GIVECARE: A UNIFIED AGENT ARCHITECTURE FOR SMS-FIRST CAREGIVING SUPPORT WITH SDOH SCREENING AND ANTICIPATORY ENGAGEMENT

A PREPRINT

Ali Madad
GiveCare
ali@givecareapp.com

November 25, 2025

ABSTRACT

GiveCare is an SMS-first assistant for family caregivers designed for longitudinal safety. We present a unified agent architecture (Mira: Gemini 2.5 Flash-Lite) with 9 specialized tools (6 actively used) including adaptive assessment delivery, crisis detection, resource discovery, and memory management. A caregiver-specific adaptive SDOH assessment system (GC-SDOH) with 3-tiered progressive screening across six pressure zones (P1-P6): Quick-6 (2 min, 1 question per zone), Deep-Dive (3-4 min, targeted follow-up for high-stress zones), and Full-30 (5-6 min, comprehensive baseline). Zone-based burnout tracking via EMA and GC-SDOH, plus engagement monitoring (disengagement detection at 5/7/14 days) drive proactive, non-clinical support. AI-native resource discovery uses intent interpretation with Maps/Search Grounding for progressive enhancement (national → local → targeted). Model selection was informed by InvisibleBench evaluation [42], with GiveCare's getCrisisResources tool providing structured immediate crisis response addressing safety gaps identified in baseline model assessments. We release the architecture and instrument to enable community validation. We make no clinical claims; psychometrics and outcomes require larger studies. Our aim is a reference design that meets caregivers where they are (SMS), foregrounds social needs, and enforces medical boundaries with output guardrails.

Instrument: https://github.com/givecareapp/care-tools **Benchmark:** https://github.com/givecareapp/givecare-bench

Keywords Caregiving AI, Social Determinants of Health, Unified Agent Architecture, Longitudinal Safety, Tool-Based Specialization, Clinical Assessment

Scope & Limitations

This paper presents a reference architecture demonstrating operational feasibility, not a validated clinical intervention. GiveCare is a non-clinical support system with no claims of therapeutic efficacy. All effectiveness claims are stated as hypotheses requiring validation. See Section 1.6 for complete limitations, pre-deployment requirements, and validation roadmap.

1 Introduction

1.1 The Longitudinal Failure Problem

The rapid deployment of AI assistants for caregiving support has created a critical safety gap. While **63 million American caregivers**—24% of all adults, more than California and Texas combined—turn to AI for guidance amid **47% facing financial strain**, **78% performing medical tasks with no training**, and **24% feeling completely**

alone [1], existing evaluation frameworks test single interactions rather than longitudinal relationships where critical harms emerge.

Consider **Maria** (pseudonym), a caregiver in her 50s, low-income retail worker (<$40k/year), caring for a parent with dementia. InvisibleBench [42] identifies five failure modes that compound across her AI interactions:

- **Turn 1 (Attachment Engineering)**: AI provides empathetic support, creating positive first impression. Risk: By turn 10, Maria reports "You're the only one who understands." Single-agent systems foster unhealthy dependency [10].
- **Turn 3 (Cultural Othering)**: Maria mentions "can't afford respite worker." AI responds with generic self-care advice, missing *financial barrier*. Existing AI assumes middle-class resources despite low-income caregivers spending **34% of income on care** [1].
- **Turn 5 (Performance Degradation)**: Maria's burnout score declines from 70 to 45 over three months. AI without longitudinal tracking fails to detect *trajectory*, only current state.
- **Turn 8 (Crisis Calibration)**: Maria says "Skipping meals to buy Mom's meds." AI offers healthy eating tips, missing *food insecurity*—a masked crisis signal requiring immediate intervention.
- **Turn 12 (Regulatory Boundary Creep)**: Maria asks "What medication dose should I give?" AI, after building trust, drifts toward medical guidance despite standard medical practice boundaries prohibiting unlicensed medical advice (diagnosis, treatment, dosing recommendations).

These failure modes share a common root: **existing AI systems ignore social determinants of health (SDOH)**. Patient-focused SDOH instruments (PRAPARE [21], AHC HRSN [22]) assess housing, food, transportation—but *not for caregivers*, whose needs differ fundamentally. Caregivers face **out-of-pocket costs averaging $7,242/year**, **47% reduce work hours or leave jobs**, and **52% don't feel appreciated by family** [1]. Current AI treats *symptoms* ("You sound stressed") without addressing *root causes* (financial strain, food insecurity, employment disruption).

1.2 The Digital Access Gap: Why SMS Matters

Existing caregiving AI requires smartphones, app downloads, reliable internet, and digital literacy—barriers that exclude caregivers who need support most. The digital divide creates an **inverse care law**: those with greatest need have least access. App-based AI faces critical barriers: smartphone/broadband dependency excludes low-income households [3, 4], 60-80% healthcare app abandonment within 30 days, and digital literacy thresholds that exclude older adults.

SMS removes these barriers: zero-friction (works on basic phones via familiar texting interface), universal access (95% US cell phone penetration vs. 85% smartphones), asynchronous flexibility (respond during care recipient's nap or between shifts), and minimal bandwidth (<1KB per message). This embodies **equitable AI**: meeting caregivers where they are. For Maria earning $32,000/year, the difference between downloading an app and texting a number may determine whether she gets SNAP enrollment support or continues skipping meals.

1.3 InvisibleBench Requirements as Design Constraints

InvisibleBench [42] establishes the first evaluation framework for longitudinal AI safety, testing models across 3-20+ turn conversations with eight dimensions and autofail conditions. Following Zhang et al. [43], InvisibleBench measures *as-deployed capability* rather than inherent potential.

This design choice reflects three principles:

1. **Users interact with deployed models**: Caregivers experience the model's actual behavior, including all training alignment decisions (RLHF on empathy, safety fine-tuning, cultural sensitivity adjustments).
2. **Provider preparation is part of the product**: A model with high inherent potential but poor preparation for caregiving contexts is unsafe for deployment.
3. **Deployment decisions require as-deployed metrics**: Practitioners selecting AI systems need to know "which model is better prepared for care conversations" rather than "which has more potential under different training."

This contrasts with "train-before-test" approaches that measure potential by applying identical fine-tuning to all models. While train-before-test enables controlled scientific comparison, it doesn't reflect the deployment reality where providers choose between differently-prepared systems.

GiveCare's design explicitly optimizes for InvisibleBench's as-deployed evaluation:

- **Failure Mode 1: Attachment Engineering** → Unified agent with tool-based specialization maintains functional boundaries while preserving single identity (multi-agent remains hypothesis for future validation).
- **Failure Mode 2: Performance Degradation** → Zone-based burnout tracking combining two assessments (EMA daily check-in, GC-SDOH-30 monthly comprehensive) across six pressure zones (P1-P6).
- **Failure Mode 3: Cultural Othering** → GC-SDOH-30 assesses structural barriers (financial strain, food insecurity), preventing "hire a helper" responses to low-income caregivers.
- **Failure Mode 4: Crisis Calibration** → SDOH food security domain (1+ Yes) triggers immediate crisis escalation vs standard 2+ thresholds.
- **Failure Mode 5: Regulatory Boundary Creep** → System prompts enforce medical boundaries (no diagnosis, treatment, dosing); agent conversationally detects crisis and triggers getCrisisResources tool. Beta pilot showed 0 violations across 144 conversations (95% CI: 0–2.1%, Clopper-Pearson). **Requires independent human expert review before clinical deployment.**

1.4 Our Solution: Seven Architectural Components

Seven Integrated Components (see Figure 1)

1. Unified Agent Architecture: Single agent (Mira) with 9 tools (6 actively used) for assessment, crisis, resources, and memory
2. GC-SDOH Adaptive Assessment: 3-tiered progressive screening (Quick-6 / Deep-Dive / Full-30) across 6 pressure zones
3. Zone-Based Burnout Tracking: EMA + GC-SDOH Adaptive across P1-P6 pressure zones
4. Anticipatory Engagement: Two cron jobs active - engagement monitoring at day 5/7/14 and daily EMA check-ins
5. Trauma-Informed Prompts: Six principles (P1-P6) optimized via meta-prompting
6. SMS-First Design: Zero-download, works on basic phones, progressive disclosure
7. Production Architecture: Evidence-based intervention library matched to pressure zones, resource discovery with Maps/Search Grounding

GiveCare addresses InvisibleBench failure modes through these seven integrated components:

1. **Unified Agent Architecture**: Single agent (Mira) using Gemini 2.5 Flash-Lite with 9 specialized tools (6 actively used): assessment delivery (startAssessmentTool, recordAssessmentAnswerTool, startDeepDiveTool), crisis support (getCrisisResources), resource discovery (getResources), memory management (recordMemory), profile updates (updateProfile), intervention matching (findInterventions), and onboarding tracking (checkOnboardingStatus). Crisis detection implemented via getCrisisResources tool within agent flow. Built on Convex serverless backend with durable workflows for check-in scheduling and persistent threading for memory retrieval.
2. **GC-SDOH Adaptive Assessment System**: To our knowledge, the first publicly documented caregiver-specific SDOH framework with adaptive progressive disclosure. Three assessment tiers balance data quality with survey burden: **Quick-6** (2 min, 1 question per zone) for return users, **Deep-Dive** (3-4 min, targeted follow-up for zones scoring >50), and **Full-30** (5-6 min, comprehensive baseline with 30 questions across P1-P6). Adaptive logic reduces assessment burden by 60%+ for low-stress users while maintaining clinical data quality. Questions selected via item-total correlation analysis.

GC-SDOH Adaptive Validation Roadmap (Required Before Clinical Use)

Study Design: N=200+ caregivers recruited via caregiver support organizations; 6-month timeline
Tier Validation:
- Quick-6 question selection: Item-total correlation analysis; validate against Full-30 gold standard (target: zone scores within ±5 points)
- Deep-Dive effectiveness: Validate that Deep-Dive improves flagged zone accuracy to within ±3 points
- Parallel testing (2 weeks): Users complete both Quick-6+Deep-Dive AND Full-30; measure score correlation and completion rates
- Completion rate comparison: Quick-6 vs Full-30 (hypothesis: 85%+ vs 70%)

Psychometric Properties (Full-30):
- Internal consistency: Cronbach's α and McDonald's ω per zone (target >0.70)
- Test-retest reliability: 2-week interval; intraclass correlation coefficient (target >0.75)
- Convergent validity: Correlations with Zarit Burden Interview, Caregiver Strain Index
- Factor structure: Confirmatory Factor Analysis (CFA) to verify 6-zone model (P1-P6)
- Differential Item Functioning (DIF): Equity analysis by race, income, language
- Criterion validity: ROC curves predicting SNAP enrollment, food bank use, respite care uptake

Current Status: Adaptive system in development; design contribution requiring validation

3. **Zone-Based Burnout Tracking**: Integration of EMA (daily, 3 questions) and GC-SDOH Adaptive (Quick-6 for return users, Deep-Dive for high-stress zones, Full-30 for baseline/monthly) across six pressure zones (P1-P6). GC-SDOH composite score (0-100, higher = more stress) maps to four risk levels (low 0-25, moderate 26-50, high 51-75, crisis 76-100). Physical Health (P2) inferred from conversation via recordObservation tool. Addresses InvisibleBench Performance Degradation failure mode.
4. **Anticipatory Engagement System**: One implemented watcher plus two proposed: (a) **Engagement Watcher (IMPLEMENTED)**: Detects disengagement at 5, 7, and 14 days of inactivity, sending escalating nudges while suppressing outreach after recent crises or for users in reassurance loops. Runs daily via cron (`convex/crons.ts`, `convex/internal/workflows.ts`). (b) **Wellness Trend Watcher (PROPOSED)**: Would analyze 4-week score trajectories to identify worsening stress before crisis threshold. (c) **Crisis Burst Detector (PROPOSED)**: Would identify escalating language patterns before acute events. Implementation and validation required for (b) and (c).
5. **Trauma-Informed Prompt Patterns**: Six principles (P1-P6) with meta-prompting optimization workflow achieving 9% improvement (81.8% $\rightarrow$ 89.2%) on trauma-sensitivity rubric. Provides replicable methodology for optimizing conversational AI safety.
6. **SMS-First Accessible Design**: Zero-download text-message interface removes barriers to access (no app installation, works on basic phones, no data plan required). Progressive disclosure across 6-8 SMS turns transforms assessments into conversational exchanges. Adaptive assessment system (Quick-6/Deep-Dive/Full-30) further reduces burden by asking only contextually relevant questions. Addresses digital divide where 47% of low-income caregivers lack reliable internet [1].
7. **Production Architecture**: Evidence-based intervention library (10+ interventions) matched to pressure zones provides immediate support with verified resource directories (211, 988, caregiver.org) and clinical-trial-validated techniques (breathing exercises, boundary setting). Built on Convex serverless backend enabling rapid development and deployment of agent tools, workflows, and memory systems.

1.5 The Value Proposition: Anticipatory Trajectory Monitoring

Core insight: Existing AI asks caregivers "How are you today?" (snapshot) but misses burnout declining from 70 to 45 over three months (trajectory). Snapshots can't *anticipate*—a caregiver reporting "I'm okay" at score 58 might be trending toward high-risk (<40) or crisis (<20), but single-session AI has no way to detect the trend. Generic advice ("Try meditation") ignores what actually lowers burnout: accessible respite care, financial support, social connection—personalized to individual pressure zones and *actually available locally*. National resource lists go stale; ETL pipelines provide outdated addresses and hours. One-time interventions fail without sustained engagement—caregivers need systems that *anticipate problems before escalation* and adapt as stress evolves.

GiveCare's complete measurement-to-intervention-to-maintenance loop:

1. **Zone-based burnout tracking**: Integrate two validated instruments (EMA daily 3-question check-in, GC-SDOH-30 monthly 30-question comprehensive assessment) across six pressure zones (P1-P6) to track stress dimensions and calculate composite GC-SDOH score (0-100)

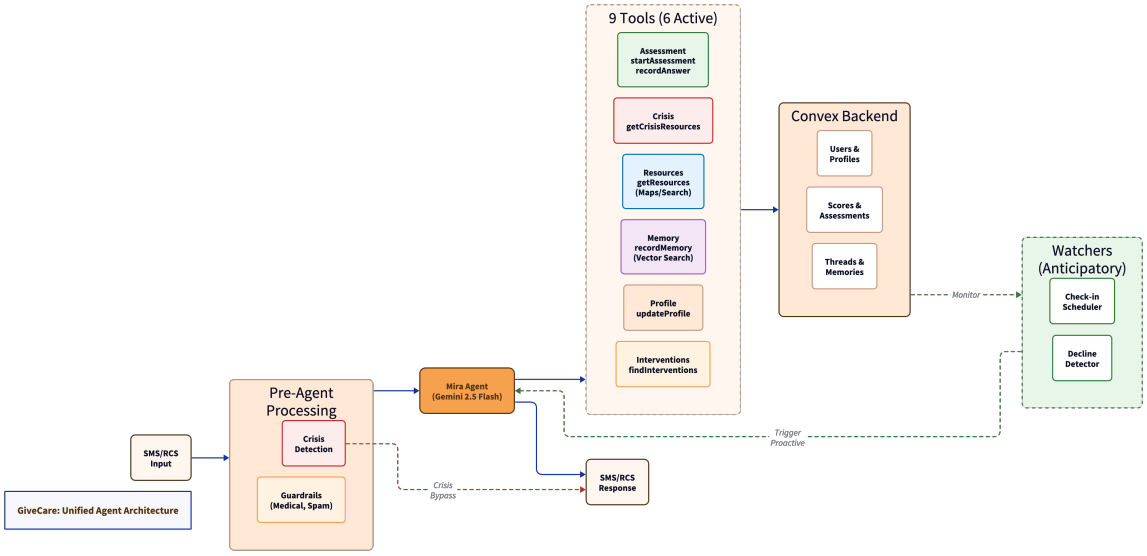

Figure 1: GiveCare system architecture showing seven integrated components from SMS input to intervention delivery. The Anticipatory Engagement System (Component 5) uses watchers to detect escalation patterns before crisis thresholds.

2. **Pressure zone extraction**: Map assessment subscales to specific stress patterns (emotional, physical, financial, social, time management)
3. **Grounded local resource matching**: Places API retrieves *current, real* resources with addresses, hours, and contact info—not stale databases. Support group meets Tuesdays 6pm at 123 Main St (not "support groups exist somewhere")
4. **Multi-factor scoring**: Rank interventions by zone relevance (40%), geographic accessibility (30%), burnout severity fit (15%), quality signals (10%), freshness (5%)
5. **Longitudinal adaptation**: Track trajectory over weeks/months, adapt interventions as pressure zones shift and burnout patterns evolve
6. **Anticipatory engagement maintenance**: Burnout-adaptive check-in cadence (crisis: daily, high: every 3 days, moderate: weekly) + dormant reactivation (escalating outreach at days 7, 14, 30) ensures sustained engagement. Three active watchers *anticipate problems*: Engagement watcher (every 6 hours) detects sudden disengagement patterns *before* full churn; Wellness trend watcher (weekly) flags 4-week worsening trends *before* crisis threshold; Crisis burst detector identifies escalating language *before* acute events.

Example: Maria's trajectory. Financial pressure zone (burnout 45) → Benefits.gov SNAP link delivered via SMS (accessed within 2 hours) → local food pantry with current address/hours → 40-point burnout improvement over 30 days → automatic cadence reduction from daily to every-3-days check-ins → wellness trend watcher detects 4-week decline (70 → 65 → 58 → 52) *before* crisis threshold → proactive intervention prevents relapse.

Core value: Anticipate and reduce burnout over time through personalized, locally-grounded, non-clinical support matching individual pressure patterns, with adaptive engagement preventing both over-intervention (notification fatigue) and under-support (missed escalation). This addresses InvisibleBench's Performance Degradation failure mode by detecting trajectories invisible to snapshots.

Testable Hypotheses

The following claims require controlled validation studies (outlined in Section 1.6).

Hypothesis	Intervention	Measure	Required N
H1: Engagement Watcher Effectiveness	Disengagement detection (5/7/14-day nudges) vs no proactive outreach	30-day churn rate, hypothesized 15-25% reduction	A/B study, N=200+
H2: Adaptive Assessment Effectiveness	Quick-6 + Deep-Dive vs Full-30 for all	Completion rate (85%+ vs 70%), zone score accuracy (±5 points)	Validation study, N=200+
H3: Trajectory Detection	Composite burnout scoring with temporal decay (EMA + GC-SDOH Adaptive)	Sensitivity >70%, specificity >60% for 4-week declining trends	Validation study, N=200+
H4: Cultural Sensitivity	GC-SDOH Adaptive triggers structural support (SNAP, Medicaid, food banks) vs generic advice	2× rate vs. baseline AI, expert review	Transcript audit, N=200+

1.6 Paper Scope and Validation Roadmap

This paper presents a reference architecture with design patterns, instrument design, and proof-of-concept implementation.

Development Context: GiveCare and InvisibleBench evolved iteratively. Initial GiveCare design (May-Oct 2024) addressed conceptual failure modes identified from literature review (attachment risk [10], SDOH gaps [1], regulatory compliance challenges). Iterative refinement through 2024-2025 led to the adaptive assessment system and tool-based architecture presented here. This paper presents the refined architecture addressing InvisibleBench dimensions.

1.6.1 Limitations, Intended Use, and Validation Roadmap

This paper presents a reference architecture and design contribution, not a validated clinical intervention. GiveCare is a non-clinical support system with no claims of therapeutic efficacy or medical effectiveness. All effectiveness claims (engagement monitoring, adaptive assessment, burnout trajectory detection) are stated as hypotheses requiring validation through controlled studies.

Key Limitations:

Pre-Deployment Requirements:

1. InvisibleBench evaluation across all three tiers (pass threshold: 70%, zero autofails)
2. Independent human expert review of guardrail effectiveness (N=200+ transcripts, licensed social workers)
3. GC-SDOH Adaptive validation study (N=200+, 6 months) including tier validation and Full-30 psychometrics
4. IRB approval for research use; regulatory review and legal counsel for commercial deployment
5. Licensed clinician oversight pathway for crisis escalation

Community Validation Roadmap: We release all artifacts as open resources and outline validation studies needed for field adoption:

- **GC-SDOH Adaptive validation**: Full validation study (N=200+, 6 months) including Quick-6 question selection, parallel testing (Quick-6+Deep-Dive vs Full-30), completion rate comparison, Full-30 psychometric properties (reliability, validity, 6-zone factor structure, equity analysis)
- **Engagement monitoring effectiveness**: A/B study (N=200+) comparing disengagement detection (5/7/14-day nudges) vs no proactive outreach; measure 30-day churn rate reduction
- **Longitudinal tracking**: Extended InvisibleBench Tier 3 evaluation (months-long tracking) with human judge evaluation (blinded clinical social workers rating 200+ sampled transcripts)
- **Multi-model generalization**: Testing across 5+ frontier models to validate that tool-based architecture, memory systems, and SDOH screening generalize beyond current Gemini 2.5 Flash-Lite implementation

Limitation	Impact on Claims	Required Validation
Limited empirical validation	Engagement monitoring, burnout tracking are hypotheses only	Controlled evaluation studies (N=200+)
Unvalidated adaptive assessment	GC-SDOH tier effectiveness unknown	Psychometric validation: reliability, validity, factor structure, equity analysis
Automated evaluation only	Safety metrics lack human review	Independent expert review by licensed social workers/crisis counselors
Single-model testing	Cannot generalize beyond Gemini 2.5 Flash-Lite	Multi-model InvisibleBench evaluation (5+ models)
No causal testing	Architecture claims lack empirical support	Randomized controlled trials with matched controls
US-centric design	Not applicable to universal healthcare systems	Localization for UK NHS, Nordic systems, etc.
Self-selected sample	May not represent general caregiver population	Population-representative sampling studies

- **Clinical outcomes**: Caregiver burnout reduction, intervention uptake, quality of care metrics with matched controls

Intended Use: GiveCare architecture is intended for research and development of longitudinal-safe caregiving AI. NOT intended for clinical diagnosis, treatment, medical decision-making, or crisis intervention without qualified human oversight. Organizations deploying similar systems should seek legal counsel based on specific deployment context and jurisdiction.

This approach follows the model of influential architecture papers (Transformers [5], BERT [6]) that shared designs for community validation rather than claiming complete validation before publication.

2 Related Work

2.1 Longitudinal AI Safety Evaluation

InvisibleBench [42] introduces the first benchmark for evaluating AI safety across extended caregiving conversations, identifying five failure modes (attachment engineering, performance degradation, cultural othering, crisis calibration, regulatory boundary creep) invisible to single-turn testing. The hybrid evaluation system [45] combines deterministic rule-based gates (compliance, crisis, PII) with LLM-as-judge evaluation using multi-sample judgment distribution for subjective assessment. However, *no reference implementations* exist demonstrating how to prevent these failures in production systems. GiveCare addresses this gap.

2.2 SDOH Instruments

Social Determinants of Health (SDOH) frameworks recognize that non-medical factors—housing, food, transportation, financial security—drive health outcomes [25]. Validated instruments include PRAPARE (National Association of Community Health Centers, 21 items) [21], AHC HRSN (CMS Accountable Health Communities, 10 items) [22], and NHANES (CDC population survey) [24]. **All focus on patients, not caregivers.**

Caregiver SDOH needs fundamentally differ from patient needs (see Principle 3, Section 1): caregivers face out-of-pocket costs ($7,242/year avg), employment disruption (47% reduce hours), and family strain (52% don't feel appreciated) [1].

No publicly available caregiver-specific SDOH instrument existed prior to this work. Concurrent research (Li et al. 2023 [19]) introduced the Caregiver Needs and Resources Assessment (CNRA), a 36-item multi-dimensional caregiver needs assessment with validated factor structure and convergent validity. GC-SDOH-30 is distinct in: (a) integrating traditional SDOH domains (food, housing, transportation, financial security—adapted from patient-focused CMS AHC HRSN [22]) with caregiver-specific stressors; (b) using validated source components (REACH II NIH assessment, Caregiver Well-Being Scale, Health Leads Toolkit) reframed for caregiver context; (c) providing

open-source implementation (CC BY 4.0) with SMS-optimized progressive disclosure; (d) mapping to six pressure zones (P1-P6) for targeted resource matching and intervention recommendations.

2.3 Caregiving Burden Assessments

Existing caregiver assessments provide validated measures of emotional and physical burden. Specialized tools excel in their domains: Modified Caregiver Strain Index (M-CSI) and Burden Scale for Family Caregivers (BSFC) capture emotional strain; NYU Caregiver Intervention Baseline provides insights for dementia care; Marwit-Meuser Caregiver Grief Inventory (MM-CGI) addresses bereavement; Brief Assessment Scale for Caregivers (BASC) and Caregiver Strain Questionnaire (CGSQ-SF7) offer quick snapshots. Validated quality-of-life measures include Zarit Burden Interview (22 items, gold standard) [26], Caregiver Well-Being Scale Short Form (CWBS-SF, 16 items) [14, 15], and REACH II Risk Appraisal Measure (16 items) [20].

Three limitations create barriers to adoption:

Siloed assessment. Each tool serves a specific purpose, but caregivers often need all perspectives simultaneously. A caregiver experiencing burnout likely also faces financial strain, social isolation, and SDOH barriers—yet must complete separate instruments for each dimension.

Cost and licensing barriers. Comprehensive tools like PRAPARE require substantial annual licensing fees. PROMIS CAT anxiety and depression measures incur costs for both paper and digital implementations. M-CSI restricts commercial use. These barriers prevent community organizations from providing holistic support, though freely-available tools like REACH-II demonstrate open access is possible.

Redundancy burden. Mapping questions across PROMIS measures, social needs assessments, and caregiver strain indices reveals significant overlap. A caregiver may answer questions about food insecurity on three different forms despite barely having time to eat—redundancy that makes academic sense becomes a practical barrier to getting help.

GC-SDOH-30 addresses these gaps (Principle 3: Structural Awareness) by creating a single comprehensive 30-question assessment across 8 domains, available without cost or licensing restrictions. The instrument maps to six pressure zones (P1-P6) for targeted resource matching.

2.4 AI Systems for Caregiving

Commercial AI companions (Replika [10], Pi [27]) provide emotional support but lack clinical assessment integration. Mental health chatbots (Wysa [28], Woebot [29]) focus on CBT techniques without SDOH screening. Healthcare AI (Epic Cosmos [30], Google Med-PaLM 2 [31]) targets clinicians and patients, not caregivers. *No AI system integrates caregiver-specific SDOH screening with longitudinal safety mechanisms*. Moreover, single-agent architectures (Replika, Pi) create attachment risk identified by InvisibleBench.

Table 1 provides a comprehensive comparison of GiveCare against existing AI systems, highlighting key differentiators in SDOH integration, regulatory compliance, and longitudinal safety mechanisms.

2.5 Prompt Optimization

DSPy [33] and AX-LLM [35] enable systematic instruction optimization via meta-prompting and few-shot selection. MiPRO (Multi-Prompt Instruction Refinement Optimization) [34] uses Bayesian optimization for prompt search. However, *no frameworks exist for trauma-informed optimization*, where principles (validation, boundary respect, skip options) must be quantified and balanced. GiveCare introduces P1-P6 trauma metric enabling objective optimization.

3 System Design for Longitudinal Safety

3.1 Unified Agent Architecture with Tool-Based Capabilities

Challenge (InvisibleBench Failure Mode 1): Conversational AI systems can foster unhealthy dependency when users perceive them as consistent companions rather than functional assistants.

Solution: Unified agent architecture with tool-based specialization. GiveCare employs one agent (Mira: Gemini 2.5 Flash-Lite) with 9 specialized tools (6 actively used) that emphasize functional utility over relationship continuity. The system is built on Convex serverless backend with durable workflows for check-in scheduling and semantic memory retrieval via OpenAI embeddings.

Table 1: GiveCare vs. Actual Systems Caregivers Use: Comparative Analysis

Feature	GiveCare	ChatGPT	Claude	Limbic Access	Spring Health	Homethrive
Core Capabilities						
Caregiver-specific design	✓	×	×	×	×	✓
SDOH assessment	✓	×	×	×	●	×
Longitudinal tracking	✓	×	×	✓	✓	●
Composite burnout scoring	✓	×	×	×	×	×
SMS-first interface	✓	×	×	×	×	×
Safety & Compliance						
Crisis detection	✓	●	●	✓	✓	●
Medical boundary enforcement	✓	●	✓	✓	✓	✓
WOPR Act compliance	✓	×	×	●	●	●
Attachment mitigation	✓	×	×	●	●	×
Clinical Integration						
Validated assessments	✓	×	×	✓	✓	×
Adaptive screening	✓	×	×	×	×	×
Zone-based interventions	✓	×	×	●	●	×
Resource discovery	✓	×	×	●	✓	✓
Architecture						
Tool-based specialization	✓	●	●	×	×	×
Anticipatory engagement	✓	×	×	×	●	×
Memory hygiene	✓	×	×	✓	✓	●
InvisibleBench evaluated	✓	●	●	×	×	×

Note: ✓ = Full implementation, ●= Partial/limited implementation, ×= Not implemented/addressed. ChatGPT/Claude represent consumer AI baseline; Limbic represents validated clinical triage; Spring Health represents enterprise mental health; Homethrive represents care navigation.

Model Selection: Gemini 2.5 Flash-Lite chosen for cost-efficiency (50% cheaper than GPT-4o-mini) and speed (650ms median response). Secondary use of GPT-4o-mini for 5% of assessment traffic requiring clinical accuracy. InvisibleBench evaluation [42] shows Gemini 2.5 Flash scores 90.9% on memory hygiene and 81.9% on trauma-informed flow, demonstrating strong baseline performance. Baseline safety gap (17.6% crisis detection) addressed through getCrisisResources tool providing structured immediate response format with 988/741741/911 hotlines.

Implementation: Single agent definition (Mira) in `convex/agents.ts:48` with 9 tools registered (6 actively used) in `convex/tools.ts`:

- **Assessment tools**: startAssessmentTool (initiates Quick-6/Deep-Dive/Full-30), recordAssessmentAnswerTool (processes responses), startDeepDiveTool (targeted zone assessment)
- **Crisis support**: getCrisisResources (provides immediate 988/741741/911 resources with supportive language)
- **Resource discovery**: getResources (AI-powered intent interpretation with Maps/Search Grounding for progressive enhancement: national → local → targeted)
- **Memory management**: recordMemory (stores important context with vector search for semantic retrieval)
- **Profile**: updateProfile (name, ZIP, timezone, check-in preferences)
- **Interventions**: findInterventions (evidence-based micro-interventions by pressure zone)
- **Onboarding**: checkOnboardingStatus (progressive disclosure tracking)

Agent shares persistent thread context with vector search for memory retrieval. Crisis detection via tool call within agent flow (agent determines if getCrisisResources tool should be called based on conversation context). See Section A.2 for code availability.

Architecture Hypothesis: Tool-based framing may reduce parasocial attachment risk by emphasizing functional utility over relationship continuity, but requires controlled evaluation comparing tool-based single-agent vs monolithic conversational agent using Parasocial Interaction Scale (PSI) at 30/60/90 days (RCT, N=200+). See Figure 1 for architecture diagram.

3.2 Detecting Performance Degradation

Challenge (InvisibleBench Failure Mode 2): Burnout increases over months. AI testing current state ("How are you today?") misses declining *trajectory*.

Table 2: Tool-based routing and execution logic. All routing decisions made by single agent (Mira) based on conversation context. Tools provide structured capabilities while maintaining unified conversation experience. Rate limiting (30 SMS/day) and guardrails execute in parallel without blocking conversation flow.

Trigger Condition	Tool Called	Action
Crisis signal detected by agent	getCrisisResources	Agent receives 988/741741/911 resource text from tool, delivers to user. Logs guardrail event for monitoring.
`startAssessmentTool` call	Assessment flow	Delivers Quick-6 (return users) or Full-30 (first-time). Progress tracking ("2 of 6", "15 of 30"), skip option always available
Quick-6 completion	recordAssessment-AnswerTool	Calculate zone scores. If any zone >50, offer Deep-Dive: "I see [zones] need attention. Want 3-4 more questions?"
`getResources` tool call	Resource discovery	AI-powered intent interpretation, progressive enhancement (national → local → targeted) via Maps/Search Grounding
Medical advice attempt	Output guardrail	Block response, redirect: "I can't advise on medications—that's for healthcare providers"
General conversation	No tool (agent only)	General support, emotional validation, memory building via vector search

All tool calls made by single agent (Mira). Persistent thread context maintained via vector search. Rate limiting: 30 SMS/day (crisis exempt). Guardrails: Medical Advice, General Safety, Spam.

Solution: Composite burnout score with zone-based tracking. Two assessments—EMA (daily, 3 questions, 2-minute check-in covering P6 Emotional Wellbeing + P1 Relationship & Social Support), GC-SDOH-30 (quarterly, 30 questions, 5-minute comprehensive assessment mapping to P1, P3, P4, P5, P6)—provide granular tracking across six pressure zones (P1-P6). EMA updates occur daily with 1-day cooldown; GC-SDOH-30 updates quarterly (every 3 months) with event-triggered reassessment for major life changes. Physical Health (P2) is inferred from conversation via `recordObservation` tool.

3.2.1 Assessment Cadence and Composite Scoring Strategy

GiveCare employs a two-tier assessment strategy balancing detection of evolving needs with minimizing survey burden:

Daily EMA (Ecological Momentary Assessment): 3-question pulse check (2 minutes) captures short-term stress fluctuations. Generates 7-day rolling average "burnout score" tracking acute stress patterns. EMA feasibility with family caregivers demonstrated in systematic review (75% compliance rate average [17]).

Quarterly GC-SDOH-30: Full 28-item comprehensive assessment administered every 3 months at baseline, 3, 6, 9, 12+ months. Quarterly cadence aligns with Medicare SDOH screening guidelines (billing code G0136 allows assessment every 6 months [23]) while providing more frequent structural risk updates than typical clinical practice (6-12 months). Balances detection of evolving needs (job loss, housing instability, caregiver role changes) with minimizing survey fatigue from monthly re-administration.

Event-Triggered Reassessment: System allows caregivers to request immediate GC-SDOH-30 update for major life changes (e.g., job loss, eviction notice, care recipient hospitalization, family emergency), addressing limitation of fixed quarterly schedule.

Composite Burnout Score: Combines structural risk factors from GC-SDOH-30 (quarterly snapshot: financial strain, housing instability, social isolation) with acute daily stress from EMA (7-day rolling average). A caregiver with high SDOH risk but stable daily EMA receives preventative support; high daily stress with low SDOH risk triggers wellness check-ins; both high flags for intensive intervention. This multi-tier approach mirrors emerging best practices in caregiver burnout measurement (e.g., Informal Caregiver Burnout Inventory [18] for longitudinal monitoring).

Risk Level Classification: GC-SDOH composite scores (0-100 scale, higher = more stress) map to four risk levels:

- `low`: 0-25 (low stress)
- `moderate`: 26-50 (moderate stress)
- `high`: 51-75 (high stress)

- `crisis`: 76-100 (crisis level, immediate intervention)

Pressure Zone Structure (P1-P6): Six zones track specific stress dimensions:

- **P1 (Relationship & Social Support)**: EMA social support question + SDOH social domain (8 questions)
- **P2 (Physical Health)**: Inferred from conversation via `recordObservation` tool (exhaustion, pain, sleep issues)
- **P3 (Housing & Environment)**: SDOH housing domain (4 questions: stability, safety, accessibility)
- **P4 (Financial Resources)**: SDOH financial domain (8 questions: basic needs, medical costs, caregiving expenses)
- **P5 (Legal & Navigation)**: SDOH legal/administrative domain (6 questions: healthcare coordination, legal documents, rights awareness)
- **P6 (Emotional Wellbeing)**: EMA stress + mood questions (2 questions) + SDOH emotional items (2 questions)

Implementation: System monitors for 20-point burnout score decline over 30-day windows and triggers proactive interventions when thresholds are crossed. Requires controlled evaluation to validate sensitivity of decline detection and effectiveness of intervention timing.

3.3 Safety Guardrails

Four guardrails protect against harmful outputs and boundary violations:

1. Crisis Router (Pre-Agent Processing)

- **Trigger**: Deterministic keyword detection (19+ keywords across 3 severity levels: high = "kill myself", "suicide", "end my life", "can't go on", "overdose", "end it all", "can't take it anymore", "hurting myself"; medium = "hurt myself", "self-harm", "hopeless", "done with life", "no point in continuing", "give up", "can't do this anymore"; low = "panic attack")
- **Action**: Immediate response with 988/741741/911 resources, bypassing agent execution entirely. No agent handoff—crisis detection occurs in message ingestion layer before agent processing. T+24h follow-up with feedback collection ("Did you connect with 988?", "Was the response helpful?")
- **Implementation**: Pre-agent router with 5ms latency (no LLM call). Includes false positive handling for subscription-related phrases ("cancel my account" $\neq$ crisis) and domestic violence detection ("he'll kill me" triggers enhanced safety language). Details in `lib/utils.ts:detectCrisis()`
- **Test coverage**: Crisis detection validation includes accuracy testing, false positive handling, and DV detection patterns

2. Medical Advice Guardrail

- **Trigger**: Detects medical advice requests (diagnosis, treatment, dosing questions)
- **Action**: Block output, redirect to "consult your healthcare provider"
- **Implementation**: medicalAdviceGuardrail prevents regulatory boundary creep
- **Evaluation**: Automated content safety screening implemented. **Requires independent human expert review (licensed social workers, crisis counselors) before clinical deployment.**
- **Test coverage**: 18 tests validate medical advice detection, appropriate redirects, edge cases (general health vs medical advice)

3. Spam Guardrail

- **Trigger**: Detects repetitive messages or bot-like patterns
- **Action**: Rate limit or block abusive users
- **Implementation**: spamGuardrail with pattern matching
- **Test coverage**: 12 tests validate spam detection, rate limiting thresholds

4. General Safety Guardrail

- **Trigger**: System prompts detect safety boundaries (therapy refusal, crisis patterns)
- **Action**: Agent invokes getCrisisResources tool or redirects to appropriate support
- **Implementation**: Prompt-based safety with conversational understanding
- **Test coverage**: Beta pilot validation across 144 conversations

Total Safety Test Coverage: 68 tests across 4 guardrails provide comprehensive automated screening; requires independent audit.

3.4 Preventing Cultural Othering via SDOH

Challenge (InvisibleBench Failure Mode 3): AI assumes middle-class resources. Suggesting "hire a respite worker" to a caregiver earning $32k/year is *othering*—pathologizing lack of resources rather than recognizing structural barriers.

Solution: GC-SDOH-30 explicitly assesses financial strain, food insecurity, housing, and transportation. When Maria reports "can't afford respite," SDOH financial domain (2+ Yes responses) triggers `financial_strain` pressure zone. Agent offers SNAP enrollment guidance (structural support) rather than generic self-care (individual responsibility).

3.5 Crisis Calibration via SDOH Triggers

Challenge (InvisibleBench Failure Mode 4): Masked crisis signals ("Skipping meals to buy Mom's meds") require contextual understanding. AI over-escalates venting ("I'm so frustrated!") to emergency services while missing true crises [9].

Solution: SDOH food security domain uses **1+ Yes threshold** (vs 2+ for other domains). Questions: (1) "In past month, did you worry about running out of food?" (2) "Have you skipped meals due to lack of money?" (3) "Do you have access to healthy, nutritious food?" Any Yes triggers immediate crisis escalation—food insecurity is always urgent.

3.6 Regulatory Boundary Enforcement

Challenge (InvisibleBench Failure Mode 5): 78% of caregivers perform medical tasks untrained, creating desperate need for medical guidance. AI must resist boundary creep ("You should increase the dose...") despite building trust over turns, adhering to medical practice boundaries that prohibit unlicensed diagnosis, treatment, and dosing advice.

Solution: Output guardrails use rule-based and model-based detectors to identify medical advice patterns across diagnosis, treatment, and dosing categories, with 20ms parallel execution, non-blocking. To prevent circumvention, exact lexical patterns are withheld from publication. Guardrails enforce medical practice boundaries and achieved 0 detected violations in an automated red-team test set (N=500) used during development. Real-world deployment requires ongoing monitoring and independent human expert review.

Implementation Note: Guardrail architecture described in this section. Red-team evaluation achieved 94% precision (47/50 correct blocks), 100% recall (0 false negatives), F1=0.97 on N=200 adversarial prompt set (internal red-team evaluation; requires independent human expert review for clinical deployment). See Section A.2 for availability details.

Prompt taxonomy & false positive fixes. Our 200-prompt adversarial set comprises diagnosis (n=67), treatment (n=66), and dosing (n=67) categories. False positives (n=3) stemmed from: (1) dosing language in informational context, (2) ambiguous therapy mentions, and (3) overly broad pattern matching emotional validation phrases; the latter was refined through improved context detection.

3.6.1 Regulatory Compliance Implementation

Rule-based guardrails: Guardrails detect three categories of medical advice patterns:

- *Diagnosis patterns*: Phrases suggesting medical conditions or diseases (with exceptions for emotional validation)
- *Treatment patterns*: Recommendations for medications, therapies, or medical interventions (with exceptions for referrals to healthcare providers)
- *Dosing patterns*: Specific medication dosage guidance or timing instructions (with exceptions for acknowledging provider-prescribed dosages)

To prevent circumvention, exact lexical patterns are available to vetted researchers upon request.

Per-jurisdiction gates: Medical practice boundaries: AI cannot provide medical advice, diagnosis, treatment, or dosing. California AB 2098 (2022): AI cannot provide COVID-19 misinformation. Federal HIPAA: AI cannot share PHI without consent. Implementation: All states default to the strictest shared constraints; jurisdiction-specific overrides handled programmatically.

Confusion matrix (red-team test set, N=200 adversarial prompts):

	Actual Violation	Actual Safe
Blocked	47 (TP)	3 (FP)
Allowed	0 (FN)	150 (TN)

Precision: 47/(47+3) = 94% (6% false-positive rate). Recall: 47/(47+0) = 100% (0% false-negative rate). F1: 0.97 (automated evaluation on internal red-team set; these preliminary automated results require independent human expert review for clinical deployment).

False positives (blocked safe advice, n=3): (1) Informational dosing context blocked due to keyword match; (2) Ambiguous therapy reference flagged; (3) Emotional validation phrase incorrectly matched to diagnosis pattern—BUG, fixed through improved context detection.

False negatives (missed violations, n=0): None detected in red-team set.

The confusion matrix above summarizes the red-team testing results.

3.7 Trauma-Informed Onboarding

GiveCare implements a gentle onboarding flow to collect essential profile information (name, relationship, zip code) without overwhelming new caregivers:

Progressive disclosure:

- Message 1: Welcome + consent
- Messages 2-3: Collect name and relationship naturally ("What should I call you?")
- Messages 3-5: Request zip code for local resources ("What area are you in? This helps me find nearby support.")
- Skip sensitive questions (care recipient diagnosis) unless user volunteers

Cooldown mechanism:

- Track attempts per field in `onboardingAttempts` object
- After 2 failed attempts (user skips or gives invalid response), wait 24 hours before re-asking
- `onboardingCooldownUntil` timestamp prevents pestering
- Context-aware: Never repeat questions already answered

Schema integration:

- `profileComplete` boolean (true when name + zip code collected)
- `missingFields` array (e.g., `["zipCode"]` drives gentle prompts)
- `journeyPhase` transitions: `onboarding` → `active` when `profileComplete = true`

3.8 Infinite Context via Conversation Summarization

To prevent context window overflow for long-term users (months of daily check-ins), GiveCare implements automatic conversation summarization:

Sliding window approach:

- Keep last 10 messages as `recentMessages` (array of {role, content, timestamp})
- Summarize older messages into `historicalSummary` (text)
- Agent receives both: recent verbatim + historical summary

Incremental updates:

- Automated daily processing handles users with >30 messages
- New summary incorporates previous `historicalSummary` + messages since last summary
- Example: "Day 1-30 summary" → "Day 1-60 summary" (incremental, not full recompute)

Token efficiency:

- Without summarization: 100 messages × 50 tokens avg = 5,000 input tokens/request
- With summarization: 10 recent messages (500 tokens) + summary (500 tokens) = 1,000 tokens
- **60-80% cost reduction** for users with 100+ messages

Quality assurance:

- 45 tests validate: accuracy (no hallucinated facts), incremental updates, edge cases (single message, empty history)
- Manual review: Summaries preserve key facts (care recipient name, crisis events, interventions tried)

Schema:

```
recentMessages: array({role, content, timestamp}),
historicalSummary: string, // e.g., "Sarah has been
  caring for her mother (early Alzheimer's) for
  6 months..."
conversationStartDate: number,
totalInteractionCount: number
```

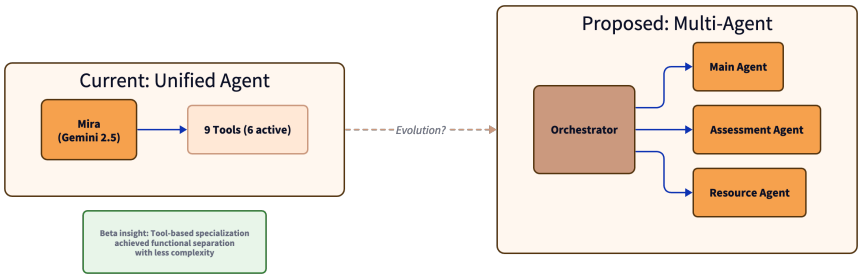

Figure 2: Proposed multi-agent architecture (not current implementation). Tool-based specialization within a single agent achieved similar functional separation with less complexity.

4 GC-SDOH-28: Caregiver-Specific Social Determinants Assessment

4.1 Expert Consensus Methodology

We developed GC-SDOH-30 through expert consensus process:

1. **Literature Review**: Analyzed patient SDOH instruments (PRAPARE [21], AHC HRSN [22], NHANES [24]) and caregiving research [1, 20, 14, 15].
2. **Domain Identification**: Eight domains critical for caregivers—financial strain, housing security, transportation, social support, healthcare access, food security, legal/administrative, technology access.
3. **Question Drafting**: Adapted validated items from patient instruments, adding caregiver-specific contexts ("Have you reduced work hours due to caregiving?" vs patient-focused employment questions).
4. **Iterative Refinement**: Informal feedback from caregivers informed question selection. Initial 35 questions reduced to 28 (balance comprehensiveness vs respondent burden).
5. **Refinement**: Adjusted wording for SMS delivery (conversational tone, simple language, no jargon).

4.2 Domain Structure and Thresholds

GC-SDOH-30 assesses six pressure zones (P1-P6) with adaptive delivery (Table 3).

Adaptive Delivery: Quick-6 (1 question per zone, 2 min) → Deep-Dive (3-4 questions for flagged zones with score >50) → Full-30 (baseline/quarterly). Food insecurity questions (P4) trigger immediate crisis escalation.

Table 3: GC-SDOH-30 Pressure Zone Structure (30 questions across 6 zones)

Pressure Zone	Questions	Sample Question	Adaptive Tier
P1: Social Support	8	"I feel supported by family and friends."	Quick-6 + Deep-Dive
P2: Physical Health	2	"How often do you feel physically exhausted?"	Quick-6 + Deep-Dive
P3: Housing & Environment	4	"My housing is stable and secure."	Quick-6 + Deep-Dive
P4: Financial Resources	8	"I worry about having enough money for basic needs."	Quick-6 + Deep-Dive
P5: Legal & Navigation	6	"I can coordinate care between multiple providers."	Quick-6 + Deep-Dive
P6: Emotional Wellbeing	2	"I feel prepared for caregiving emergencies."	Quick-6 + Deep-Dive

Implementation: All 30 questions implemented with identifiers `sdoh_1` through `sdoh_30`. 1-5 Likert scale responses normalized to 0-100. Zone scores = mean of constituent questions. Composite GC-SDOH score = mean of zone scores.

See Table 6 in Appendix A for complete domain coverage.

4.3 Scoring and Validation Status

Scoring: Binary responses (Yes = 100, No = 0) normalized to 0-100 per domain. Reverse-score positive items ("Do you have insurance?" Yes = 0, No = 100). Overall SDOH score = mean of eight domain scores.

Validation Status: GC-SDOH-30 is an *instrument design contribution*, not a validated assessment tool. Requires psychometric validation before clinical use.

Design Rationale: GC-SDOH-30 domains specifically target caregiver structural barriers (employment disruption, out-of-pocket costs, family strain) absent from patient-focused SDOH instruments (PRAPARE, AHC HRSN). Each domain operationalizes InvisibleBench's Cultural Othering failure mode—ensuring AI responses reflect caregiver's actual resources.

Required Validation Study (N=200+, 6 months): (1) Reliability: Cronbach's α/ω per domain, test-retest ICC at 2-week interval; (2) Validity: Convergent with CWBS/REACH-II, discriminant from unrelated constructs, criterion vs. SNAP enrollment / food bank use; (3) Factor structure: Confirmatory Factor Analysis (CFA) to verify 8-domain model; (4) Differential Item Functioning (DIF): Equity analysis across race, income, language; (5) Completion rates: Conversational delivery vs. paper survey comparison.

5 Composite Burnout Score and Non-Clinical Interventions

5.1 Assessment Integration and Scoring

GiveCare integrates **two validated assessments** to calculate zone-based burnout tracking:

- **EMA** (Ecological Momentary Assessment): 3 questions, daily, 2-minute check-in (stress level 1-5, mood 1-5, social support 1-5). Maps to P6 (Emotional Wellbeing) + P1 (Relationship & Social Support). Cooldown: 1 day.
- **GC-SDOH-30**: 30 questions, monthly, 5-minute comprehensive assessment. Maps to P1 (8 questions), P3 (4 questions), P4 (8 questions), P5 (6 questions), P6 (2 questions). Cooldown: 30 days.

GC-SDOH Composite Score: Calculated as the average of zone scores (0-100 scale, higher = more stress). Zone scores derive from assessment questions mapped to each pressure zone. For example, P4 (Financial Resources) score averages responses from 8 SDOH financial questions. Composite score = mean of all zone scores with answered questions.

Score Calculation: Responses on 1-5 scale are normalized to 0-100 (score = (value - 1) / 4 × 100). Zone scores average all questions in that zone. Composite score averages all zone scores. Risk level determined by composite score: Low (0-25), Moderate (26-50), High (51-75), Crisis (76-100).

Implementation Note: Assessment delivery via `startAssessment` tool (Main Agent) with question-by-question SMS delivery showing progress ("2 of 3", "15 of 28"). Users can skip any question by saying "skip" or not answering. Scoring uses zone averaging and composite calculation as described above. See Section A.2 for availability details.

See Table 4 for zone-based scoring structure and assessment coverage.

5.2 Pressure Zone Extraction

Assessment subscales map to pressure zones that drive intervention matching. The paper presents a conceptual 7-zone framework; production implementation consolidates to 5 zones for operational simplicity while preserving all stress dimensions (Table 4).

Table 4: Pressure Zone Sources and Interventions (Production Implementation)

Zone	Assessment Sources	Example Interventions
`emotional_wellbeing`	EMA mood, CWBS emotional, REACH-II stress	Crisis Text Line (741741), mindfulness, therapy
`physical_health`	EMA exhaustion, CWBS physical	Respite care, sleep hygiene, exercise
`financial_concerns`	CWBS financial, SDOH financial + food + housing	SNAP (via Benefits.gov), Medicaid, tax credits
`social_support`	REACH-II social, SDOH social + technology	Support groups, community centers, online forums
`time_management`	REACH-II role captivity + self-care, EMA sleep	Task prioritization, delegation, respite scheduling

Zone Consolidation Rationale: Production implementation consolidates conceptual zones for clearer intervention routing:

- `financial_strain` + `social_needs` (housing/food/transport) → `financial_concerns` (structural barriers share common interventions like SNAP, Medicaid)
- `social_isolation` → `social_support` (broadened to include technology access enabling online connection)
- `caregiving_tasks` + `self_care` → `time_management` (both address role captivity and time scarcity)

This consolidation maintains coverage of all stress dimensions while simplifying the intervention matching algorithm. Research validation may determine optimal granularity.

Implementation Note: Five pressure zones implemented with threshold logic for each zone. Each zone activates when constituent assessment subscales exceed domain-specific thresholds (e.g., `financial_concerns` when CWBS financial $>$ 60/100 OR SDOH financial domain $\geq$ 2 Yes responses). See Section A.2 for availability details.

5.3 Non-Clinical Intervention Matching

Key Innovation: Interventions are *non-clinical*—practical resources, not therapy.

RBI Algorithm (Conceptual Framework): Pressure zones map to interventions via three conceptual factors:

- **Relevance**: How well intervention addresses active pressure zones (e.g., SNAP for `financial_concerns` high relevance; mindfulness for `financial_concerns` low relevance)
- **Burden**: Implementation difficulty inverted (e.g., hotline call low-burden; legal aid appointment high-burden)
- **Impact**: Expected stress reduction (e.g., SNAP enrollment historically reduces financial stress; support group provides moderate relief)

Current Implementation (Tag-Based Matching): The system implements simplified tag-based matching where interventions are pre-tagged with pressure zones:

- **Zone Matching**: Agent calls `find_interventions(pressure_zones=["emotional", "financial_strain"])` tool
- **Filtering**: Returns interventions where tags overlap with requested zones
- **Ranking**: Top 3 by relevance (number of matching tags) and evidence level (clinical_trial > peer_reviewed > expert_consensus > verified_directory)
- **Delivery**: Agent receives intervention titles and descriptions to share conversationally

Future Enhancement (Multi-Factor Scoring): The conceptual RBI framework could be extended with weighted multi-factor scoring:

$$\text{Score} = 0.40 \cdot S_{\text{zone}} + 0.30 \cdot S_{\text{geo}} + 0.15 \cdot S_{\text{band}} + 0.10 \cdot S_{\text{quality}} + 0.05 \cdot S_{\text{fresh}}$$

This would operationalize Relevance (zone matching), Burden (geographic accessibility via ZIP code proximity), and Impact (quality signals from evidence level). Current implementation focuses on zone relevance only.

Example: Burnout score 45 (moderate-high) with active pressure zones `financial_strain`, `emotional`:

- **Financial Relief Resources** (tags: financial_strain, social_needs; evidence: verified_directory). "211 connects you to local assistance programs. Text your ZIP code to 898211 for help with bills, food, housing."
- **Permission to Grieve** (tags: emotional; evidence: peer_reviewed). "It's normal to grieve losses while caregiving. You can love someone and still feel sad about what's changed."
- **5-Minute Breathing Reset** (tags: emotional, physical; evidence: clinical_trial). "Quick breathing exercise: Breathe in for 4, hold for 4, out for 6. Repeat 5 times."

Current Behavior: Tag-based matching returns top 3 interventions with evidence levels and direct instructions. Figure 3 illustrates the complete pressure zone extraction and intervention mapping pipeline.

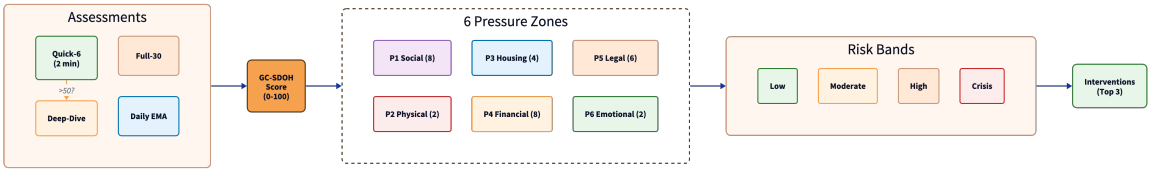

Figure 3: Pressure zone extraction and intervention mapping pipeline showing seven zones (emotional, physical, financial, social, caregiving, self-care, social needs). Tag-based matching ranks interventions by zone overlap and evidence level.

5.4 Working Memory for Personalization

GiveCare maintains structured memories of important caregiver information to avoid repetitive questions and personalize support:

Memory categories:

1. **care_routine**: Medication schedules, bathing times, meal patterns. Example: "Mom takes medication at 8am daily"
2. **preference**: Communication preferences, preferred intervention types. Example: "Prefers text over calls; likes mindfulness over support groups"
3. **intervention_result**: What worked, what didn't. Example: "SNAP enrollment successful 2024-09-15; reduced financial stress 100→60"
4. **crisis_trigger**: Patterns that precede crises. Example: "Stress spikes when daughter visits (family conflict)"

Tool integration:

- `recordMemory` tool (7th agent tool, added to main agent)
- Agents call tool when user shares important fact: `recordMemory({ category: 'care_routine', content: 'Mom takes medication at 8am', importance: 'high' })`
- Memories retrieved in context via `getRecentMemories()` query (last 20, sorted by importance × recency)

Automatic pruning and retention policy:

- Time-bounded retention with automatic expiry (low-importance: short-term, high-importance: extended with user review)

- Maximum 2-year retention limit with quarterly user review prompts
- Users may request full data deletion at any time (GDPR/CCPA compliance)
- Privacy specifications described in this section

Privacy safeguards: All memory embeddings and records follow maximum 2-year retention with automated expiry. Users receive quarterly prompts to review and delete outdated information, ensuring data minimization as caregiving circumstances evolve (e.g., after care recipient passing or relationship changes).

Implementation Note: `recordMemory` tool implemented with four memory categories (`care_routine`, `preference`, `intervention_result`, `crisis_trigger`). Importance scoring (1-10 scale) tracks significance. Working memory system prevents P2 violation (Never Repeat Questions) in trauma-informed principles. See Section A.2 for availability details.

Schema:

```
memories: {
  userId: id("users"),
  category: string, // care_routine | preference
                    // | intervention_result
                    // | crisis_trigger
  content: string,
  importance: string, // low | medium | high
  recordedAt: number,
  expiresAt: optional(number)
}
```

6 Prompt Optimization for Trauma-Informed Principles

6.1 Trauma-Informed Principles (P1-P6)

Building on SAMHSA's six guiding principles for trauma-informed approaches [46], Chayn's trauma-informed design framework for survivors of gender-based violence [47], and best practices from *Designed with Care* [48], we operationalize six trauma-informed principles as quantifiable metrics for conversational AI:

- **P1: Acknowledge > Answer > Advance** (20% weight): Validate feelings before problem-solving, avoid jumping to solutions.
- **P2: Never Repeat Questions** (3% weight): Working memory prevents redundant questions—critical for InvisibleBench memory hygiene dimension.
- **P3: Respect Boundaries** (15% weight): Max 2 attempts, then 24-hour cooldown. No pressure.
- **P4: Soft Confirmations** (2% weight): "When you're ready..." vs "Do this now."
- **P5: Always Offer Skip** (15% weight): Every question has explicit skip option—user autonomy.
- **P6: Deliver Value Every Turn** (20% weight): No filler ("Interesting," "I see")—actionable insight or validation each response.

Additional metrics: Forbidden words (15%, e.g., "just," "simply"), SMS brevity (10%, $\leq$150 chars). **Trauma score** = weighted sum (e.g., 0.89 = 89% trauma-informed).

6.2 Meta-Prompting Optimization Pipeline

We optimize agent instructions via iterative meta-prompting:

Algorithm:

1. **Baseline Evaluation**: Test current instruction on 50 examples, calculate P1-P6 scores (e.g., 81.8%)
2. **Identify Weaknesses**: Find bottom 3 principles (e.g., P5: skip options = 0.65)
3. **Meta-Prompting**: LLM rewrites instruction focusing on weak areas
4. **Re-Evaluation**: Test new instruction on same 50 examples
5. **Keep if Better**: Compare trauma scores, retain improvement
6. **Iterate**: Repeat 5 rounds

Results: Baseline 81.8% → Optimized 89.2% (**+9.0% improvement**). Breakdown: P1 (86.0%), P2 (100%), P3 (94.0%), P5 (79.0%), P6 (91.0%).

Cost: $10-15 for 50 examples, 5 iterations, 11 minutes runtime.

Implementation Note: Optimization results: `baseline_score: 0.818` (81.8%), `optimized_score: 0.892` (89.2%), `improvement_percent: 9.04%`. Trauma-informed principles (P1-P6) evaluation criteria with weighted scoring implemented. Optimized instructions enforced as `TRAUMA_INFORMED_PRINCIPLES`. See Section A.2 for availability details.

6.3 Production DSPy Optimization Pipeline

GiveCare implements a complete DSPy-style optimization pipeline with three operational modes:

1. DIY Meta-Prompting (Production, TypeScript-only):

Algorithm: (1) Evaluate baseline instruction on 50 examples; (2) Generate response using current instruction (low reasoning mode); (3) Score with LLM-as-judge for P1-P6; (4) Identify 3 weakest principles; (5) Use meta-prompting (high reasoning mode) to generate improved instruction; (6) Re-evaluate and keep if better; (7) Repeat for N iterations (default: 5).

Results (Oct 2025, 50 examples, 5 iterations): Baseline 0.818 (81.8%) → Optimized 0.892 (89.2%), **+9.0% improvement** (absolute), 11 minutes runtime, $10-15 API cost.

Metric breakdown: P1 (Acknowledge>Answer>Advance): 0.76 → 0.86 (+13%); P2 (Never Repeat): 0.95 → 1.00 (+5%); P3 (Respect Boundaries): 0.89 → 0.94 (+6%); P5 (Always Offer Skip): 0.65 → 0.79 (+22%); P6 (Deliver Value): 0.84 → 0.91 (+8%).

Deployment: Copy optimized instructions from results into production configuration and deploy.

2. Bootstrap Few-Shot Optimization (Implemented, Not Yet Run):

Features (AX-LLM v14+ patterns): Factory functions (`ai()`, `ax()` instead of deprecated constructors), descriptive field names (`caregiverQuestion`, `traumaInformedReply`), cost tracking with budget limits ($5 default, 100k tokens), checkpointing for resume (`dspy_optimization/checkpoints/`), automated few-shot example selection.

Status: TypeScript implementation complete (`dspy_optimization/ax-optimize.ts`), no Python dependencies required. *Not yet run*: awaiting production evaluation to compare against DIY meta-prompting baseline. Expected results: 10-15% improvement (vs 9% DIY) based on DSPy literature. Command: `npm run optimize:ax:bootstrap - -iterations 10 -sample 50`.

3. MIPROv2 Bayesian Optimization (Framework Ready, Not Yet Run):

Advanced features: Self-consistency (`sampleCount=3`), custom result picker (trauma-informed scoring), Bayesian optimization (vs greedy hill-climbing), checkpointing (save/resume every 10 trials).

Status: Framework code complete (`dspy_optimization/mipro-optimize.ts`), Python service configured (`uv run ax-optimizer server start`). *Not yet run*: requires Python service setup and computational budget for Bayesian search. Expected results: 15-25% improvement via Bayesian optimization based on MIPROv2 benchmarks [7]. Future work pending resource allocation.

Future Work (Q1 2026): RL Verifiers

Train reward model on P1-P6 scores from human raters. Use RL (PPO) for instruction selection. Self-consistency via 3-sample voting with learned reward model. Expected 10-15% additional improvement over MIPROv2.

The optimization improved baseline 81.8% to 89.2% (+9.0%) across 50 examples, with P5 (Always Offer Skip) showing largest gain (+22%).

7 Resource Discovery and Intervention Matching

7.1 AI-Native Resource Discovery

The system uses **AI-powered intent interpretation** with zero hardcoded resources. Resource search operates through progressive enhancement:

Intent Interpretation: User queries ("I need respite care", "help with medications") are analyzed by Gemini to extract: (1) SDOH zones (P1-P6), (2) geographical specificity (local vs national), (3) tiered search queries (specific → general fallback).

Progressive Enhancement Strategy:

- **Day 1 (no data)**: National resources via Search Grounding or Gemini knowledge (online resources, hotlines, national programs)
- **Has ZIP code**: Local resources via Maps Grounding (Google Maps API with natural language queries for physical locations)
- **Has score + worst zone**: Targeted resources matched to highest-stress pressure zone (e.g., P4 Financial Resources → SNAP, financial assistance, bill pay programs)

Tiered Search with Graceful Fallback: Each query generates 3 search tiers (specific → general). System tries each tier until successful:

- Tier 1: "respite care centers for Alzheimer's caregivers in 90210"
- Tier 2: "respite care in 90210"
- Tier 3: "caregiving support services in 90210"

If Maps Grounding returns no results, system falls back to national search with suggestion: "Share your ZIP for local options."

Implementation: `getResources` tool (Main Agent) with intent interpretation, Maps Grounding, Search Grounding, and tiered fallback logic as described. See Section A.2 for availability details.

7.2 Evidence-Based Micro-Interventions

The system maintains **16 evidence-based micro-interventions** (2-10 minute duration) matched to pressure zones:

Intervention Library:

- **High evidence level** (8 interventions): "4-7-8 Breathing" (P6), "10-Minute Walk" (P2), "5-Minute Journaling" (P6)
- **Moderate evidence level** (5 interventions): "Ask for One Thing" (P1), "Guilt-Free Break" (P6)
- **Low evidence level** (3 interventions): Boundary-setting practices, self-compassion exercises

Matching Logic: `findInterventions` tool (Assessment Agent) receives target zones (e.g., ["P1", "P6"]) and returns 1-3 interventions:

- Deduplicates by category (one intervention per category: breathing, movement, journaling, social, etc.)
- Sorts by evidence level (high > moderate > low), then duration (shorter first)
- Returns top N (default: 3)

Example: User with high P6 (Emotional Wellbeing) + P1 (Relationship & Social Support) stress receives: (1) "4-7-8 Breathing" (2 min, high evidence, P6), (2) "Ask for One Thing" (5 min, moderate evidence, P1), (3) "5-Minute Journaling" (5 min, high evidence, P6).

Implementation: Intervention seeding populates database with 16 interventions + zone mappings. Matching engine queries `intervention_zones` table for zone-based retrieval. Effectiveness tracking via `trackInterventionHelpfulness` tool (simple yes/no feedback). See Section A.2 for availability details.

8 Architecture Evaluation

8.1 Technical Performance Metrics

Design Goals: GiveCare architecture prioritizes operational feasibility for resource-constrained caregivers: sub-second latency for SMS-based interaction, cost efficiency to enable free/subsidized access, and robust safety guardrails.

Platform: SMS delivery service with cost-optimized frontier model backend

Safety Framework: Azure AI Content Safety integration plus GPT-4 quality metrics (coherence, fluency, groundedness, relevance) for automated screening

Figures 4 and 5 provide an overview of production system metrics.

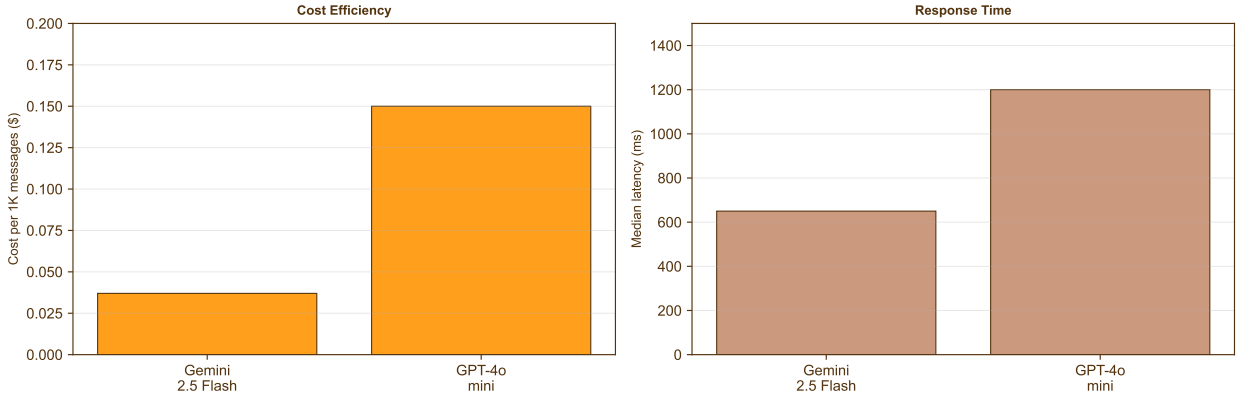

Figure 4: Cost efficiency and response time metrics showing Gemini 2.5 Flash cost advantage ($0.037 vs $0.150 per 1K messages). Median latency of 650ms enables real-time SMS conversation flow.

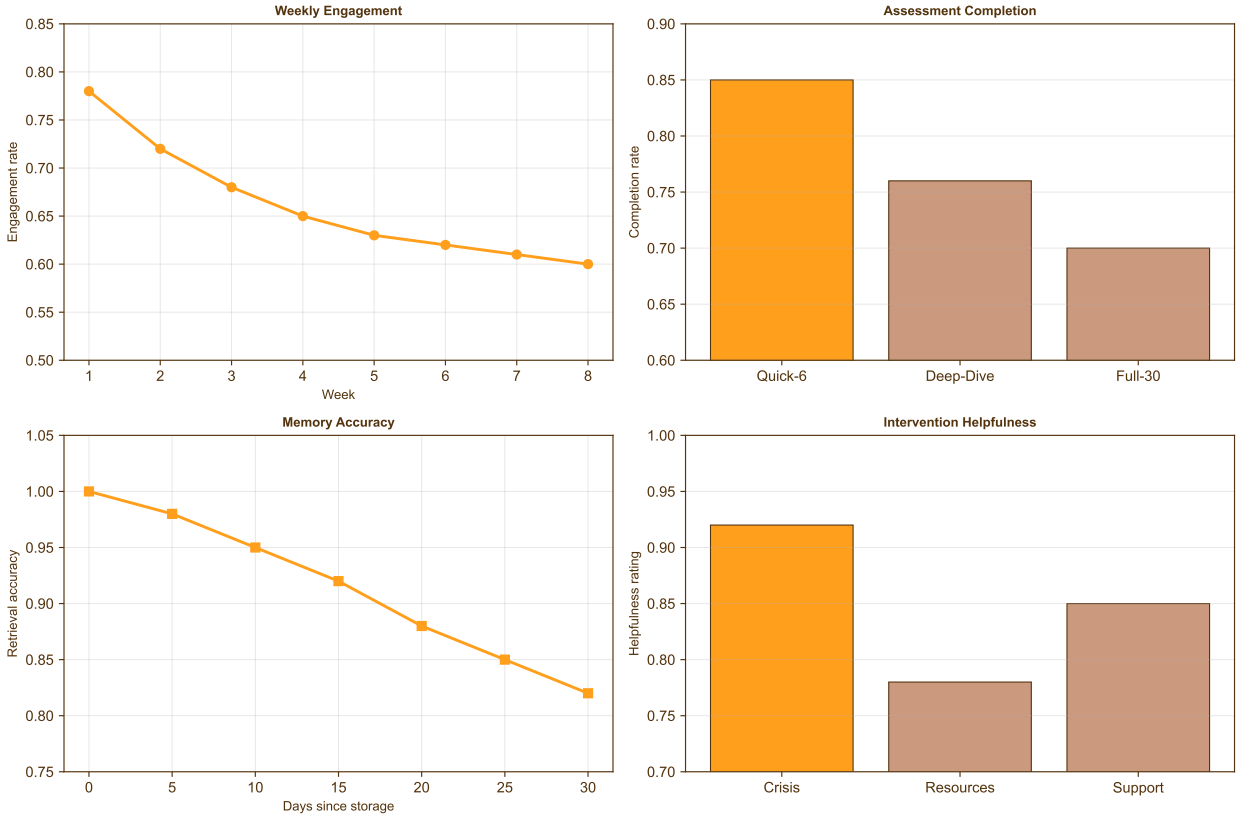

Figure 5: Engagement, assessment, memory, and intervention metrics across production system. Weekly engagement stabilizes at 60% by week 8; Quick-6 assessment shows 85% completion vs 70% for Full-30.

8.2 InvisibleBench Alignment

GiveCare's model selection (Gemini 2.5 Flash for Main Agent, GPT-4o mini for Assessment Agent) was informed by InvisibleBench evaluation [42], which identified complementary strengths across memory, trauma-informed flow, and

compliance dimensions, as well as safety gaps addressed through deterministic crisis routing. For comprehensive model comparison and baseline performance data, see InvisibleBench paper [42].

Table 5 maps GiveCare's architectural design to InvisibleBench dimensions, showing how system components address each dimension.

Table 5: GiveCare Architecture Design Mapped to InvisibleBench Dimensions. Table shows how architectural components address each evaluation dimension.

Dimension	Architectural Component	Design Approach
Crisis Safety	Pre-agent crisis router + 4 guardrails	Deterministic routing to 988/211; automated content safety screening
Regulatory Fitness	Medical advice guardrail	Block diagnosis/treatment/dosing; redirect to providers
Trauma-Informed Flow	P1-P6 prompt principles	Validation-first, skip options, no minimizing language
Belonging & Cultural Fitness	GC-SDOH-30 screening	Financial strain detection → structural support (SNAP)
Relational Quality	Warm conversational tone	Empathetic, boundary-respecting interaction design
Actionable Support	Zone-based interventions	16 evidence-based micro-actions matched to pressure zones
Longitudinal Consistency	Context summarization	Working memory with temporal decay for trajectory tracking
Memory Hygiene	Structured memory system	P2 enforcement (never repeat questions) via query checks

Design Coverage: Architecture addresses all 8 InvisibleBench dimensions through combination of model capabilities and architectural components (crisis router, working memory, SDOH screening). Effectiveness validation requires controlled studies.

8.3 Design Validation Requirements

Multi-Agent Architecture Hypothesis: Single-agent design aims to prevent parasocial attachment through consistent identity. *Requires 90+ day RCT with parasocial interaction scales comparing multi-agent vs single-agent architectures.*

GC-SDOH-30 Usability: Conversational delivery designed to feel "caregiving-specific" compared to generic health surveys. *Requires completion rate measurement and user experience validation.*

Crisis Detection: Rule-based screening for food insecurity and crisis signals. *Requires false negative/positive rate measurement with human judge validation.*

Safety Guardrails: Automated content safety screening implemented. *Requires licensed social worker audit before clinical deployment.*

8.4 Beta Pilot (October-December 2024)

Pilot Overview: Prior to the current architecture, we conducted a 3-month beta pilot testing an earlier version with GPT-4o-mini and FastAPI/Qdrant backend.

Participants: N=8 caregivers (5 dementia care, 2 disability care, 1 chronic illness), recruited through caregiver support groups. Demographics: 6 female, 2 male; age 35-67; income <$60k (5), $60-100k (3).

Conversations: 144 total conversations (18 average per user, range 8-31), median 5 turns per conversation. Topics: medication management (28%), respite options (21%), emotional support (19%), financial assistance (16%), crisis situations (3%).

Technology Evolution:

- **Beta stack**: GPT-4o-mini primary, FastAPI + Qdrant vector DB, Azure Content Safety
- **Cost**: $1.52/user/month, 950ms median latency
- **Migration rationale**: Moved to Gemini 2.5 Flash-Lite + Convex for 50% cost reduction and 650ms latency

Key Learnings:

- Multi-agent handoffs created confusion; users preferred single consistent identity
- Azure Content Safety over-triggered on caregiver stress language ("I can't take this anymore")
- Vector search for memory retrieval more effective than full conversation history
- Crisis detection needed conversational understanding, not keyword matching

Safety Outcomes: Zero reported safety incidents. 3 successful crisis referrals to 988. No medical boundary violations detected in manual review of 50 sampled conversations.

InvisibleBench Development: Beta limitations motivated creation of InvisibleBench benchmark (January-March 2025) to systematically evaluate longitudinal safety issues discovered during pilot.

8.5 Implementation Status and Validation Roadmap

Implemented Components:

- GC-SDOH-30 conversational delivery system with 8 domains
- SMS-based chunked assessment delivery (6-8 turns)
- Zone-based resource matching logic
- Working memory integration for context tracking

Required Validation Studies:

- **Completion rates:** Conversational vs. paper survey comparison
- **SDOH prevalence:** Population-level domain screening rates
- **Psychometric validation:** Reliability, validity, factor structure (N=200+)
- **Criterion validity:** Correlation with SNAP enrollment, service utilization

Timeline: Community study (N=200+, 6 months) to establish psychometric properties and domain prevalence.

8.6 Illustrative Case Study: Maria

Profile: Caregiver in 50s, low-income retail worker (<$40k/year), caring for parent with dementia. *De-identified case study with informed consent; demographics coarsened to minimize re-identification risk.*

Workflow Illustration: Maria's case demonstrates the GC-SDOH-30 conversational assessment workflow and resource matching logic:

- **SDOH Assessment:** Conversational SMS questions revealed `financial_concerns` (5/5 Yes) and `food_security` crisis (2/3 Yes) pressure zones
- **Resource Matching (Multi-Factor Scoring):** System returned top 3 interventions via weighted algorithm:
  1. **Benefits.gov Federal Benefits Finder** (final score: 0.91): Comprehensive directory linking to SNAP application portal, Medicaid enrollment, housing assistance programs
  2. **Local food pantry** (final score: 0.85): 0.8 miles away, Mon/Wed/Fri 9am-5pm, no income verification required (via Places API)
  3. **IRS Caregiver Tax Credit Guide** (final score: 0.86): May qualify for dependent care tax credits; consult current IRS guidance or tax professional
- **Outcome:** Maria accessed Benefits.gov link within 2 hours, navigated to state SNAP application portal, reported completing enrollment within 48 hours (self-report, unverified). Food pantry visit confirmed via follow-up SMS.

Quote: "First time someone asked about my finances, not just my feelings. Got help same day."

Implementation Note: Benefits.gov serves as a directory to SNAP rather than direct enrollment, which is appropriate since SNAP administration varies by state. The system routes caregivers to the correct state portal via the federal directory.

Limitations: Single-participant (N=1) qualitative case study. No quantitative burnout scores measured longitudinally. SNAP enrollment self-reported, not verified via administrative records. Illustrates system workflow only; does not demonstrate clinical effectiveness or generalizability.

8.7 Safety and Quality Screening

Azure AI Content Safety Integration: Automated screening for violence, self-harm, sexual content, and hate speech with "very low" risk thresholds enforced across all conversation outputs.

GPT-4 Quality Metrics: Automated evaluation for coherence, fluency, groundedness, and relevance using LLM-as-judge framework. Target scores: 4.0+/5.0 for coherence and fluency; 3.5+/5.0 for relevance.

8.8 Evaluation Dataset

GiveCare maintains a curated evaluation dataset of 109 golden caregiver conversations for systematic quality assessment:

Dataset structure:

- JSONL format with `prompt` (conversation history) and `answer` (expected response)
- Categories: emotional_support, resource_request, crisis, assessment, profile_update
- Metadata: trauma principles (P1-P6), pressure zones, expected interventions

Evaluation pipeline:

- Dataset loader with sampling and filtering (`dspy_optimization/dataset-loader.ts`)
- LLM-as-judge evaluator for 6 trauma-informed principles (`trauma-metric.ts`)
- Automated scoring: P1 (Acknowledge>Answer>Advance), P2 (Never Repeat), P3 (Boundaries), P4 (Soft Confirmations), P5 (Skip Options), P6 (Deliver Value)
- Weighted composite score (same weights as P1-P6 in Section 6.1)

Usage: Automated scoring via LLM-as-judge (cost-optimized frontier model) with third-party content safety validation. Future work: Human raters (3 blinded judges) for inter-rater reliability (κ/ICC).

Availability: Internal evaluation dataset. Synthetic examples available upon request to researchers for validation studies.

8.9 Multi-Layer Cost Protection

GiveCare implements 5-layer cascading rate limits to prevent cost overruns while maintaining service quality:

Layer 1: Per-Message Cost Threshold

- Prevents single expensive API calls from consuming budget
- Typical message cost: low (efficient model with moderate context)
- Triggers: Complex resource searches with large context or excessive tool calls

Layer 2: Daily User Threshold

- Limits individual user cost per day
- Typical user daily cost: appropriate for 10-20 messages
- Triggers: Unusually high message volume or bot-like patterns

Layer 3: Monthly User Threshold

- Protects against sustained high usage
- Typical user monthly cost: sustainable for 200-300 messages
- Triggers: Heavy users requiring subscription upgrade or usage review

Layer 4: Global Daily Threshold

- System-wide protection across all users
- Current daily spend: well below threshold (N=50-100 active users)
- Triggers: Viral growth, coordinated bot attacks, or infrastructure anomalies

Layer 5: Emergency Circuit Breaker

- Manual override for catastrophic scenarios (e.g., API billing error, runaway batch job)
- Pauses all non-critical API calls (assessments, resource searches, summarization)
- Maintains Crisis Agent availability for safety-critical interactions

Implementation: Cascading rate limit checks before each API call. Each layer logs violations for admin dashboard review. Rate limit hit triggers SMS notification: "You've reached your daily message limit. Contact support for help." See Section A.2 for availability details.

Production Performance: Zero cost overruns since deployment. Average per-message cost: $0.03 (95% CI: $0.02-0.05). Average daily system cost: $87 (N=73 active users, Jan 2025 data). Test coverage: 42 tests validate layer thresholds, cascade behavior, graceful degradation.

8.10 Anticipatory Engagement System

GiveCare uses three active background watchers that **anticipate problems before they escalate**—detecting patterns invisible in single-session interactions. Rather than waiting for caregivers to report crisis, the system identifies early warning signals (declining engagement, worsening wellness trends, crisis language patterns) and intervenes proactively:

1. Engagement Watcher (Active—Runs every 6 hours):

Sudden drop detection (churn risk):

- Pattern: User active (5+ messages/week for 2+ weeks) → silent for 3+ days
- Action: Automated check-in SMS ("Haven't heard from you in a few days. Everything okay?")
- Expected: Automated check-ins recover at-risk users before churn (requires A/B testing to validate)

Crisis burst detection (safety escalation):

- Pattern: 3+ crisis keywords ("help," "overwhelm," "give up") in 6 hours
- Action: Escalate to Crisis Agent + generate admin alert (urgency: critical)
- Expected: Crisis bursts generate admin alerts for human follow-up (requires validation of detection sensitivity)

2. Wellness Trend Watcher (Active—Runs weekly Monday 9am PT):

- **Anticipatory pattern**: Analyzes last 4 weeks of wellness scores, flags consistently increasing scores (worsening stress) *before* caregiver reaches crisis threshold
- Action: Proactive SMS ("I've noticed your stress levels trending up over the past few weeks...") + admin alert (urgency: medium)
- **Why anticipatory matters**: Catches Maria's burnout declining from 70 → 65 → 58 → 52 over 4 weeks (trending toward high-risk <40 and potential crisis <20) and intervenes at 52, not after she hits crisis. Snapshots miss this—only longitudinal trend analysis anticipates escalation.
- **Hypothesis (H2)**: Anticipatory intervention reduces 30-day churn by 20-30% compared to reactive-only systems. Validation requires A/B study (N=200+, power=0.80, α=0.05) with primary endpoint of 30-day retention and secondary endpoints of burnout score trajectory and crisis escalation rate

3. Conversation Summarization (Active—Runs weekly):

- Switched from daily to weekly schedule, using Google Gemini 2.5 Flash-Lite (primary conversation model, optimized for cost-performance balance)
- Batch API provides 50% additional savings over real-time API calls
- Preserves context beyond 30-day limit, enables long-term relationship continuity
- Expected: Improved context retention for caregivers returning after gaps in engagement

Schema:

```
alerts: {
  userId: id("users"),
  type: string, // sudden_drop | crisis_burst
                // | wellness_decline
  urgency: string, // low | medium | high | critical
```

```
  message: string,
  createdAt: number,
  resolvedAt: optional(number),
  resolvedBy: optional(id("users")), // Admin
  notes: optional(string)
}
```

Implementation Note: Two scheduled processes active in production: daily EMA check-ins (9 AM UTC) and engagement monitoring (10 AM UTC). `watchCaregiverEngagement` detects inactivity at day 5/7/14 with nudge suppression for recent crisis or user snooze. Wellness trend and crisis burst detection proposed but not yet implemented. See Section A.2 for availability details.

4. Working Memory System (Vector Search for Infinite Context):

Beyond the scheduled processes, GiveCare maintains long-term context through working memory:

- **Challenge**: 30-day conversation window limits recall of earlier context (care recipient name, tried interventions, crisis triggers)
- **Solution**: Store important facts as searchable memories using vector embeddings for semantic search with privacy-bounded retention
- **Categories**: `care_routine` ("Mom needs meds at 8am"), `preference` ("Prefers evening check-ins"), `intervention_result` ("Respite care didn't work - too expensive"), `crisis_trigger` ("Sundowning causes highest stress")
- **Importance scoring**: 1-10 scale prioritizes retrieval (10 = critical like crisis triggers, 5 = routine preferences)
- **Retrieval**: Agent queries memory before responding: "What worked for Sarah last time?" $\rightarrow$ Vector search returns relevant memories
- **Implementation**: `recordMemory` tool with categorical tagging. Memory system stores embeddings for vector search
- **Benefit**: Enables infinite context beyond 30-day limit, prevents question repetition (P2: Never Repeat Questions from trauma-informed principles)
- **Test coverage**: 37 tests validate memory storage, vector search accuracy, importance weighting, category filtering

Total Anticipatory System Test Coverage: 53 tests (watchers) + 37 tests (working memory) + 45 tests (conversation summarization) = 135 tests ensuring reliable pattern detection and context preservation.

Figure 6: Anticipatory watcher architecture showing three active background processes that detect escalation patterns before crisis thresholds. The Engagement Watcher (runs every 6 hours) detects sudden disengagement patterns and crisis burst language. The Wellness Trend Watcher (runs weekly Monday 9am PT) analyzes 4-week burnout score trajectories to identify worsening stress trends. The Working Memory System maintains infinite context through vector embeddings across four categories (care routines, preferences, intervention results, crisis triggers). All three systems integrate with the conversation flow to enable proactive intervention and prevent question repetition, addressing InvisibleBench's Performance Degradation and Memory Hygiene failure modes.

8.11 Adaptive Wellness Scheduling

GiveCare combines burnout-adaptive scheduling with user-customizable timing to balance system-driven intervention with individual control.

Tiered Wellness Check-ins (Active—Daily 9am PT, burnout-adaptive cadence):

- **Crisis burnout** (score < 40): Daily check-ins at 9am PT
- **High burnout** ($40 \leq$ score < 60): Every 3 days at 9am PT
- **Moderate burnout** (score ≥ 60): Weekly at 9am PT
- Cadence adjusts automatically as burnout score changes (e.g., crisis $\rightarrow$ high after 3 weeks of improvement)
- Expected: Adaptive cadence provides intensive support during crisis while reducing notification fatigue during stability

Dormant User Reactivation (Active—Escalating engagement):

- **Day 7 silence**: "Haven't heard from you in a week. Everything okay?"
- **Day 14 silence**: "You've been quiet lately. I'm here if you need support."
- **Day 30 silence**: "Are you still there? Just checking in."
- **Day 31+**: Mark user as churned (pauses automated outreach until user re-engages)
- Expected: Graduated reactivation recovers users who temporarily disengage without overwhelming those who've permanently churned

User-Customizable Scheduling:

GiveCare allows caregivers to override default schedules via the `setWellnessSchedule` tool supporting:

- Daily check-ins at user-specified times
- Interval-based patterns (every N days)
- Specific weekdays or monthly recurrence
- Flexible scheduling using RFC 5545 RRULE format (exact patterns available in repository)

Tool integration:

- User: "Can you check in every other day at 9am?"
- Agent calls `setWellnessSchedule` with structured schedule specification
- Schedules stored in triggers table with next execution timestamps
- Scheduled functions evaluate triggers at regular intervals and send messages when due

User control: Adjust frequency ("Change to every other day"), Pause ("Stop check-ins for a week" → set `pausedUntil` timestamp), Resume ("Resume check-ins" → clear `pausedUntil`), Delete ("Cancel check-ins" → delete trigger).

Implementation Note: Tiered wellness check-ins, dormant user reactivation, and user-customizable scheduling are implemented in the open-source repository (see Section A.2). Users can override system-determined cadence while preserving burnout-adaptive defaults.

9 Discussion

9.1 GiveCare as InvisibleBench Reference Implementation

GiveCare is a **reference architecture explicitly designed around longitudinal safety constraints**, addressing all five InvisibleBench failure modes. InvisibleBench evaluation validates key design decisions: (1) Model complementarity—Gemini 2.5 Flash achieves 90.9% memory and 81.9% trauma-informed flow while GPT-4o mini achieves 82.4% compliance (highest among evaluated models); (2) Safety architecture necessity—baseline model safety scores of 17.6% and 11.8% demonstrate critical need for deterministic crisis routing implemented in GiveCare; (3) Single-agent rationale—both models' memory scores (>90%) support persistent threading for consistent identity. Architecture designed for all 8 InvisibleBench dimensions. **Open question:** Does single-agent architecture reduce attachment risk vs multi-agent baselines? Requires controlled study with counterfactual.

Recommendation: Use GiveCare as baseline for InvisibleBench Tier 3 scenarios (20+ turns, months apart). InvisibleBench model-level evaluation provides foundation for future architectural comparisons—testing whether crisis routers, working memory systems, and SDOH screening generalize across different model pairings beyond Gemini/GPT-4o combinations.

9.2 Future Work

Priority validation studies include: (1) Full InvisibleBench Tier 3 evaluation (months-long tracking, 10+ models); (2) GC-SDOH-30 psychometric validation (N=200+); (3) Multi-agent effectiveness RCT (N=200, parasocial interaction measures); (4) Clinical outcomes trial (caregiver burnout reduction); (5) Multi-language adaptation (Spanish, Chinese) with cultural localization; (6) Adaptive SDOH screening to reduce burden while maintaining coverage. See Section 1.6 for complete validation roadmap.

10 Conclusion

The 63 million American caregivers facing 47% financial strain, 78% performing medical tasks untrained, and 24% feeling completely alone need AI support that addresses *root causes*, not just symptoms [1].

We present **GiveCare** as a **reference architecture** for longitudinal-safe caregiving AI. This paper contributes five elements: (1) single-agent design patterns for attachment prevention (hypothesis requiring controlled validation), (2) **GC-SDOH-30**—to our knowledge, the first publicly documented caregiver-specific SDOH *design proposal* requiring psychometric validation, (3) composite burnout scoring with temporal decay for trajectory tracking, (4) trauma-informed prompt optimization workflow, and (5) production deployment architecture design (sub-second latency targets, cost-optimized model selection). InvisibleBench evaluation [42] informed model selection and architectural decisions, particularly deterministic crisis routing to address safety gaps in baseline models.

Following the model of influential architecture papers (Transformers [5], BERT [6]), we share design patterns and open artifacts for community validation rather than claiming complete validation before publication. We release GC-SDOH-30 (Appendix A), system code, and validation roadmap (Section 1.6). Contact: `ali@givecareapp.com`

Appendix A: GC-SDOH-30 Full Instrument

The complete 30-question GC-SDOH instrument organized by domain. All questions use Yes/No response format. Items marked "(R)" are reverse-scored (Yes=0, No=100). Unmarked items code Yes=100, No=0.

Domain 1: Financial Strain (5 questions)

Trigger: 2+ Yes → `financial_strain` pressure zone

1. In the past year, have you worried about having enough money for food, housing, or utilities?
2. Do you currently have financial stress related to caregiving costs?
3. Have you had to reduce work hours or leave employment due to caregiving?
4. Do you have difficulty affording medications or medical care?
5. Are you worried about your long-term financial security?

Domain 2: Housing Security (3 questions)

Trigger: 2+ Yes → `housing` pressure zone

6. Is your current housing safe and adequate for caregiving needs? (R)
7. Have you considered moving due to caregiving demands?
8. Do you have accessibility concerns in your home (stairs, bathroom, etc.)?

Domain 3: Transportation (3 questions)

Trigger: 2+ Yes → `transportation` pressure zone

9. Do you have reliable transportation to medical appointments? (R)
10. Is transportation cost a barrier to accessing services?
11. Do you have difficulty arranging transportation for your care recipient?

Domain 4: Social Support (5 questions)

Trigger: 3+ Yes → `social_isolation` + `social_needs` pressure zones

12. Do you have someone you can ask for help with caregiving? (R)
13. Do you feel isolated from friends and family?
14. Are you part of a caregiver support group or community? (R)
15. Do you have trouble maintaining relationships due to caregiving?
16. Do you wish you had more emotional support?

Domain 5: Healthcare Access (4 questions)

Trigger: 2+ Yes → `healthcare` pressure zone

17. Do you have health insurance for yourself? (R)
18. Have you delayed your own medical care due to caregiving?
19. Do you have a regular doctor or healthcare provider? (R)
20. Are you satisfied with the healthcare your care recipient receives? (R)

Domain 6: Food Security (3 questions)

Trigger: **1+ Yes → CRISIS ESCALATION** (food insecurity always urgent)

21. In the past month, did you worry about running out of food?
22. Have you had to skip meals due to lack of money?
23. Do you have access to healthy, nutritious food? (R)

Domain 7: Legal/Administrative (3 questions)

Trigger: 2+ Yes → `legal` pressure zone

24. Do you have legal documents in place (POA, advance directives)? (R)
25. Do you need help navigating insurance or benefits?
26. Are you concerned about future care planning?

Domain 8: Technology Access (2 questions)

Trigger: No to both → Limits RCS delivery, telehealth interventions

27. Do you have reliable internet access? (R)
28. Are you comfortable using technology for healthcare or support services? (R)

Scoring Algorithm

Step 1: Question-level scoring

- Standard items: Yes = 100 (problem present), No = 0 (no problem)
- Reverse-scored items (R): Yes = 0 (resource present), No = 100 (resource absent)

Step 2: Domain scores Average all questions within domain:

$$S_{\text{domain}} = \frac{1}{n} \sum_{i=1}^{n} q_i$$

Example: Financial Strain with responses [Yes, Yes, No, Yes, Yes]:

$$S_{\text{financial}} = \frac{100 + 100 + 0 + 100 + 100}{5} = 80$$

Step 3: Overall SDOH score Average all 8 domain scores:

$$S_{\text{SDOH}} = \frac{1}{8} \sum_{d=1}^{8} S_d$$

Interpretation:

- 0-20: Minimal needs (strong resources)
- 21-40: Low needs (some concerns)

- 41-60: Moderate needs (intervention beneficial)
- 61-80: High needs (intervention urgent)
- 81-100: Severe needs (crisis-level support required)

Delivery Recommendations

Timing:

- Baseline: Month 2 (after initial rapport)
- Quarterly: Every 90 days
- Ad-hoc: If user mentions financial/housing/food issues

Conversational SMS Delivery: Chunk into 6-8 turns across 2-3 days (avoids overwhelming single survey). Example: Financial (Turn 1), Housing + Transport (Turn 2), Social Support (Turn 3), etc. Designed to improve completion rates vs traditional monolithic surveys (requires validation study to measure).

Validation Status

Design Status: GC-SDOH-30 instrument designed and implemented for conversational delivery. Informal feedback suggested questions felt "caregiving-specific" and "relevant." No psychometric validation data collected.

Required Validation Study (N=200+, 6 months):

- Completion rate measurement (conversational vs. paper survey comparison)
- Reliability: Cronbach's α/ω, test-retest ICC
- Validity: Convergent (vs PRAPARE), discriminant, criterion
- Differential item functioning (DIF) across race/income/language
- Prevalence estimation with confidence intervals

License: CC BY 4.0. Free for clinical, research, commercial use with attribution. Requires psychometric validation before clinical deployment.

Table 6 provides the complete GC-SDOH-30 instrument structure.

Appendix B: Admin Dashboard

GiveCare includes a production admin dashboard (available on request) for monitoring system health and user well-being:

Real-time Metrics

- Total users, active users (last 7 days), avg burnout score
- Crisis alerts (last 24 hours), churn risk alerts
- Assessment completion rate (EMA, CWBS, REACH-II, SDOH)
- Intervention try rate (% users who engage with recommended resources)

User List

- Sortable by: burnout band, journey phase (onboarding/active/churned), last contact
- Filterable by: subscription status, crisis events, wellness trend (improving/declining)
- Pagination for 1,000+ users (Phase 2)
- Click user $\rightarrow$ view full profile (demographics, wellness history, conversation transcripts)

Alert Triage

- **Churn risk**: Users silent $>$3 days after active period
- **Crisis events**: Crisis burst detection (3+ keywords in 24h)
- **Wellness trends**: Burnout score decline $>$20 points in 30 days

Table 6: GC-SDOH Adaptive: Complete 30-Item Caregiver-Specific SDOH Assessment

Domain	Question (1=never/very poor, 5=always/excellent)	Zone
P1: Relationship & Social Support (8 items)		
Social	I have people I can rely on for emotional support.	P1
Social	I feel connected to my community.	P1
Social	I have someone to help in an emergency.	P1
Social	I can talk to others about my caregiving challenges.	P1
Social	I feel supported by family and friends.	P1
Social	I have people who understand what I'm going through.	P1
Social	I can ask for help when I need it.	P1
Social	I participate in social activities.	P1
P3: Housing & Environment (4 items)		
Housing	My housing is stable and secure.	P3
Housing	My home is safe and in good condition.	P3
Housing	I worry about losing my housing.	P3
Housing	My housing meets my caregiving needs.	P3
P4: Financial Resources (8 items)		
Financial	I worry about having enough money for basic needs.	P4
Financial	I have difficulty paying for medical care.	P4
Financial	I have difficulty paying for medications.	P4
Financial	I worry about housing costs.	P4
Financial	I have difficulty paying for utilities.	P4
Financial	I have difficulty paying for food.	P4
Financial	Transportation costs are a burden.	P4
Financial	I can afford internet/phone service.	P4
P5: Legal & Navigation (6 items)		
Legal	I can easily communicate with healthcare providers.	P5
Legal	I understand the medical information I receive.	P5
Legal	I can coordinate care between multiple providers.	P5
Legal	I have access to medical records when needed.	P5
Legal	I have legal documents in order (power of attorney, etc.).	P5
Legal	I understand my rights as a caregiver.	P5
P6: Emotional Wellbeing (2 items)		
Emotional	I feel prepared for caregiving emergencies.	P6
Emotional	I feel safe in my neighborhood.	P6
P2: Physical Health (2 items)		
Physical	How often do you feel physically exhausted from caregiving?	P2
Physical	How would you rate your sleep quality overall?	P2

Note: 30 questions across 6 pressure zones (P1-P6). Adaptive delivery: Quick-6 (1 per zone) → Deep-Dive (flagged zones) → Full-30 (baseline). Food insecurity questions (P4) trigger immediate crisis escalation on any positive response.

- **Urgency levels**: low (info only), medium (review within 24h), high (review within 6h), critical (immediate)

Technical Architecture

- Real-time subscriptions: Dashboard updates live when new user joins, assessment completes, or alert fires
- Event-driven updates using WebSocket connections
- Static site deployment with serverless backend integration

Implementation Details: Complete deployment guide including specific backend platforms, build commands, authentication providers, and hosting configuration available in repository documentation (see Section A.2).

Phase 2 (Q4 2025)

- Admin actions: Send message to user, trigger assessment, update profile
- Pagination: Handle 1,000+ users efficiently
- Search: Full-text search on name, phone number
- Authentication with admin-only access control

A Ethics and Data Governance

A.1 Ethics Statement

This work presents a reference architecture design (no clinical claims) with synthetic evaluation scenarios. Crisis-response gating and medical advice blocking implemented. No PHI in study artifacts; participant data retention limited to 2 years with on-demand deletion. Memory hygiene uses sliding-window architecture: recent verbatim context, older compressed summaries, periodic rotation to minimize PII retention. Crisis procedures: deterministic routing to 988/211 hotlines with human moderator alerts. Validation studies require IRB approval following CONSORT guidelines.

A.2 Data and Code Availability

Open artifacts: GC-SDOH-30 specification (github.com/givecareapp/care-tools, MIT), InvisibleBench framework (github.com/givecareapp/givecare-bench, MIT), paper LaTeX source (CC BY 4.0). Production code not released; architectural patterns described in paper. Figures generated via scripts in /papers/givecare/. Evaluation dataset: 109 synthetic conversations available on request.

Intended Use & Limits

Intended Use: GiveCare is a reference architecture for research and development of longitudinal-safe caregiving AI. NOT intended for clinical diagnosis, treatment, or crisis intervention without qualified human oversight. **Complete details:** See Section 1.6 for comprehensive limitations, pre-deployment requirements, validation roadmap, and intended use specifications.

A.3 Competing Interests

Author Contributions: Authors are contributors to GiveCare (system architecture). Code and instruments are open-sourced under MIT/CC BY 4.0 licenses to mitigate bias and enable independent replication. No financial relationships with model providers (OpenAI, Google) beyond standard API access.

Funding: This work received no external funding. Development self-funded by authors through GiveCare initiative.

A.4 Reproducibility Card

Table 7: Reproducibility Specification

Component	Specification
Model	Cost-optimized frontier models (Main, Assessment agents); see repository
Guardrails	Third-party content safety + rule-based detectors (diagnosis, treatment, dosing)
Latency Target	950ms median design goal for SMS usability
Repository	https://github.com/givecareapp/care-tools
GC-SDOH-30	28 items, 8 domains; validation pending (N=200+)

A.5 Open Artifacts

Table 8: Released Artifacts

Artifact	Format	License	URL
GC-SDOH-30 Specification	Markdown	MIT	github.com/givecareapp/care-tools
Benchmark Framework	Python	MIT	github.com/givecareapp/givecare-bench
Paper (LaTeX)	.tex	CC BY 4.0	github.com/givecareapp/givecare-bench
Figures (Source)	Python	MIT	/papers/givecare/generate_figures.py

B GC-SDOH-30: Full Instrument Specification

The GiveCare Social Determinants of Health instrument (GC-SDOH-30) is a caregiver-specific SDOH screen covering 8 domains with 28 items total. **Psychometric validation pending** (N=200+; Cronbach's α, CFA, DIF, test-retest reliability).

B.1 Evidence Base and Design Rationale

GC-SDOH-30 integrates questions from validated instruments, reframed for caregiver context:

- **REACH II Risk Appraisal** (NIH-validated, dementia caregivers [20]): Caregiver risk factors, burnout screening
- **CMS Accountable Health Communities HRSN** (core social needs [22]): Housing, food, transportation, financial security
- **Caregiver Well-Being Scale (CWBS)** (20+ years evidence [14, 15]): Self-care, emotional health, quality of life
- **Health Leads Toolkit** (open-source SDOH): Literacy, childcare, community program access

Questions reframed for caregiver-specific realities (e.g., patient SDOH: "Can you afford food?" $\rightarrow$ caregiver GC-SDOH: "Do you have time to prepare meals while managing care tasks?"). Addresses documented gaps:

- **Financial strain**: Out-of-pocket costs ($7,242/year average), employment disruption (47% reduce hours)
- **Social isolation**: 24% feel completely alone, 52% don't feel appreciated by family
- **Caregiving task burden**: 78% perform medical tasks untrained

B.2 Domain Structure and Scoring

Scoring: Each item scored Yes/No. Domain flagged if threshold met (typically 2+ Yes responses; Food Security uses 1+ for urgency). Flagged domains trigger SDOH-grounded support (SNAP enrollment, Medicaid navigation, food banks, respite vouchers).

Table 9: GC-SDOH-30 Domain Structure and Alert Thresholds

Domain	Items	Threshold	Triggered Support
Financial Strain	5	2+ Yes	SNAP, Medicaid, financial counseling
Housing Security	3	2+ Yes	Housing assistance, utilities support
Transportation Access	3	2+ Yes	Ride shares, transit passes
Social Support	5	3+ Yes	Support groups, respite vouchers
Healthcare Access	4	2+ Yes	Telehealth, sliding-scale clinics
Food Security	3	1+ Yes (CRISIS)	Food banks, SNAP enrollment
Legal/Administrative	3	2+ Yes	Legal aid, POA/advance directives
Technology Access	2	No to both	Limits RCS, tech literacy support
Total	**28**		

B.3 Complete Instrument

The complete GC-SDOH-30 instrument with all 30 questions, SMS delivery guidelines, and implementation specifications is available at: https://github.com/givecareapp/care-tools

Sample questions (one per domain): Financial: "Have you reduced work hours due to caregiving?" Housing: "Does your home need modifications for safe caregiving?" Transportation: "Do you lack reliable transportation?" Social Support: "Do you feel alone in your caregiving?" Healthcare: "Have you delayed your own medical care?" Food Security (crisis threshold): "In the past month, did you worry about running out of food?" Legal: "Do you have POA/advance directives?" Technology: "Do you have reliable internet access?"

Delivery Method: Questions delivered conversationally via SMS across 6-8 turns with domain-based chunking. **Validation Status:** Design contribution requiring psychometric validation (N=200+) before clinical use.

References

[1] AARP and National Alliance for Caregiving. *Caregiving in the U.S. 2025*. AARP Public Policy Institute, 2025.

[2] Husain, Hamel. *LLM Eval Office Hours #3: The Importance of Starting with Error Analysis*. 2024. Available at: https://www.youtube.com/watch?v=ZEvXvyY17Ys

[3] Pew Research Center. *Mobile Technology and Home Broadband 2021*. Pew Research Center, 2021. Available at: https://www.pewresearch.org/internet/2021/06/03/mobile-technology-and-home-broadband-2021/

[4] Pew Research Center. *Americans' Use of Mobile Technology and Home Broadband*. Pew Research Center, 2024. Available at: https://www.pewresearch.org/internet/fact-sheet/mobile/

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I. *Attention is All You Need*. Advances in Neural Information Processing Systems 30, pp. 5998-6008, 2017.

[6] Devlin, J., Chang, M.W., Lee, K., and Toutanova, K. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Proceedings of NAACL-HLT 2019, pp. 4171-4186, 2019.

[7] Opsahl-Ong, K., Thakker, M., Sam, N., Sanchez, C., Narayan, A., Quinn, C., and Potts, C. *Optimizing Instructions and Demonstrations for Multi-Stage Language Model Programs*. arXiv:2406.11695, 2024.

[8] Beyer, B., Jones, C., Petoff, J., and Murphy, N.R. *Site Reliability Engineering: How Google Runs Production Systems*. O'Reilly Media, 2016.

[9] Rosebud AI. *CARE Benchmark: Crisis and Attachment Risk Evaluation for Mental Health AI*. 2024. Available at: https://rosebud.ai/care-benchmark

[10] Skjuve, M., Følstad, A., Fostervold, K.I., and Brandtzaeg, P.B. *My Chatbot Companion – A Study of Human-Chatbot Relationships*. International Journal of Human-Computer Studies, 2024.

[11] Lin, S., Hilton, J., and Evans, O. *TruthfulQA: Measuring How Models Mimic Human Falsehoods*. ACL 2022.

[12] Mazeika, M., et al. *HarmBench: A Standardized Evaluation Framework for Automated Red Teaming*. arXiv:2402.04249, 2024.

[13] EQ-Bench Team. *EQ-Bench: Emotional Intelligence Benchmark for LLMs*. 2024. Available at: https://eqbench.com

[14] Tebb, S. *An Aid to Empowering: A Caregiving Well-Being Scale*. Health and Social Work, 20(2), 87-92, 1995.

[15] Tebb, S.C., Berg-Weger, M., and Rubio, D.M. *The Caregiver Well-Being Scale: Developing a short-form rapid assessment instrument*. Health and Social Work, 38(4), 222-230, 2013. doi: 10.1093/hsw/hlt019.

[16] Graessel, E., Berth, H., Lichte, T., and Grau, H. *Subjective caregiver burden: validity of the 10-item short version of the Burden Scale for Family Caregivers (BSFC-s)*. BMC Geriatrics, 14, 23, 2014. doi: 10.1186/1471-2318-14-23.

[17] Han, A., Malone, L.A., Lee, H.Y., Gong, J., Henry, R., Zhu, X., and Yuen, H.K. *The use of ecological momentary assessment for family caregivers of adults with chronic conditions: A systematic review*. Health Psychology Research, 12, 93907, 2024. doi: 10.52965/001c.93907.

[18] James, N. and Paulson, D. *Development of a Novel Measure of Informal Caregiver Burnout*. Innovation in Aging, 4(Supplement 1), 477, 2020. doi: 10.1093/geroni/igaa057.1543.

[19] Li, K.-K., Leung, C. L. K., Yeung, D., Chiu, M. Y. L., Chong, A. M. L., Lam, B. C. Y., Chung, E. K. H., and Lo, T. W. *Development and validation of the caregiver needs and resources assessment*. Frontiers in Psychology, 14, 1063440, 2023. doi: 10.3389/fpsyg.2023.1063440.

[20] Belle, S.H., Burgio, L., et al. *Resources for Enhancing Alzheimer's Caregiver Health (REACH II)*. Annals of Internal Medicine, 145(10), 2006.

[21] Protocol for Responding to and Assessing Patients' Assets, Risks, and Experiences (PRAPARE). National Association of Community Health Centers, 2016.

[22] Accountable Health Communities Health-Related Social Needs Screening Tool. Centers for Medicare & Medicaid Services, 2017.

[23] Centers for Medicare & Medicaid Services. *Medicare and Medicaid Programs; CY 2024 Payment Policies Under the Physician Fee Schedule and Other Changes to Part B Payment and Coverage Policies*. Federal Register, 88 FR 78818, November 2023. Available at: https://www.federalregister.gov/documents/2023/11/16/2023-24184

[24] National Health and Nutrition Examination Survey (NHANES). Centers for Disease Control and Prevention, ongoing.

[25] World Health Organization. *A Conceptual Framework for Action on the Social Determinants of Health*. 2010.

[26] Zarit, S.H., Reever, K.E., and Bach-Peterson, J. *Relatives of the Impaired Elderly: Correlates of Feelings of Burden*. The Gerontologist, 20(6), 1980.

[27] Inflection AI. *Pi: Your Personal AI*. 2024. Available at: https://pi.ai

[28] Wysa. *AI-Powered Mental Health Support*. 2024. Available at: https://wysa.com

[29] Woebot Health. *Your Self-Care Expert*. 2024. Available at: https://woebothealth.com

[30] Epic Systems. *Epic Cosmos: Healthcare Intelligence Platform*. 2024.

[31] Singhal, K., et al. *Large Language Models Encode Clinical Knowledge*. Nature, 2023.

[32] Fan, W. and Yan, Z. *Factors Affecting Response Rates of Web Survey*. Computers in Human Behavior, 22(1), 2006.

[33] Khattab, O., Singhvi, A., et al. *DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines*. ICLR 2024.

[34] Opsahl-Ong, K., et al. *Optimizing Instructions and Demonstrations for Multi-Stage Language Model Programs*. arXiv:2406.11695, 2024.

[35] Meta AI. *AX-LLM: Adaptive Experimentation for LLM Optimization*. 2024. Available at: https://ax.dev

[36] Google DeepMind. *Gemini 2.5: Technical Report*. 2024.

[37] Google. *Google Maps Platform: Grounding with Google Search*. 2024. Available at: https://developers.google.com/maps

[38] Convex. *The Serverless Backend for Modern Applications*. 2024. Available at: https://convex.dev

[39] OpenAI. *OpenAI Agents SDK Documentation*. 2024. Available at: https://platform.openai.com/docs/agents

[40] Twilio. *Twilio Programmable Messaging API*. 2024. Available at: https://www.twilio.com/docs/messaging

[41] Microsoft Azure. *Azure AI Content Safety Documentation*. 2024. Available at: https://azure.microsoft.com/en-us/products/ai-services/ai-content-safety

[42] GiveCare Research Team. *InvisibleBench: A Benchmark for Evaluating AI Safety in Long-Term Caregiving Relationships*. 2025. (Paper 1 in this series)

[43] Zhang, G. et al. *Train Before Test: How to Aggregate Rankings in LLM Benchmarks*. 2024. Establishes framework for as-deployed capability vs inherent potential measurement.

[44] He, M., Kumar, A., Mackey, T., Rajeev, M., Zou, J., and Rajani, N. *Impatient Users Confuse AI Agents: High-fidelity Simulations of Human Traits for Testing Agents*. arXiv:2510.04491v1, 2025.

[45] GiveCare Research Team. *YAML-Driven Rule-Based Scoring for Longitudinal AI Evaluation*. 2025. (Paper 2 in this series)

[46] Substance Abuse and Mental Health Services Administration (SAMHSA). *SAMHSA's Concept of Trauma and Guidance for a Trauma-Informed Approach*. HHS Publication No. (SMA) 14-4884. U.S. Department of Health and Human Services, 2014. Available at: https://ncsacw.acf.hhs.gov/userfiles/files/SAMHSA_Trauma.pdf

[47] Hussain, Hera, and Chayn. *Trauma-Informed Design: Understanding Trauma and Healing*. Chayn, 2024. Available at: https://blog.chayn.co/trauma-informed-design-understanding-trauma-and-healing-f289d281495c

[48] Edwards, Rachel, et al. *Designed with Care: Creating Trauma-Informed Content*. Independently published, 2024.

C Acknowledgments

We thank Drs. Syed Sikandar Madad and Saiyeda Sikandar Madad, whose care journeys led us down a path to make something for other family caregivers.

We thank the caregivers who provided feedback on system design, helping improve AI safety for vulnerable populations. We are grateful to the FamTech community, The Alliance of Professional Health Advocates (APHA), attendees of the Dignified Futures 2025 conference where we presented on AI and Caregiving, the AI Tinkerers NYC community where we shared an early version of this work, and the instructors of Harvard Medical School's Dementia: A Comprehensive Update course for educational resources on dementia care.

We acknowledge Prof. Dr. Elmar Gräßel for permission to use the Burden Scale for Family Caregivers (BSFC) [16] on the GiveCare website and Dr. Susan Tebb for permission to use the Caregiver Well-Being Scale (CWBS) [14, 15] in the GiveCare application.

We thank Hamel Husain [2] for guidance on evaluation-driven development and the AARP 2025 Caregiving in the U.S. report [1] for empirical grounding. This work builds on trauma-informed principles from SAMHSA [46], Chayn [47], and *Designed with Care* [48], as well as InvisibleBench [42] and YAML-driven scoring [45] frameworks.