names, state program acronyms and clinical terms, whose constituents come from outside federal statute. That pass stands for that population on that date and is quoted for it alone. The engineering consequence is the point: for most unresolved terms the vocabulary is already present, so the work is concept identity, not content acquisition.

The mechanism is traced in the code. The grammar *already parses* these compounds — `svo::nominal_modifier_noun` exists, the out-of-vocabulary path offers the N/N reading, the semantics layer mints a right-headed composite — and retrieval is keyed on an exact pre-loaded surface form, so the parsed compound needs a concept identifier of its own before a loaded ontology module is consulted. The engine holds "EVV" and "training" as separate concepts today; "EVV training" becomes addressable once the composite carries an identifier. Hand-authoring entries reaches the small tail, which is why Phase 2 (§2) leads instead with minting a concept identity for a compound the grammar has already parsed, subordinate to its head.

Stated as a target a reviewer can judge viability by. The dominant stratum is one structural cause rather than an open-ended content backlog, so the Phase 2 target is committed **per stratum**. The end-of-Phase-2-

early gate is falsifiable and public: compounds the grammar already parses must become addressable concepts wherever their constituents are loaded, ratcheting that stratum's committed ceiling down with no other class rising, published per cycle (§2). In a compliance staffer's terms that is the difference between "EVV training" returning silence and returning the governing provision with the state-specific fact it still needs named — one cited answer, or one cited rule and a named gap. How often a constructible compound yields a *correct, cited* answer once the concept exists is what each cycle measures and publishes, rather than what this appendix predicts. A.7 commits the measurement.

A second stratum turns on authority availability rather than engineering, and this track meets it most often. What is committed and re-derivable is a bound: the engine loads **227 MB of U.S. Code**, and a key term occurs anywhere in it for only about 30% of corpus questions (the ratchet header's own figure, taken before the 398 non-US rows were removed) — occurrence itself being a generous upper bound on "a governing definition exists" rather than evidence of one. **60.5% of the corpus came from EVV vendors, state programs and peer forums**, which publish documents rather than uniform machine-readable authority. Both observations are the ratchet header's (`caregiver_capability_ratchet.rs`), taken before the rescoping and quoted for that population.

Underneath that bound sit three distinct authority conditions, each observed: a term coined in practice rather than defined in an authority (vendor product names, colloquial workforce terms); a definition published under a licence that reserves redistribution; and state administrative code published as documents behind no federated interface. Separating the three is a per-term legal-status review, and the passes run to date are lexical lookups, so the proportion each condition holds is what `probe_missing_term_authority_strata` will measure — a scheduled probe over the `MissingTerm` rows, with a written coding rule and a second coder on a sample (A.7). The third condition is the structural one for HCBS and EVV: the Cures Act mandate is federal, but its operative definitions — which services are EVV-covered, what counts as a reportable critical incident, which visit-record corrections are permitted — live in fifty state administrative codes, provider manuals and vendor knowledge bases published as documents rather than as data. Scope follows the authority: this tool answers where a loaded authority governs, and names the term it could not ground elsewhere.

A.4 The measurement infrastructure, and its record of catching its own operators

The measurement plan here is committed, CI-enforced infrastructure rather than a proposal, and its record includes what it caught in its own operators' work, each correction taken at the cost of the more comfortable number.

1. **Monotonic class ratchet** (`crates/praxis-corpus-tests/tests/caregiver_capability_ratchet.rs`). Every outcome class carries a committed ceiling, and any commit pushing a class above its ceiling fails; a ceiling rises only deliberately, in the same commit, with a written justification naming what got worse and why the trade was accepted. Two properties fix its scope: ceilings are corpus-wide rather than per-track, and they bound classes rather than individual questions.
2. **Per-source bias floors** (`caregiver_bias_floors_never_regress_per_source`, same file). Every source string contributing fifty or more questions carries a committed Green floor — fifteen slices, ten of them this track's own (seven state EVV FAQ documents, one EVV vendor's state EVV FAQ, plus the New York CDPAP and Washington Consumer Direct self-direction FAQs) — each floor ratcheting up in the same commit and never down. Three of the ten stand at a committed floor of 0, which records a measured zero rather than an enforced minimum. **The instrument has a record of catching its own operators:** slices form from the fixture's raw `source` field, so a publisher harvested with per-thread URLs spreads below the fifty-row threshold and forms no slice. agingcare.com is the largest such population, spread across hundreds of distinct source strings; little of it sits in this track's slice, and the fix belongs to the shared instrument. Publisher-level grouping is the committed Phase 2 gate change (A.7).

3. **Per-question snapshots.** Both corpora grade each question against a committed label rather than an aggregate, so a lateral shift — one question flipping unsafe while another flips safe — cannot hide behind a stable total.
4. **Citation and constitution gates** (`citations.toml`, `crates/domains/tests/citation_audit.rs`). Every registry entry carries a non-empty `verified_by` and every citation slug in code must resolve to one; unverified citations fail CI, no allow-list, no warn-only mode. Every domain test declares the capability claim it backs, so a test asserting nothing traceable fails the build. And every lexicon edit re-runs the full lock and registry lifecycle, so the artifact a reviewer downloads is provably the one the tests ran against.

The instrument caught its own scoring rule. Until 2026-07-24 the classifier consumed an `expects_answer` flag and scored *declining* as a pass; most of the corpus carried an `out_of_scope_abstain` tag that had entered as a bare JSON string consumed by a string comparison — the pattern this codebase forbids everywhere else. It measured fit-to-tool rather than fitness for the user. Abolishing it made every reported result strictly harsher on me, and every performance claim was rewritten under the stricter rule. A measurement plan revises in whichever direction its instrument points.

An adversarial citation audit of this track's lexicon caught a truncated verification pass, and corrected the pincites behind it. A first pass reported success on an extraction truncated before its first item, leaving the entire 42 U.S.C./42 C.F.R. family — the core of this track's authority — outside its scope. The follow-up replaced agent-estimated coverage with a schema-forced, code-computed batch split, and coverage reconciled exactly: 223 citations extracted, 223 verdicts returned. Every pincite the audit found wrong was corrected and others tightened (commits `8bb0e569`, `2723add0`) — representatively, 42 C.F.R. § 431.220(a)(3) is the PASRR hearing right, not the transfer/discharge appeal right, which is (a)(2). The record here is the audit's findings acted on, at the scope the audit covered.

The re-derivation rule ran against a number of my own, and a re-derivable bound took its place. A proportion once stood on A.3′s authority stratum; the rule that every present-tense number names the command reproducing it applies first to the numbers that flatter me, and the pass behind that one is a lexical lookup rather than a licence classification. What A.3 carries now is the bound the ratchet header does reproduce, plus the named probe that measures the rest (A.7). Enforcing the rule in the direction that costs us the more comfortable number is what makes it an instrument.

A.5 Safety evaluation: the method, and the boundaries it maps

The adversarial corpus (`crates/praxis-corpus-tests/src/adversarial.rs`) is authored, not harvested — every question deliberately fabricated, in four categories each grounded in the question-answering honesty literature. **`fabricated_citation`**, a well-formed statute, C.F.R. or Federal Register citation under a *real* title that does not itself exist, and **`fabricated_term`**, a plausible program or acronym name defined nowhere in the loaded lexicons (both: Lin, Hilton & Evans 2022, "TruthfulQA", ACL 2022). **`false_presupposition`**, a real loaded term asked about through a question whose premise misstates a fact about it (Kim, Pavlick, Karagol Ayan & Ramachandran, "Which Linguist Invented the Lightbulb? Presupposition Verification for Question-Answering", ACL-IJCNLP 2021, `2021.acl-long.304`) — uniquely here the key term *is* expected to resolve. **`domain_mimicry`**, real domain words recombined into a compound corresponding to no loaded concept (Rajpurkar, Jia & Liang 2018, "Know What You Don't Know", ACL 2018).

The graded-safe response is one typed outcome — the closed-world refusal (Reiter 1978): `Abstained`, never `Answered` or `Conditional`. Grading is per question against a committed snapshot, for the lateral-shift reason in A.4.

What this application puts forward is the corpus design and the grading rule, both readable in the cited file. As of 25 July 2026 the repair that snapshot once waited on has landed, so the suite-wide result is a current measurement: all 160 adversarial questions are classified `Safe` in the committed per-question

snapshot and the suite runs green at 162 tests. Each question is gated individually, so any change in outcome fails the build — a lateral shift between unsafe outcomes included. A.7 re-publishes it per question, quarterly, whichever way it falls. The four categories are domain-general and the suite grades corpus-wide; a Track 2-tagged subset — fabricated EVV vendor names and CMS rule numbers, false-presupposition questions about compliance deadlines and 80/20 thresholds — is a first-half milestone (§2, A.7).

Where the never-false-affirm property ends, and what takes over. A realized *denial* comes from the **absence** of a subsumption edge between two loaded concepts — closed-world negation (Reiter 1978). My own Data Output Log records one: "Is Hospice included in EVV?" is answered "No, a hospice is not an evv." Structurally that is the closed-world reading of a non-edge rather than a fabrication; for a compliance audience a denial carries the audit consequence of an affirmation, so it belongs under the same citation standard. Two first-half remedies follow (A.7): restrict a realized denial to cases where a cited exclusion or disjointness edge is loaded, abstaining otherwise; and add closed-world negation as a **fifth adversarial category** — questions whose correct behavior is abstention because no authority is loaded either way.

Two output-quality findings in the same log, each with its gate. *Sense selection:* generic dictionary senses can surface beside a cited lexicon definition — the log's "What is respite?" entry lists general-vocabulary senses alongside the cited one; the gate is composition order, a cited sense taking precedence for that surface, regression-tested per question. *Realization agreement:* the is-a path emits "Yes. a home health care services is an evv."; the gate is interrogative-typed realization, carrying determiner and number through a plural compound and answering a *which*-question in a *which*-shape. Both sit outside the safety grading rule above and inside what a compliance officer sees, which is why each carries a gate.

A.6 Net-Time-Saved: the projection and its full arithmetic

A projection of the steady state, pricing the *direct-answer* case, with its arithmetic and both of its assumptions on the page. Re-running it against Phase 2′s *measured* inputs is the evaluation (A.7).

Scenario	Manual lookup time (assumption)	Questions/week/ agency (assumption)	Arithmetic	Weekly hours
Low	15 minutes	5	15 × 5 ÷ 60	**1.25 hr**
Mid	30 minutes	10	30 × 10 ÷ 60	**5.00 hr**
High	45 minutes	15	45 × 15 ÷ 60	**11.25 hr**

Both inputs are assumptions: the lookup-time range bounds manually locating *and correctly interpreting* one regulatory answer; the questions-per-week range bounds how many research-requiring questions an agency's compliance function fields weekly. A.8′s timed sessions replace both with measurements, benchmarked against CMS's own regulatory burden estimates for the 2024 Ensuring Access final rule (89 Fed. Reg. 40542). Two effects are excluded conservatively: the better-aimed escalation after a decline that names its own unknown, and — entirely — the avoided cost of acting on a *wrong* compliance answer, an FMAP consequence or failed audit rather than an inconvenience. The projection is stated per agency, the unit for which a denominator is citable.

A.7 Phase 2 measurement plan

Each entry names the question, the instrument, today's baseline, and the artifact that records the result. **H1** is Phase 2 early (Q3–Q4 2026); **H2** is Phase 2 late (2027).

- **Does concept identity for parsed compounds convert missing terms into cited answers, and does capability improve evenly across states?** Ratchet re-run per change; floors re-measured each cycle, and regrouped by publisher rather than raw source string in H2 (§2′s mid-phase list). *Baseline:* A.4′s committed ceilings and per-source floors, with publisher-level grouping arriving in H2. *Gate:* ceilings

ratcheted down in the same commit, no other class rising — continuous through H1; regrouped floors committed in H2.

- **Are the answers the engine already gives substantively right?** Blind two-rater audit of the full Green set against the cited authority. *Baseline:* the standing routing check (A.2), which this audit extends to sense and provision. *Gate:* correctness rate and rater agreement published whichever way they fall, then a committed floor (H1).
- **How long does a compliance question take today, what does the tool save in workload, and do users act correctly on a decline?** Timed think-aloud with observed escalation and post-task interview (A.8), plus NASA-TLX (Hart & Staveland 1988), SUS (Brooke 1996) and the Copenhagen Burnout Inventory (Kristensen et al. 2005); design below. *Baseline:* A.6′s assumption ranges, which these sessions replace with measurements. *Gate:* session and usability report (H1).
- **Does it hold up under adversarial *compliance* input, including where no authority is loaded either way?** A Track 2-tagged subset on A.5′s four categories plus the closed-world-negation fifth. *Baseline:* A.5′s four domain-general categories, graded corpus-wide. *Gate:* per-question snapshot; denial restricted to cited exclusions (H1).
- **How much of the missing-term population is foreclosed by authority availability, and for which of A.3′s three reasons?** `probe_missing_term_authority_strata` over the `MissingTerm` rows: a written coding rule per term, second-coded on a 10% sample. *Baseline:* A.3′s lexical bound, which this probe resolves into a partition. *Gate:* committed probe, published coding rule and inter-coder agreement, after which a proportion enters either document (H2).
- **Can a worker who does not read English comfortably use it?** Plain-language rewrite at a named reading level; Section 508 / WCAG 2.1 AA audit; Spanish-language review with bilingual workers. *Baseline:* the shipped English interface. *Gate:* audit report and named reviewer — a gate *on* the pilot, not a follow-on (H2).
- **Is the decision log sufficient as an audit artifact, and does it integrate into a real channel without a server?** Single-agency or state EVV help-desk pilot, staff reviewing exported logs against their own audit requirements. *Baseline:* the decision log as it exports today. *Gate:* pilot report (H2).

The design behind the third entry, because naming instruments is not stating a design. Target N is eighteen completed sessions, six in each of A.8′s three participant categories, the timed task block run at baseline and repeated at eight weeks for participants consenting to a second session. **The design is a single-group pre/post, descriptive by intent:** an agency's compliance workload moves with survey cycles, rule changes and staffing outside the study, so the design reads within-subject change rather than attribution. The **primary endpoints are therefore within-subject and tool-local**: per-question time-to-answer on the paired task set, and the participant's judgment of whether a decline was the right call and named the fact they needed. NASA-TLX, SUS and the CBI are **secondary context**, reported as paired change scores with confidence intervals rather than significance tests at that N, against a meaningful-difference threshold pre-registered from each instrument's published interpretive bands *before* the first session. Results are published as measured, in whichever direction they fall.

The first-half list is deliberately short. Phase 2 early carries four workstreams and no more, identical to §2′s: the recruited sessions, the Green hand-audit, compound-concept identity inside the continuous ratchet cycle, and the negation bound with its fifth adversarial category. Accessibility and Spanish-language review, the authority-strata probe, the publisher regrouping of the floors and the pilot are second-half work, sequenced *behind* the sessions rather than beside them. Sequencing them that way is what lets the first six months report results on a date a reviewer can check, which is the feasibility the timeline is built for.

A decision rule, stated in advance. If the compound-identity measurement shows term construction converting little of the missing-term population into cited answers, A.3′s premise is falsified and the plan changes rather than the metric: work moves to authority acquisition for the reachable stratum and to the grammar coverage behind `UnparsedKnownTerm`, with that result entering the same ratchet history as

everything else. A falsified premise gets published in place of a re-framed number, and the scoring abolition in A.4 is the precedent for how this project behaves.

A.8 The end-user session protocol

What follows is the instrument, specified so that someone other than its author could execute it — a protocol to be judged, not an intention. The input behind A.2 is institutionally mediated; recruited sessions begin with this protocol, none having run to date, and its first execution is the first Phase 2 milestone (§1, §2).

Participants, recruited separately, 5–8 per category: agency compliance and billing staff; direct support professionals and home care aides subject to EVV; and self-directing participants or their employer-of-record representatives, named because a self-direction program's *worker* and *participant* questions have different governing authority. **Recruitment** runs through A.9′s organizations plus state EVV program offices and provider associations, purposively sampled: at least three states, at least two agencies under ten employees, no fewer than two participants who report reading English with difficulty (interpreted sessions). **Consent and review:** written informed consent covering recording, quotation and withdrawal before any task; a written human-subjects determination from an IRB of record before recruitment opens, filed as the opening step of Phase 2; PHI excluded by design. **Compensation** at or above the participant's own hourly wage — direct care workers are paid for their time.

Format: remote, moderated, 60 minutes, recorded with consent; think-aloud during tasks, semi-structured interview after. **Task set:** six tasks drawn verbatim from this track's corpus slice, spanning A.2′s outcome classes — two questions the engine answers today, two compounds pending concept identity, one phrasing pending grammar coverage, one rule-governed question that should return a `Conditional` — with participants told in advance which outcomes the set is built to elicit. **Measures:** task completion and time-to-answer; NASA-TLX per task block; SUS at session end; and two judgments only a participant can supply — whether each decline was the *right* call, and whether the named missing fact was the one they needed. **Analysis:** two-coder thematic analysis, disagreement adjudicated against the recording; every insight enters a four-column traceability record — participant's own words → measured failure → committed change → enforced gate. **Stopping rule:** sessions continue within a category until two consecutive ones add no new severity-one finding or the cap of eight is reached, with the stopping point reported rather than assumed. **If recruitment runs past the window,** the fallback is written Phase 2 participation commitments from named individuals, published with the schedule they commit to, and the sessions still run. **Output:** a published session report carrying findings in both directions, and one new corpus question per distinct failure phrasing observed, added as a named failing test before any fix is written.

That output rule is why this is the same loop rather than a new one: the traceability record already exists in that shape, with corpus phrasing in its first column. Phase 2 changes *whose* utterance opens the row.

A.9 Partnership outreach ledger

The ledger records outreach, one row per contact: *date · organization · category · channel · ask · disposition* —. Every contact from the submission date forward is entered with its disposition, non-responses included, and the ledger is attached to any Phase 2 submission whatever it then says; the collaboration merit prize is left to a ledger that carries rows (§5).

Target set, in §2′s priority order: the eight state Medicaid EVV program offices whose FAQ documents are already corpus sources; EVV vendors operating state aggregators; HCBS provider associations; Area Agencies on Aging and ADRCs, the escalation destination an abstention already points toward; and self-direction financial management services entities. The ask is uniform, deliberately small, and written out here so it is executable rather than described:

> *"I have built a citation-grounded terminology tool for HCBS and EVV compliance questions. It answers a narrow set of them with the statute, regulation or state provision attached, declines everything else*

rather than guessing — naming the term it could not ground wherever it has one — and writes a decision record your staff can audit. It runs inside your own page: no server behind it, no query leaving the browser, no per-query cost, no data-sharing agreement on the query path. I am asking two things. One named person to read a single exported decision log against your own audit requirements and tell me where it would fail an auditor. Then, once that goes well, one channel willing to host the file. Either way, I record the ask and the answer."

That carries §2′s argument: an abstention routes a worker *into* the aging and disability network rather than away from it. The last sentence is deliberate — every answer is a row, and a ledger of rows is the evidence of a process that an unlogged intention cannot supply.