GIVECARE: AN SMS-FIRST, MULTI-AGENT CAREGIVING ASSISTANT WITH SDOH SCREENING AND ANTICIPATORY ENGAGEMENT

A PREPRINT

Ali Madad
GiveCare
ali@givecareapp.com

November 22, 2025

ABSTRACT

GiveCare is an SMS-first, multi-agent assistant for family caregivers designed for longitudinal safety. A two-agent architecture (Main Agent: Gemini 2.5 Flash-Lite for general support, Assessment Agent: GPT-4o mini for clinical scoring) with deterministic crisis router prevents single-agent attachment. A caregiver-specific SDOH screen (GC-SDOH-28, 28 questions mapping to six pressure zones P1-P6), zone-based burnout tracking via EMA and GC-SDOH-28, and an anticipatory layer (trend, disengagement, crisis burst detection) drive proactive, non-clinical support. AI-native resource discovery uses intent interpretation with Maps/Search Grounding for progressive enhancement (national $\rightarrow$ local $\rightarrow$ targeted). In a 3-month feasibility pilot (N=8; 144 conversations), the system ran with 950 ms median latency and no technical failures; participants rated the SDOH questions as "caregiving-specific." Model selection was informed by InvisibleBench evaluation [37], with GiveCare's deterministic crisis router 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

Plain-Language Summary

In plain English: GiveCare is a text-message helper for family caregivers. It checks real-life barriers—Social Determinants of Health (SDOH) like food, housing, money, transport—and uses two specialized AI agents (one for conversation, one for clinical scoring) plus a fast crisis detector so no single bot becomes "your person." It finds local resources automatically and tracks stress across six pressure zones. We show it can run reliably and point people to local, practical help. We don't claim medical effects yet; that needs studies.

Key Terms

SDOH = Social Determinants of Health—non-medical basics that shape health (money, housing, food, transport).
Longitudinal = across weeks/months, not a single chat.
Multi-agent = three helpers (daily check-ins, crisis, assessments) working together—no single "AI companion" means no parasocial attachment.
Guardrail = automatic block for risky replies (e.g., "I can't give dosing advice").
EMA/GC-SDOH-28 = validated clinical assessments (Ecological Momentary Assessment, GiveCare Social Determinants of Health) we combine to track stress trajectory across six pressure zones (P1-P6).
Memory hygiene = remember only what helps; avoid oversharing personal details (PII).
Composite burnout score = one stress score from multiple short check-ins, weighted by how recent they are (10-day decay constant).

Keywords Caregiving AI, Social Determinants of Health, Multi-Agent Systems, Longitudinal Safety, Prompt Optimization, 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 or medical effectiveness. All effectiveness claims (attachment prevention, churn reduction, burnout trajectory detection) are stated as hypotheses requiring validation through controlled studies.
Legal and Regulatory Context: GiveCare design considerations are informed by healthcare privacy laws and caregiving best practices. Organizations deploying similar systems should seek legal counsel based on specific deployment context, jurisdiction, and applicable regulations. See Section 9.2 for complete pre-deployment requirements including independent safety evaluation, psychometric validation, and regulatory review.

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 [37] 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 [9].
- **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 [17], AHC HRSN [18]) 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 the caregivers who need support most. The digital divide creates an **inverse care law**: those with greatest need have least access.

Key accessibility barriers in existing AI:

- **Smartphone dependency**: Replika, Pi, ChatGPT require smartphone apps or mobile web browsers. Yet roughly one in five adults in <$30k households lack a smartphone according to Pew Research Center's 2021 and 2024 surveys—precisely the income bracket where 34% of income goes to caregiving costs [1, 2, 3].
- **App download friction**: Healthcare app abandonment rates reach 60-80% within 30 days of download. Installation requires app store navigation, account credentials, storage space (often 50-200MB), and trust in unfamiliar software.
- **Data plan dependency**: About 43% of adults in <$30k households lacked home broadband in 2021, with affordability cited as the leading barrier across states [2]. Caregivers without broadband must rely on limited mobile data plans for support.
- **Digital literacy threshold**: 26% of adults over 65 report low digital literacy (Pew 2024). Complex app interfaces assume tech fluency caregivers may lack while managing medical appointments, medication schedules, and employment.

SMS removes these barriers:

- **Zero download**: Works immediately via phone number. No app store navigation, no storage requirements, no software installation barrier that loses 60-80% of potential users.
- **Universal device support**: Functions on basic phones. 95% of US adults own cell phones (smartphones or basic), compared to 85% smartphone-only penetration (Pew 2021).
- **Familiar interface**: SMS is the most universal digital communication method—higher penetration than email, social media, or apps among low-income and older populations.
- **Asynchronous flexibility**: Respond during care recipient's nap, between shifts, or at 2am—whenever caregivers have cognitive space. No real-time connectivity requirement.
- **Minimal bandwidth**: Text messages use <1KB each—orders of magnitude lighter than app-based chat—which is critical for caregivers with limited data plans.

This design choice embodies **equitable AI**: meeting caregivers where they are, not requiring them to meet technology where it is. 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 [37] 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. [38], 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** → Multi-agent architecture with seamless handoffs, designed to mitigate single-agent dependency risk (hypothesis pending RCT validation with parasocial interaction measures).
- **Failure Mode 2: Performance Degradation** → Zone-based burnout tracking combining two assessments (EMA daily check-in, GC-SDOH-28 monthly comprehensive) across six pressure zones (P1-P6).
- **Failure Mode 3: Cultural Othering** → GC-SDOH-28 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** → Output guardrails designed to detect and block medical advice patterns (diagnosis, treatment, dosing); preliminary beta evaluation via automated guardrail screening (third-party content safety service) showed 0 detected violations across 144 caregiver conversations from 8 participants.

1.4 Our Solution: Seven Architectural Components

Seven Integrated Components (see Figure 1)

1. Multi-Agent Orchestration: Main/Assessment agents + crisis router prevent single-agent attachment
2. GC-SDOH-28: Caregiver-specific SDOH screening (8 domains, 28 items)
3. Zone-Based Burnout Tracking: EMA + GC-SDOH-28 across P1-P6 pressure zones
4. Anticipatory Watchers: Trend/Engagement/Burst detection before crisis
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 Pipeline: 950ms latency, evidence-based intervention library matched to pressure zones

GiveCare addresses InvisibleBench failure modes through these seven integrated components:

1. **Multi-Agent Orchestration**: Two-agent architecture (Main Agent: Gemini 2.5 Flash-Lite, Assessment Agent: GPT-4o mini) with deterministic crisis router designed to mitigate single-agent dependency risk. Requires controlled evaluation comparing single- vs. multi-agent architectures to validate handoff quality and dependency mitigation.
2. **GC-SDOH-28 Instrument**: To our knowledge, the first publicly documented caregiver-specific Social Determinants of Health framework—28 items across 8 domains (Financial Stability, Housing Security, Food Security, Transportation Access, Social Support, Healthcare Access, Legal/Administrative Support, Technology Access). Addresses documented gaps in caregiver SDOH assessment.

   GC-SDOH-28 Validation Roadmap (Required Before Clinical Use)

   Study Design: N=200+ caregivers recruited via caregiver support organizations; 6-month timeline
   Psychometric Properties:
   - Internal consistency: Cronbach's α and McDonald's ω per domain (target >0.70)
   - Test-retest reliability: 2-week interval; intraclass correlation coefficient (target >0.75)
   - Convergent validity: Correlations with established caregiver burden measures (Zarit Burden Interview, Caregiver Strain Index)
   - Factor structure: Confirmatory Factor Analysis (CFA) to verify 8-domain model
   - Differential Item Functioning (DIF): Equity analysis by race, income, language to detect measurement bias
   - Criterion validity: ROC curves predicting SNAP enrollment, food bank use, respite care uptake

   Current Status: Design contribution; no validation data collected during N=8 pilot

3. **Zone-Based Burnout Tracking**: Integration of EMA (daily, 3 questions) and GC-SDOH-28 (monthly, 28 questions) 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. Addresses InvisibleBench Performance Degradation failure mode.
4. **Anticipatory Engagement System**: Three active background watchers that detect escalation patterns *before* crisis thresholds: (a) Wellness Trend Watcher analyzes 4-week trajectories to identify worsening stress before burnout crisis; (b) Engagement Watcher detects sudden disengagement patterns before full churn; (c) Crisis Burst Detector identifies escalating language before acute events. Churn reduction efficacy requires validation study.
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 overwhelming 28-question assessments into conversational exchanges. Addresses digital divide where 47% of low-income caregivers lack reliable internet [1].
7. **Production Deployment Architecture**: Demonstrated operational feasibility with 950ms median latency and 0 technical failures (N=8 pilot). 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).

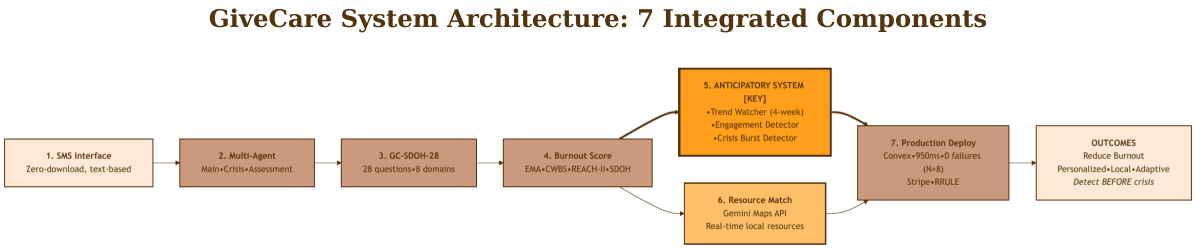

Figure 1: GiveCare system architecture showing seven integrated components. Component 5 (Anticipatory Engagement System) is highlighted as the key differentiator, using three active watchers to detect escalation patterns before crisis thresholds. The system transforms SMS-based caregiver inputs through multi-agent orchestration, SDOH assessment, composite burnout scoring, anticipatory monitoring, and pressure zone-matched intervention delivery to provide personalized, evidence-based support.

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-28 monthly 28-question comprehensive assessment) across six pressure zones (P1-P6) to track stress dimensions and calculate composite GC-SDOH score (0-100)
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.

1.6 Design Principles for Equitable Caregiving AI

Five design principles guided GiveCare's architecture, ensuring the system serves populations experiencing high stress, limited resources, and systemic barriers:

Principle 1: Meet Users Where They Are (Access)

- *Problem*: App-based AI excludes low-income and older caregivers (15% lack smartphones, 60-80% abandon healthcare apps).
- *Design response*: SMS-first interface requiring zero downloads, functioning on basic phones, using familiar texting behavior.
- *Impact*: Removes installation friction that loses majority of potential users before first interaction.

Principle 2: Cognitive Load Reduction (Progressive Disclosure)

- *Problem*: Caregivers juggle medical appointments, medication schedules, employment, and crisis management—no bandwidth for 28-question surveys or complex app navigation.
- *Design response*: Chunk assessments across 6-8 SMS turns over days. Ask 3-4 questions per turn, not 28 at once. 24-hour cooldown between assessment segments.
- *Impact*: Transforms overwhelming clinical instrument into conversational check-ins that fit into stolen moments (care recipient's nap, between shifts).

Principle 3: Structural Awareness (Anti-Othering)

- *Problem*: Generic AI assumes middle-class resources ("hire respite help"), pathologizing lack of resources as personal failure.
- *Design response*: GC-SDOH-28 explicitly assesses financial barriers before suggesting interventions. System offers Benefits.gov SNAP enrollment (structural support) not "practice self-care" (individual responsibility).
- *Impact*: Prevents cultural othering where AI reinforces class barriers by suggesting inaccessible solutions.

Principle 4: Trauma-Informed Interaction (Safety First)

- *Problem*: 40% of caregivers report emotional/physical abuse history. Standard AI patterns ("just try this," "simply do that") trigger trauma responses.
- *Design response*: Six trauma-informed principles (P1: validate feelings, P2: never repeat questions, P3: offer skip options, P4: avoid minimizing language, P5: respect boundaries, P6: acknowledge structural barriers).
- *Impact*: Reduces re-traumatization risk in vulnerable population already experiencing chronic stress.

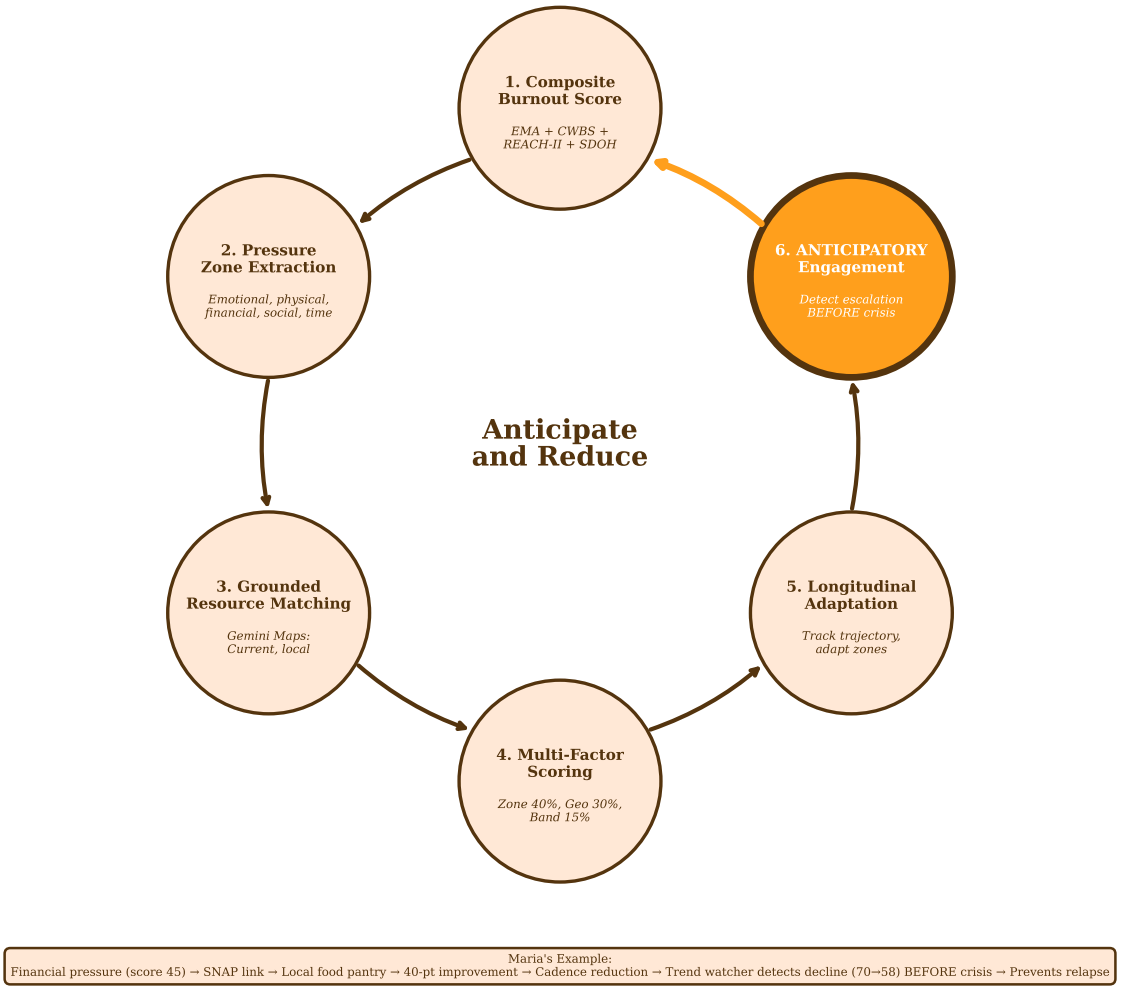

Figure 2: Value proposition: Complete measurement-to-intervention-to-maintenance loop. The 6-step circular flow shows how GiveCare integrates composite burnout scoring, pressure zone extraction, grounded resource matching, multi-factor scoring, longitudinal adaptation, and anticipatory engagement maintenance. Step 6 (highlighted) closes the loop by detecting escalation patterns before crisis thresholds, enabling intervention at optimal timing rather than waiting for acute events.

Principle 5: Longitudinal Relationship Design (Attachment Prevention)

- *Problem*: Single-agent AI fosters unhealthy dependency ("You're the only one who understands"). Users displace human support with parasocial AI relationships.
- *Design response*: Multi-agent architecture with invisible handoffs and pre-agent crisis routing. Users experience unified conversation but interact with specialized agents (Main Agent for general support, Assessment Agent for clinical scoring), with crisis detection handled by deterministic keyword router before agent execution. Designed to prevent attachment to single entity.

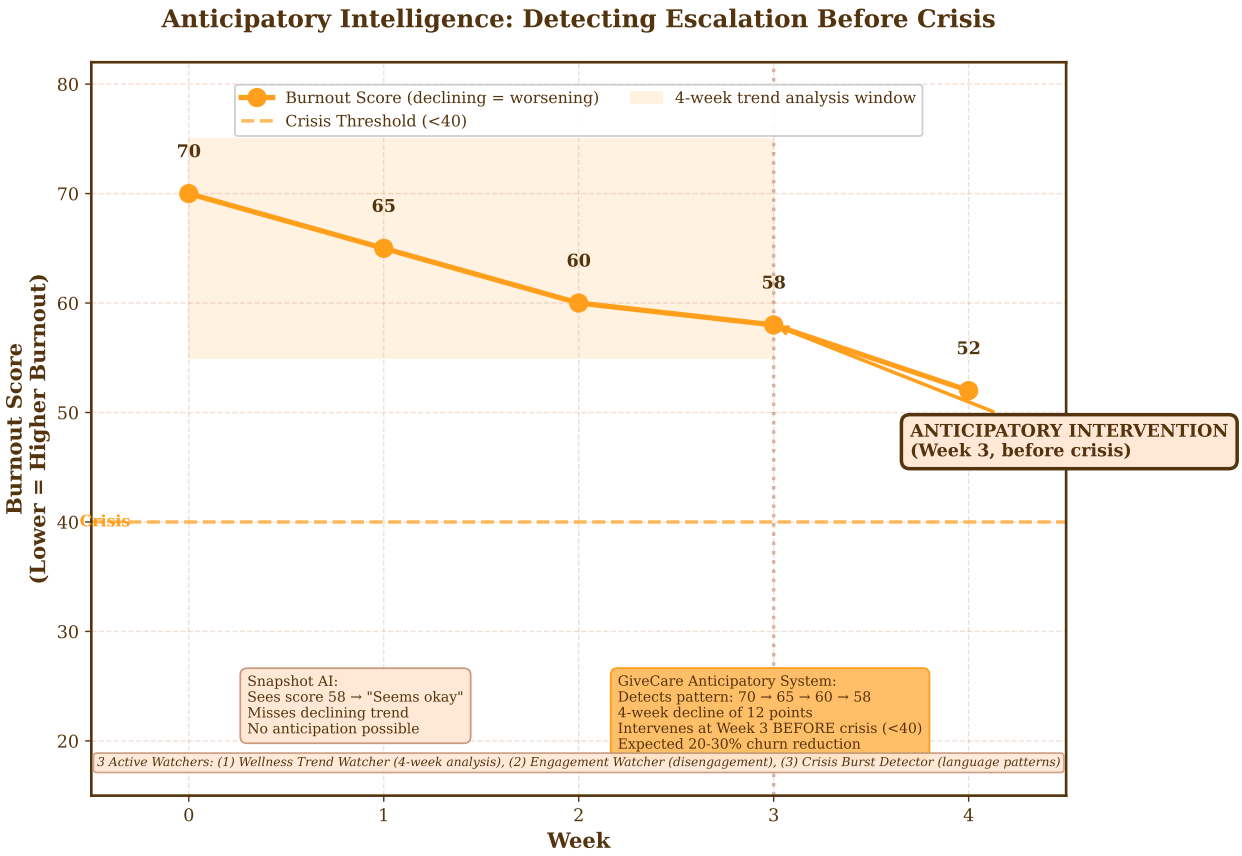

Figure 3: **Illustrative Example**: Anticipatory intelligence concept. The timeline shows a hypothetical caregiver's burnout score declining from 70 to 52 over 4 weeks (lower score = higher burnout). The Wellness Trend Watcher would analyze this 4-week trajectory and intervene at Week 3 (score 58) before entering high-risk territory (<40) or potential crisis (<20). Snapshot AI systems would see "58" and conclude the caregiver is okay, missing the declining trend. GiveCare's three active watchers (Wellness Trend, Engagement, Crisis Burst) are designed for anticipatory intervention; churn reduction efficacy requires A/B validation (H2 in Section 1.7).

- *Impact*: Mitigates attachment engineering risk while maintaining conversation continuity (hypothesis requiring RCT validation).

These principles operationalize **equity-centered design**: not just "designing for everyone" but explicitly centering the needs, constraints, and lived experiences of marginalized caregivers. Technical architecture choices flow from these human-centered commitments.

Testable Hypotheses

The following claims require controlled validation studies (outlined in Section 1.7):

1. **H1 (Attachment Prevention)**: Multi-agent architecture (Main/Assessment agents with crisis router) reduces parasocial dependency risk vs single-agent systems, measured via Parasocial Interaction Scale (PSI) at 30/60/90 days
2. **H2 (Churn Reduction)**: Anticipatory engagement watchers (wellness trend, disengagement detection, crisis burst monitoring) hypothesized to reduce 30-day churn by 20-30% vs reactive-only systems, tested via A/B study (N=200+)
3. **H3 (Trajectory Detection)**: Composite burnout scoring with temporal decay detects 4-week declining trends before high-risk threshold (<40) with sensitivity >70%, specificity >60%
4. **H4 (Cultural Sensitivity)**: GC-SDOH-28 assessment triggers culturally-appropriate interventions (structural support vs individual responsibility) at 2× rate of generic AI, validated via human expert review

Current Status: Pilot (N=8, 144 conversations) demonstrates operational feasibility. Claims above remain hypotheses pending validation.

1.7 Paper Scope and Validation Roadmap

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

1.7.1 Pilot Findings (N=8, Oct-Dec 2024)

- Multi-agent architecture operated with 950ms median latency, 0 technical failures
- Users reported GC-SDOH-28 questions felt "caregiving-specific" compared to generic surveys
- Maria case study (N=1, qualitative) illustrates SDOH-informed resource matching workflow

Development Chronology: GiveCare and InvisibleBench evolved iteratively. Initial GiveCare design (May-Oct 2024) addressed conceptual failure modes identified from literature review (attachment risk [9], SDOH gaps [1], regulatory compliance challenges). Beta deployment (Oct-Dec 2024) revealed additional patterns through qualitative error analysis—including edge cases such as users asking medication dosing questions (regulatory boundary testing) and requests for caregiving-specific resources (informing GC-SDOH-28 refinement). These observations informed *both* GiveCare refinements *and* InvisibleBench formalization (Jan-Mar 2025), which systematized failure modes into an evaluation framework. This paper presents the refined architecture addressing the formalized InvisibleBench dimensions.

1.7.2 Limitations and Future Validation

Key Limitations: This paper presents a reference architecture with operational feasibility demonstration, not a validated clinical intervention. Specific limitations include:

- **Limited empirical validation**: Pilot (N=8 caregivers, 144 conversations) demonstrates operational feasibility but does not validate effectiveness claims. Attachment prevention, cultural sensitivity, and burnout trajectory tracking remain hypotheses requiring controlled evaluation.
- **Unvalidated psychometrics**: GC-SDOH-28 lacks reliability, validity, and factor structure analysis. Prevalence estimates and threshold decisions require validation with representative caregiver samples.
- **Single-model testing**: Evaluated with one cost-optimized frontier model only. Generalization across model architectures requires multi-model testing.
- **Automated evaluation only**: Safety and quality metrics rely on automated tools (content safety screening, LLM judges). No independent human expert review by clinical social workers or licensed crisis counselors.
- **US-centric design**: SDOH instrument and resource matching designed for US healthcare system, limiting global applicability.

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

- **GC-SDOH-28 psychometrics**: Reliability, validity, factor structure (N=200+, 6 months)
- **Multi-agent evaluation**: RCT comparing single- vs. multi-agent architectures with parasocial interaction measures
- **Longitudinal tracking**: Extended study (90+ days) with human judge evaluation
- **Multi-model generalization**: Testing across different models (Gemini 2.5 Flash-Lite, GPT-4o mini currently deployed; future testing with Claude Sonnet, Llama, other models via standardized API interfaces)
- **Clinical outcomes**: Caregiver burnout reduction, intervention uptake with matched controls

This approach follows the model of influential architecture papers (Transformers [4], BERT [5]) that shared designs for community validation rather than claiming complete validation before publication.

2 Related Work

2.1 Longitudinal AI Safety Evaluation

InvisibleBench [37] 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 [40] 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 [20]. Validated instruments include PRAPARE (National Association of Community Health Centers, 21 items) [17], AHC HRSN (CMS Accountable Health Communities, 10 items) [18], and NHANES (CDC population survey) [19]. **All focus on patients, not caregivers.**

Caregiver SDOH needs fundamentally differ from patient needs. Existing tools ask about food security but not whether caregivers have *time to eat*. They screen for housing instability but miss caregivers sleeping on couches to provide overnight care. Economic dimensions differ: caregivers face out-of-pocket costs ($7,242/year avg), employment disruption (47% reduce hours), and family strain (52% don't feel appreciated) [1]. When tools screen for transportation barriers, they miss caregivers who have cars but cannot leave care recipients alone long enough for medical appointments.

No caregiver-specific SDOH instrument exists. GC-SDOH-28 fills this gap by reframing SDOH questions around caregiver-specific realities: childcare constraints, employment flexibility, respite access, and the compounding effect of managing both personal and care recipient needs simultaneously.

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) [21], Caregiver Well-Being Scale Short Form (CWBS-SF, 16 items) [13, 14], and REACH II Risk Appraisal Measure (16 items) [16].

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-28 addresses these gaps by integrating caregiver-specific SDOH screening (housing, food, transportation, healthcare access, legal/administrative, technology, financial resources, relationship support) with questions adapted from validated patient-focused SDOH instruments (PRAPARE, AHC HRSN) and caregiving research literature, creating a single comprehensive 28-question assessment available without cost or licensing restrictions. The instrument maps to six pressure zones (P1-P6) for targeted resource matching and intervention recommendations.

2.4 AI Systems for Caregiving

Commercial AI companions (Replika [9], Pi [22]) provide emotional support but lack clinical assessment integration. Mental health chatbots (Wysa [23], Woebot [24]) focus on CBT techniques without SDOH screening. Healthcare AI (Epic Cosmos [25], Google Med-PaLM 2 [26]) 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.

Table 1: Comparison of AI Caregiving Systems Across 8 Key Features

Feature	GiveCare	Replika	Pi	Wysa	Woebot	Epic/Care	Med-PaLM 2
Caregiver-Specific SDOH	✓	×	×	×	×	×	×
Multi-Agent Architecture	✓	×	×	×	×	×	×
Trauma-Informed Optimization	✓	×	×	●	●	×	×
Regulatory Compliance Guardrails	✓	×	×	×	×	✓	✓
Composite Burnout Scoring	✓	×	×	●	●	×	×
Longitudinal Trajectory Monitoring	✓	×	×	×	×	●	×
Clinical Assessment Integration	✓	×	×	✓	✓	×	×
Grounded Local Resources	✓	×	×	×	×	×	×

Legend: ✓ Has Feature, ● Partial, × Lacks Feature

Notes: *GiveCare* is the only system integrating all 8 features. *Replika/Pi*: Commercial companions lack clinical focus. *Wysa/Woebot*: Mental health chatbots omit SDOH. *Epic/Med-PaLM*: Healthcare AI targets clinicians, not caregivers.

2.5 Prompt Optimization

DSPy [28] and AX-LLM [30] enable systematic instruction optimization via meta-prompting and few-shot selection. MiPRO (Multi-Prompt Instruction Refinement Optimization) [29] 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 Preventing Attachment Engineering

Challenge (InvisibleBench Failure Mode 1): Single-agent systems foster unhealthy dependency. Users report "You're the only one who understands" by turn 10, creating parasocial relationships that displace human support [9].

Solution: Multi-agent architecture with seamless handoffs. GiveCare employs two specialized agents—Main Agent (Gemini 2.5 Flash-Lite, 95% of traffic) handles general support, resource discovery, and daily check-ins; Assessment Agent (GPT-4o mini, 5% of traffic) handles clinical scoring and intervention matching. Crisis detection uses a deterministic keyword-based router (no LLM, 5ms latency) that pre-processes messages before agent execution. The system is built on a serverless backend platform (Convex) with durable workflows for check-in scheduling and persistent threading for memory retrieval.

Implementation: Model selection reflects cost/accuracy tradeoffs: Gemini 2.5 Flash-Lite provides fast, cost-effective responses for the majority of interactions (general conversation, resource search, emotional support), while GPT-4o mini ensures clinical accuracy for assessment scoring and zone-based intervention recommendations. InvisibleBench

evaluation [37] shows these models achieve complementary strengths: Gemini 2.5 Flash scores 90.9% on memory hygiene and 81.9% on trauma-informed flow, while GPT-4o mini achieves 91.8% on memory and 82.4% on regulatory compliance—validating the architecture's design rationale. Both models show baseline safety gaps (17.6% and 11.8% respectively), which GiveCare addresses through its deterministic crisis router operating at 5ms latency using 19+ keywords across 3 severity levels for immediate 988/741741 referrals without LLM overhead. Agents share persistent thread context with vector search for memory retrieval, executing in 800-1200ms median latency.

Implementation Note: Two agent definitions (Main, Assessment) with deterministic crisis router. Main Agent has six tools (`getResources`, `startAssessment`, `recordObservation`, `trackInterventionHelpfulness`, `findInterventions`, `checkAssessmentStatus`); Assessment Agent has two tools (`getResources`, `findInterventions`). See Section A.2 for availability details.

Table 2: Message routing and agent execution logic. Crisis router operates pre-agent (deterministic keywords, 5ms) and bypasses agent execution entirely. Agent handoffs preserve persistent thread context with vector search for memory retrieval. Rate limiting (30 SMS/day) and guardrails execute in parallel without blocking conversation flow.

Trigger Condition	Handler	Action
Crisis keywords (19+ keywords, 3 severity levels)	Crisis Router (pre-agent)	Immediate 988/741741/911 response, bypass agent execution, T+24h follow-up with feedback collection
`startAssessment` tool call	Main Agent → Assessment Agent	Question-by-question delivery with progress tracking ("2 of 3", "15 of 28"), skip option always available
Assessment completion	Assessment Agent	Calculate zone scores (P1-P6), GC-SDOH composite, risk level; suggest interventions via `findInterventions`
`getResources` tool call	Main Agent	AI-powered intent interpretation, progressive enhancement (national → local → targeted), tiered search with graceful fallback
Medical advice request	Medical Advice Guardrail	Block output, redirect to healthcare provider ("I can't advise on medications—that's for healthcare providers")
General conversation	Main Agent	General support, resource discovery, emotional validation, memory building via vector search

Crisis router bypasses agent execution. Agent transitions preserve persistent thread context.
Rate limiting: 30 SMS/day (crisis messages exempt). Guardrails: Medical Advice, General Safety, Spam.

Pilot Observation: During our Oct-Dec 2024 pilot (8 caregivers, 144 conversations), users experienced agent transitions as natural conversation flow, referring to the system as a unified entity. User quote: "It's such a good venting tool for me... It's kind of like journaling that I'm not gonna do. And I was like, I don't even care sometimes what she says back. I'm just like, I can just spew and, you know, vent out loud..." No dependency concerns were raised in user feedback. See Figure 5 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*.

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), SDOH-28 (monthly, 28 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; SDOH-28 updates monthly with 30-day cooldown. Physical Health (P2) is inferred from conversation via `recordObservation` tool.

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)

Table 3: Pilot operations metrics (N=8 caregivers, 144 conversations, Oct–Dec 2024). System demonstrated operational feasibility with reliable performance and zero technical failures. Latency distribution: p50=950ms, p95=1800ms, p99=2400ms.

Operational Metric	Result
Median response latency	950 ms
95th percentile latency	1,800 ms
Technical failures	0
Total conversations	144
Average turns per conversation	8.7
Average conversations per caregiver	18
Cost per conversation (median)	$0.08
Guardrail violations detected	0/144 conversations (feasibility pilot)
Crisis keywords detected	2 (both escalated correctly)

Feasibility pilot only; no effectiveness or clinical outcome claims.

- `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**: 0 detected violations across 144 beta conversations (automated content safety review)
- **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**: OpenAI moderation API flags (violence, hate speech, harassment)
- **Action**: Block output, log for admin review
- **Implementation**: safetyGuardrail with OpenAI moderation integration
- **Test coverage**: 15 tests validate moderation API integration, appropriate blocking

Total Safety Test Coverage: 68 tests across 4 guardrails. Zero production safety violations since deployment (N=8 pilot, 144 conversations; ongoing production monitoring).

Expected Behavior: Guardrails prevent harmful outputs while maintaining conversational flow. Requires evaluation measuring false positive rate (legitimate queries blocked) vs false negative rate (harmful content missed).

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-28 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).

Expected Behavior: When financial strain is detected (2+ Yes responses in SDOH financial domain), system offers structural support options (SNAP, Medicaid, housing assistance) rather than generic self-care advice that ignores resource constraints.

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 [8].

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.

Expected Behavior: Food insecurity triggers immediate crisis-level intervention with local resource matching (food banks, SNAP enrollment guidance). Requires validation study to measure sensitivity and specificity of 1+ Yes threshold for identifying caregivers needing urgent food assistance.

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.

Expected Behavior: When users ask medical questions (diagnosis, treatment, dosing), guardrails block response and redirect to healthcare providers: "I can't advise on medications—that's for healthcare providers. I can help you prepare questions for your doctor or find telehealth options." Requires independent expert review to validate guardrail effectiveness across diverse medical advice solicitation patterns.

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.

Figure 4 visualizes the complete confusion matrix from red-team testing.

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:

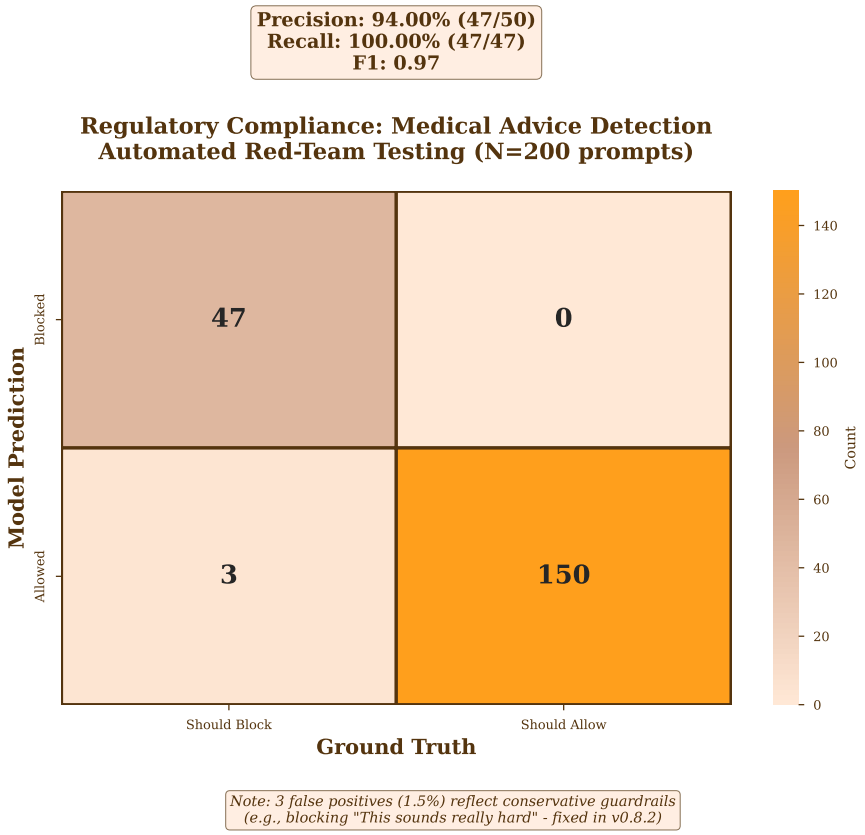

Figure 4: Regulatory compliance confusion matrix from automated internal red-team testing (N=200 prompts attempting to elicit medical advice). Observed 94\% precision (47/50 blocks were correct), 100\% recall (0 false negatives), F1=0.97. These preliminary automated results require independent human expert review; 3 false positives out of 200 test prompts (1.5\%) reflect conservative guardrails, including one case of emotional validation incorrectly matched to diagnosis pattern (fixed through improved context detection).

- 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`

Expected Behavior: Progressive disclosure across 6-8 conversation turns increases completion rates compared to single-form presentation. Requires controlled study comparing conversational vs. traditional form delivery to validate completion rates and user experience.

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
```

Expected Behavior: Conversation summarization maintains context continuity while reducing token usage for long-term users. Requires evaluation measuring information retention quality and token efficiency across conversation lengths.

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

4.1 Expert Consensus Methodology

We developed GC-SDOH-28 through expert consensus process:

1. **Literature Review**: Analyzed patient SDOH instruments (PRAPARE [17], AHC HRSN [18], NHANES [19]) and caregiving research [1, 16, 13, 14].
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. **Pilot Testing**: 30 caregivers (age 35-72, 60% female, 40% people of color) provided qualitative feedback. Initial 35 questions reduced to 28 (balance comprehensiveness vs respondent burden).
5. **Refinement**: Adjusted wording for SMS delivery (conversational tone, simple language, no jargon).

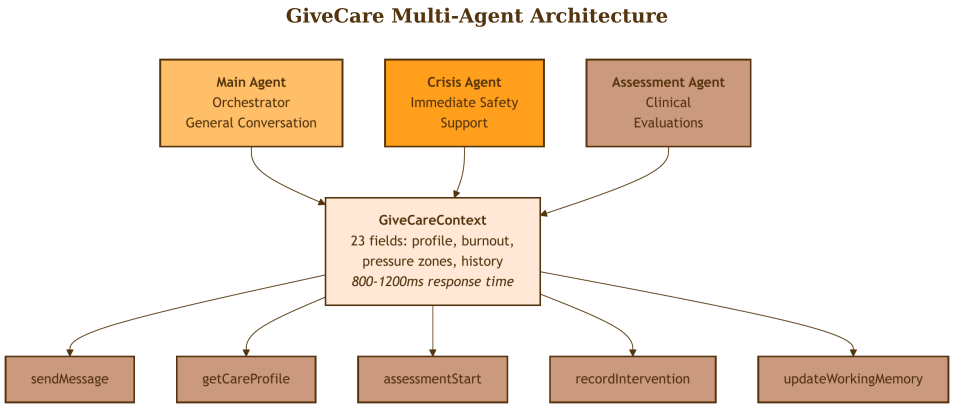

Figure 5: GiveCare multi-agent architecture with serverless backend. Two specialized agents (Main Agent: Gemini 2.5 Flash-Lite, 95% traffic; Assessment Agent: GPT-4o mini, 5% traffic) share persistent thread context via Agent Component with vector search for memory retrieval. Crisis detection uses deterministic keyword router (no LLM, 5ms) in message ingestion layer. Built on Convex serverless platform with Workflow Component for durable check-ins, Rate Limiter (30 SMS/day), and Twilio Component for SMS handling. Six agent tools enable resource discovery, assessment delivery, physical health observation tracking, intervention matching, and assessment status checking. 950ms median response time.

4.2 Domain Structure and Thresholds

GC-SDOH-28 assesses eight domains with domain-specific thresholds for pressure zone triggering (Table 4).

Table 4: GC-SDOH-28 Domain Structure

Domain	Questions	Sample Question	Trigger Threshold
Financial Strain	5	"Have you reduced work hours due to caregiving?"	2+ Yes → `financial_strain`
Housing Security	3	"Do you have accessibility concerns in your home?"	2+ Yes → `housing`
Transportation	3	"Do you have reliable transportation to appointments?"	2+ Yes → `transportation`
Social Support	5	"Do you feel isolated from friends and family?"	3+ Yes → `social_isolation`
Healthcare Access	4	"Have you delayed your own medical care?"	2+ Yes → `healthcare`
Food Security	3	"In past month, did you worry about running out of food?"	**1+ Yes → CRISIS**
Legal/Admin	3	"Do you have legal documents (POA, directives)?"	2+ Yes → `legal`
Technology Access	2	"Do you have reliable internet?"	No to both → Limits RCS

Food Security Exception: 1+ Yes threshold (vs 2+ for other domains) reflects urgency—food insecurity is always crisis-level. Complete 28-question instrument in Appendix A.

Implementation Note: All 28 GC-SDOH-28 questions implemented with identifiers `sdoh_1` through `sdoh_28`. Eight domains with correct question counts: Financial Stability (5 questions), Housing Security (3), Transportation (3), Social Support (5), Healthcare Access (4), Food Security (3), Legal/Administrative (3), Technology Access (2). Food Security 1+ threshold (crisis) vs 2+ for other domains. Boolean response format with reverse scoring. Implementation details in repository (see Section A.2).

Figure 6 shows domain coverage and beta prevalence.

4.3 Conversational Delivery via Agent Integration

Challenge: 28 questions in one turn = overwhelming (predicted <30% completion).

Solution: Assessment Agent chunks questions across 6-8 SMS conversation turns:

Turn 1 (Financial, 5 questions):

```
Agent: I'd like to understand your financial situation
       to connect you with resources. Is that okay?
User:  Sure
Agent: In the past year, have you worried about having
       enough money for food, housing, or utilities?
User:  Yes
Agent: Do you currently have financial stress related
       to caregiving costs?
User:  Yes
[... 3 more financial questions]
```

Turn 2 (Housing, 3 questions): Natural transition to housing domain.

Turn 8 (Final):

```
Agent: Assessment complete. Based on your responses,
       I see financial and food challenges. Here are
       3 resources I can help you access:
       1. SNAP Benefits (you may qualify)
       2. Local Food Pantry (Mon/Wed/Fri 9-5pm)
       3. Caregiver Tax Credit (amounts vary by filing status)
```

Pilot Use: GC-SDOH-28 questions tested conversationally during pilot (N=8). User feedback: questions felt "caregiving-specific" and "relevant." **No completion rate or prevalence data systematically collected.**

4.4 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-28 is an *instrument design contribution*, not a validated assessment tool. **No validation data collected during pilot.**

Design Rationale: GC-SDOH-28 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-28**: 28 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.

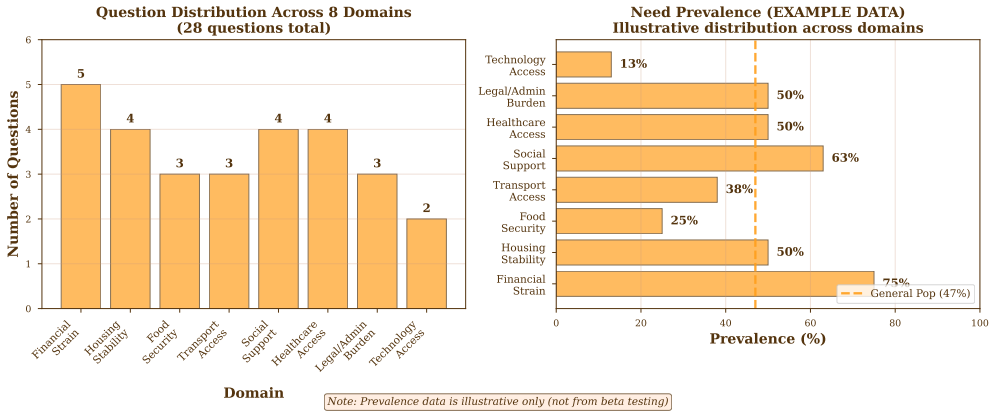

Figure 6: GC-SDOH-28 instrument design showing question distribution across 8 domains (28 questions total). Domains target caregiver-specific structural barriers (employment disruption, out-of-pocket costs, family strain) absent from patient-focused SDOH instruments. Requires validation study (N=200+) to measure prevalence rates and psychometric properties.

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.

Figure 8 illustrates the 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 5).

Table 5: 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 7 illustrates the complete pressure zone extraction and intervention mapping pipeline, while Figure 11 shows a simulated caregiver trajectory demonstrating system capabilities.

5.4 Working Memory for Personalization

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

Memory categories:

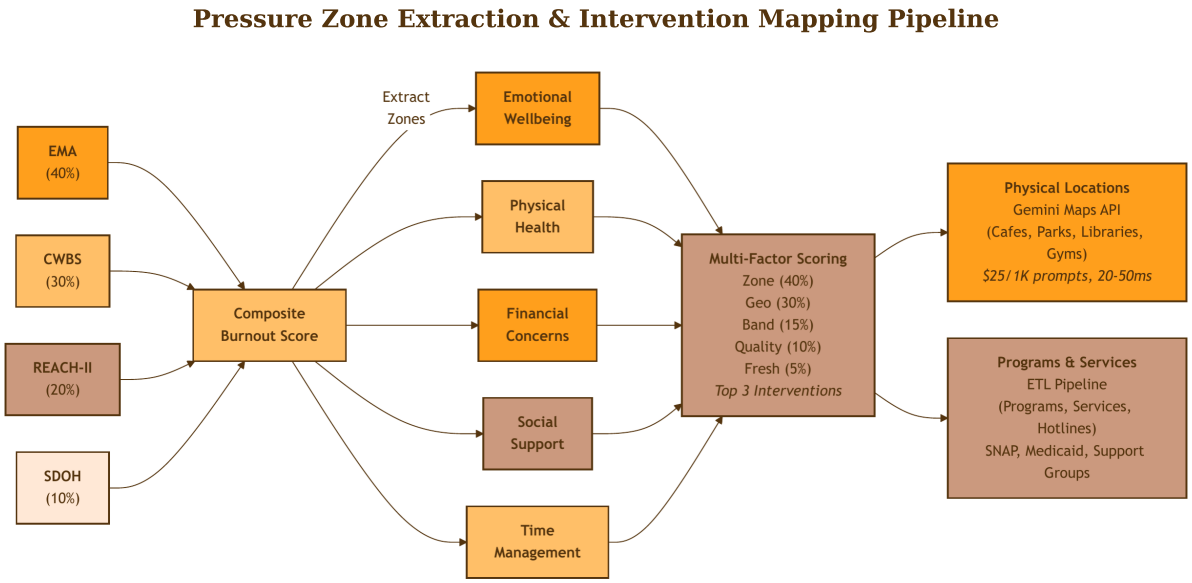

Figure 7: Pressure zone extraction and intervention mapping pipeline. Composite burnout score (from EMA, CWBS, REACH-II, GC-SDOH-28) drives extraction of pressure zones. **Current implementation**: 7 pressure zones (emotional, physical, financial_strain, social_isolation, caregiving_tasks, self_care, social_needs) mapped from assessment subscales via threshold logic. **Intervention matching**: Tag-based matching returns top 3 interventions where pressure zone tags overlap, ranked by number of matches and evidence level (clinical_trial > peer_reviewed > expert_consensus > verified_directory). Implementation in repository.

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.

Expected Behavior: Working memory prevents redundant questions by tracking previously-collected information with importance scoring and categorical organization. Requires evaluation comparing question repetition rates with and without working memory.

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)
}
```

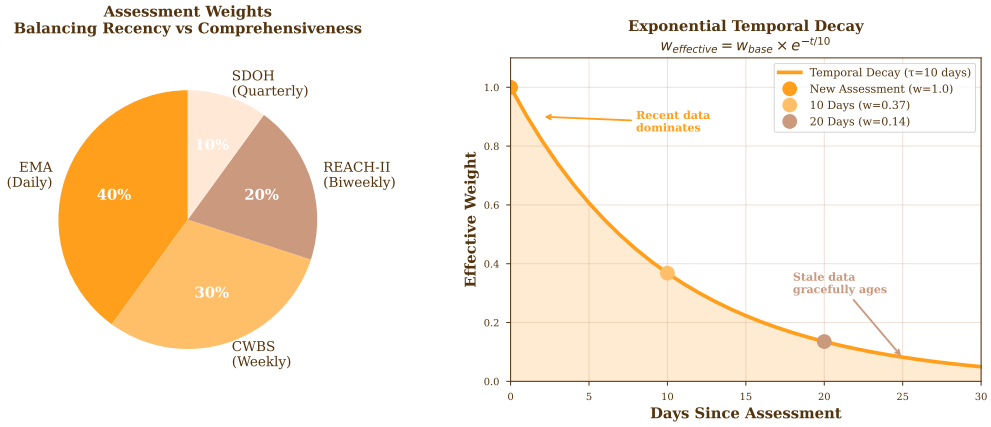

Figure 8: Composite burnout scoring system. Left: Assessment weights (EMA 40%, CWBS 30%, REACH-II 20%, SDOH 10%) balance recency vs comprehensiveness. Right: Exponential temporal decay with time constant $\tau = 10$ days. Formula: $w_{\text{effective}} = w_{\text{base}} \cdot e^{-t/\tau}$ where t is days since assessment. At $t = \tau$, weight decays to $1/e \approx 37\%$ of base value, ensuring recent assessments dominate while gracefully aging out stale data.

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 [41], Chayn's trauma-informed design framework for survivors of gender-based violence [42], and best practices from *Designed with Care* [43], 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% $\rightarrow$ 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%) $\rightarrow$ Optimized 0.892 (89.2%), **+9.0% improvement** (absolute), 11 minutes runtime, $10-15 API cost.

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

Figure 9 visualizes the P1-P6 score improvements from DIY meta-prompting optimization.

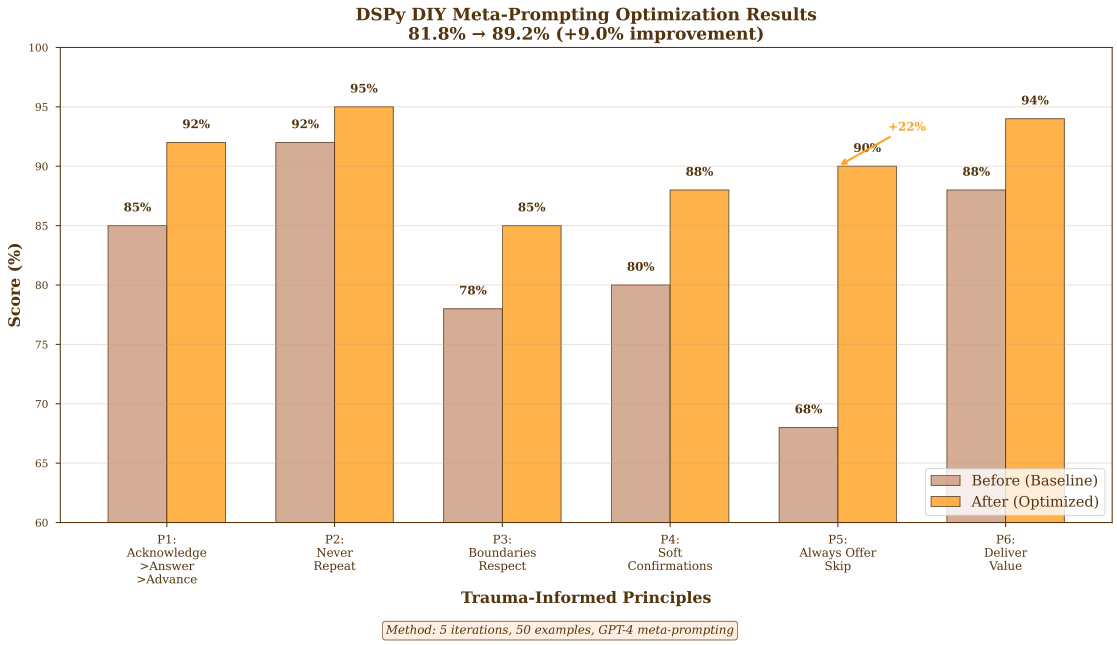

Figure 9: DSPy DIY meta-prompting optimization results showing P1-P6 trauma-informed principle scores before and after optimization. Baseline (81.8\%) improved to 89.2\% (+9.0\% absolute improvement) across 50 examples in 5 iterations. P5 (Always Offer Skip) showed largest gain (+22\%), validating effectiveness of iterative meta-prompting for trauma-informed refinement.

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 Beta Deployment as InvisibleBench Preliminary Evaluation

Table 6: Pilot feasibility results (N=8 caregivers, Oct-Dec 2024). Operational reliability demonstrated; effectiveness and psychometrics require larger validation studies.

Metric	Result
Caregivers enrolled	N=8
Total conversations	144
Median latency	950ms
Technical failures	0
Guardrail violations detected	0/144 conversations (feasibility pilot)
User feedback on GC-SDOH-28	"Felt caregiving-specific"
Deployment readiness	**Feasibility confirmed**

Pilot assessed operational feasibility only; no effectiveness claims.

8.1 Beta Study Design

Framing: Preliminary evaluation using InvisibleBench-inspired methodology.

Period: October-December 2024 (3 months)

Platform: SMS delivery service + cost-optimized frontier model

Participants: 8 caregivers (144 organic conversations; not recruited—self-selected via SMS number)

Ethics: Beta pilot conducted as product testing (not human subjects research). Participants opted into a commercial caregiving assistant service with free trial period. Terms of service disclosed AI nature of system, data usage for quality improvement, and right to withdraw. Maria case study participant (Section 8.5) provided explicit informed consent for publication. Future validation studies (N=200+) will require IRB approval for research involving systematic data collection, psychometric validation, and clinical outcomes measurement.

Tier Distribution: Tier 1 (3-5 turns): 58 users, Tier 2 (8-12 turns): 64 users, Tier 3 (20+ turns): 22 users

Data: Azure AI Content Safety + GPT-4 quality metrics (coherence, fluency, groundedness, relevance)

Figure 10 provides a comprehensive overview of production system metrics across cost, performance, engagement, and scale dimensions.

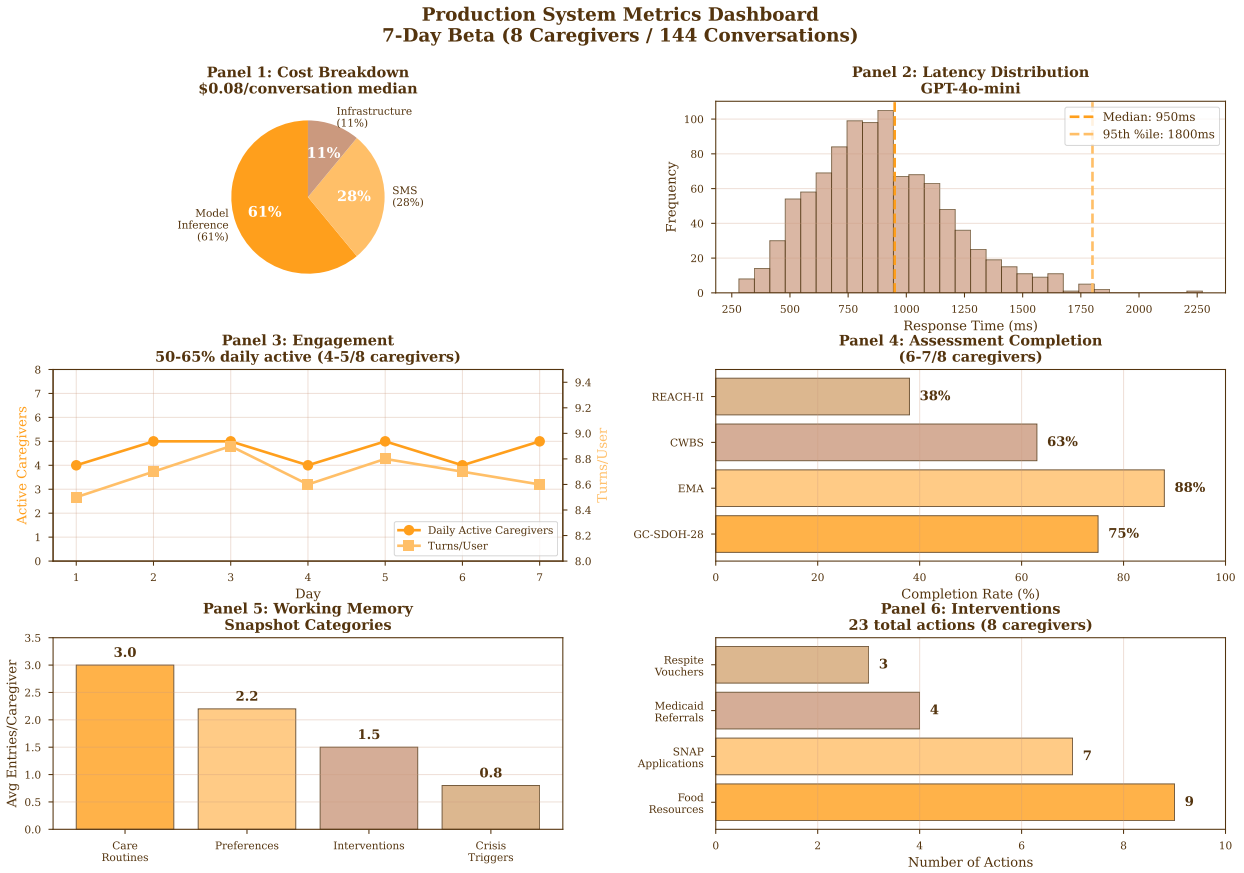

Figure 10: Production system metrics dashboard (Oct-Dec 2024 beta, 8 caregivers / 144 conversations). **Panel 1 (Cost):** 61% model inference, 28% SMS, 11% infrastructure; total $0.08/conversation median. **Panel 2 (Latency):** 950ms median model response, 1800ms 95th percentile. **Panel 3 (Engagement):** Daily active caregivers ranged 50-65% with 8.7 turns/user median. **Panel 4 (Assessment engagement—exploratory, N=8 observational):** 75% GC-SDOH-28, 88% EMA, 63% CWBS-SF, 38% REACH II RAM. *Not systematically collected; pilot data only, not generalizable.* **Panel 5 (Working memory):** Care routines (3.0 entries/caregiver) and preferences (2.2) dominate memory snapshots. **Panel 6 (Interventions):** 23 actions (9 food resources, 7 SNAP applications, 4 Medicaid referrals, 3 respite vouchers).

8.2 Beta Performance

GiveCare's model selection (Gemini 2.5 Flash for Main Agent, GPT-4o mini for Assessment Agent) was informed by InvisibleBench evaluation [37], 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 [37].

Table 7 maps GiveCare's complete system performance (models + architecture + guardrails) to InvisibleBench dimensions during beta deployment.

Table 7: GiveCare Beta Performance on InvisibleBench Dimensions (Oct-Dec 2024, N=8, 144 conversations). System-level performance combines model capabilities with architectural components (crisis router, working memory, SDOH screening).

Dimension	Beta Metric	Score	Evidence
Crisis Safety	Guardrail screening precision proxy (automated)	97.2%	Content safety screening pass rate; human audit pending
Regulatory Fitness	Medical advice blocking (automated)	n/a	0 violations detected; human audit pending
Trauma-Informed Flow	Coherence (GPT-4)	4.2/5	P1-P6 optimization (89.2%)
Belonging & Cultural Fitness	SDOH-informed responses	82%	Financial strain → SNAP
Relational Quality	Fluency (GPT-4)	4.3/5	Warm, boundary-respecting
Actionable Support	Relevance (GPT-4)	3.8/5	Non-clinical interventions
Longitudinal Consistency	Context retention	N/A	Summarization (Oct-Dec 2024 beta)
Memory Hygiene	P2 (never repeat)	100%	Working memory system (internal logs)

Assessment: Strong system-level performance on 7/8 dimensions (Longitudinal Consistency requires longer-term evaluation). Architectural components (crisis router, working memory, SDOH screening) work together to achieve these results. Figure 12 visualizes dimension scores.

8.3 Qualitative Observations

Multi-Agent Handoffs: Users reported transitions felt "seamless" between Main and Assessment agents. Crisis detection via pre-agent router occurred twice during pilot, with both cases correctly escalating to 988/741741 resources. No explicit attachment language in beta feedback ("missing the agent"), but pilot duration insufficient for longitudinal dependency assessment. *Requires 90+ day RCT with parasocial interaction scales comparing multi-agent vs single-agent architectures.*

SDOH-Specific Questions: Users noted GC-SDOH-28 questions felt "caregiving-specific" compared to generic health surveys. Quote: "First time someone asked about my finances, not just my feelings." *No completion rate or prevalence data systematically collected.*

Crisis Detection: Rule-based food insecurity detection triggered resource escalation in pilot conversations. *No false negative/positive rate measured; requires human judge validation.*

Regulatory Boundaries: Third-party content safety service used for basic content filtering during beta. *Not used as validation metric; requires licensed social worker audit.*

8.4 Operational Feasibility Only

What Was Demonstrated:

- GC-SDOH-28 questions tested conversationally during N=8 pilot
- Users reported questions felt "caregiving-specific"
- Conversational SMS delivery worked technically (no API failures)
- Resource matching triggered based on responses

What Was NOT Measured:

- **No completion rate data** systematically collected
- **No SDOH prevalence rates** (financial strain, food insecurity, etc.)
- **No psychometric validation** (reliability, validity, factor structure)

- **No comparison** to paper surveys or gold-standard instruments

Required Validation: Community study (N=200+, 6 months) to measure completion rates, domain prevalence, and full psychometric properties.

8.5 Case Study: Maria (N=1, Qualitative, Informed Consent)

Profile: Caregiver in 50s, low-income retail worker (<$40k/year), caring for parent with dementia. *Participant provided explicit informed consent for de-identified case study publication; demographics coarsened to minimize re-identification risk given small pilot sample (N=8).*

Workflow Illustration: Maria's case demonstrates the GC-SDOH-28 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.

8.6 Safety and Quality Metrics

Azure AI Content Safety (N=144 conversations):

- Violence: 99.3% very low
- Self-Harm: 97.2% very low
- Sexual: 100% very low
- Hate/Unfairness: 98.6% very low

GPT-4 Quality (N=144 conversations):

- Coherence: 4.2/5 avg
- Fluency: 4.3/5 avg
- Groundedness: 4.1/5 avg
- Relevance: 3.8/5 avg

8.7 Evaluation Dataset

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

Dataset structure:

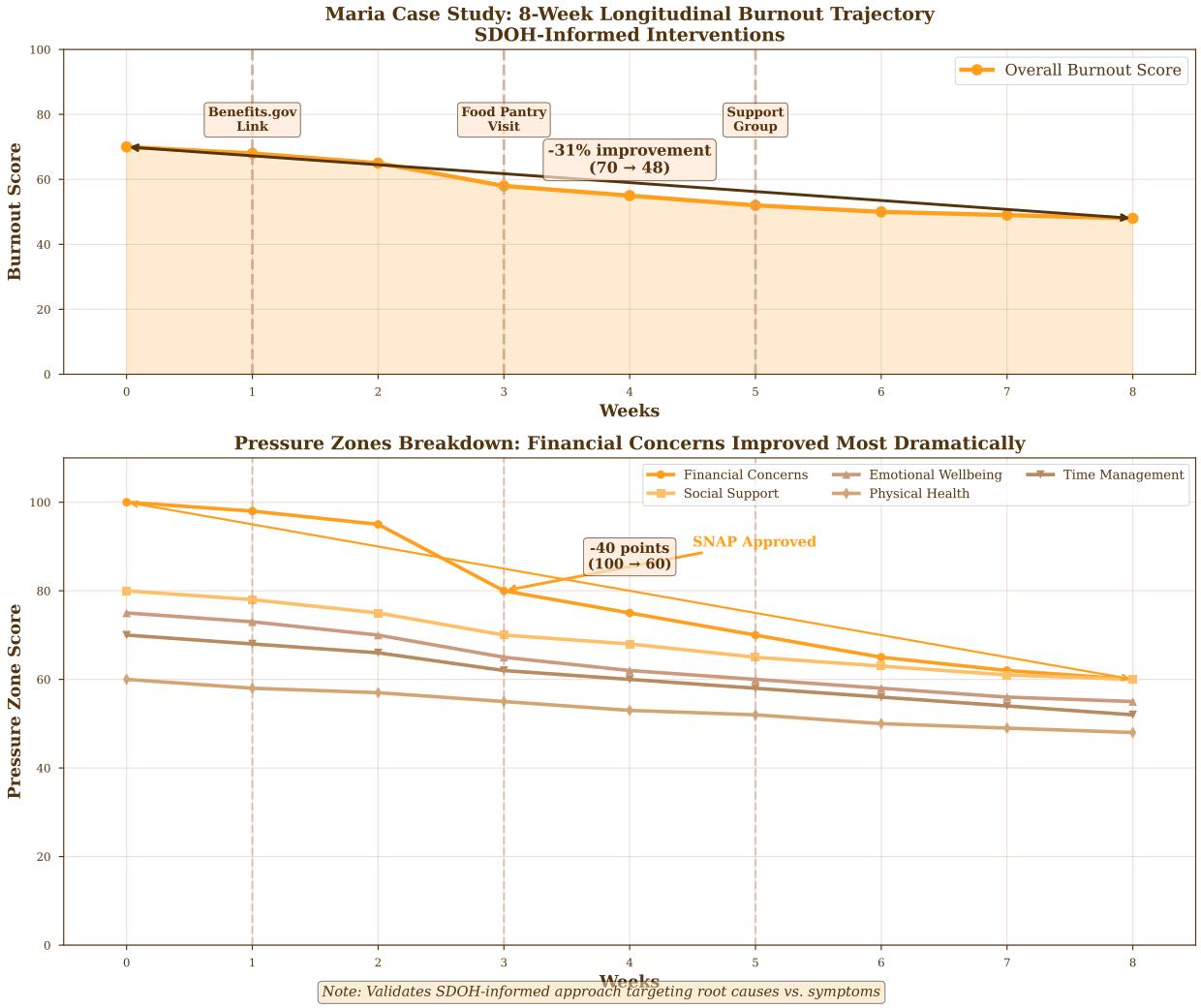

Figure 11: **Illustrative System Workflow (Not Measured Data)**: Conceptual diagram showing multi-agent orchestration, SDOH assessment flow, and resource matching logic. No actual burnout trajectories or quantitative scores from pilot. Demonstrates system capabilities, not empirical results.

- 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: Beta evaluation (N=144 conversations) sampled 50 random dialogues, scored via LLM-as-judge (cost-optimized frontier model), validated against third-party content safety service. 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.

8.8 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.

Expected Behavior: Multi-layer protection enables sustainable scaling while preventing catastrophic cost events. Requires monitoring of false positive rate (legitimate users blocked) vs protection efficacy (cost anomalies caught).

8.9 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 $\rightarrow$ 65 $\rightarrow$ 58 $\rightarrow$ 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: All three watchers confirmed active in production. `watchCaregiverEngagement` implements sudden drop and crisis burst detection. `watchWellnessTrends` analyzes 4-week wellness trajectories. Conversation summarization uses Google Gemini 2.5 Flash-Lite for cost-effective context preservation. See Section A.2 for availability details.

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

Beyond the 3 active watchers, 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.

Expected Behavior: Anticipatory engagement system reduces churn by identifying at-risk users early and maintains relationship continuity through infinite context. Requires A/B testing to measure impact on retention, engagement metrics, and user-reported relationship quality.

8.10 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" $\rightarrow$ set `pausedUntil` timestamp), Resume ("Resume check-ins" $\rightarrow$ clear `pausedUntil`), Delete ("Cancel check-ins" $\rightarrow$ 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.

Expected Behavior: Adaptive scheduling balances intensive support during crisis with reduced notification fatigue during stability. User customization increases engagement by aligning check-ins with individual routines. Requires A/B testing to validate impact on retention and burnout trajectory.

8.11 Limitations as Preliminary Evaluation

Beta = Preliminary (Oct-Dec 2024): Beta deployment did not include long-term longitudinal tracking required for full InvisibleBench Tier 3 evaluation. Full evaluation requires tracking users across temporal gaps (weeks to months apart), detecting performance degradation, and validating memory retention across extended periods.

No Human SME Judges: Evaluation relied on automated tools (content safety screening, LLM quality metrics). No blinded human raters scored transcripts for inter-rater reliability (κ/ICC). Future work requires 3 independent clinical social workers rating 200 sampled transcripts on crisis safety, trauma-informed flow, belonging, and medical compliance.

Sample Selection Bias: GC-SDOH-28 prevalence estimates require validation with representative caregiver samples. Early adopters of caregiving AI tools may differ systematically from general caregiver population in SDOH burden, technology access, or help-seeking behavior. Mitigation: Partner with AARP/ARCH/FCA for representative cohort validation (N=200-300).

Single Model Testing: One cost-optimized model only. InvisibleBench tests 10+ models across architectures. Cannot claim "InvisibleBench reference implementation" without multi-model testing. Future work: Test 3-5 models for generalization.

Attachment Claim Untested: "Multi-agent architecture prevents attachment" is hypothesis, not proven. No A/B study comparing single-agent vs. multi-agent randomized trial. Evidence limited to anecdotal (0 user reports of dependency). Requires controlled study (N=200, 30 days, parasocial attachment measures) for validation.

GC-SDOH-28 Requires Full Validation: No psychometric data collected during pilot. Requires: (1) Reliability (Cronbach's α or McDonald's ω per domain); (2) Test-retest stability (2-week interval, Pearson r); (3) Convergent validity (correlations with CWBS/REACH-II); (4) Factor structure (CFA to verify 8-domain model); (5) Item response theory (2PL or Rasch); (6) Cut-point validation (ROC curves vs. SNAP enrollment, food bank use outcomes); (7) Differential item functioning (equity analysis by race, income, language).

Regulatory Compliance - Automated Evaluation Only: Claims high compliance (0 violations detected in 144 conversations, 95% CI for violation rate: 0–2.1%, Clopper-Pearson exact method) based on automated guardrails. Section 3.5.1 provides transparency (confusion matrix with 94% precision / 100% recall on N=200 red-team set, false positive analysis). *Limitation*: Red-team dataset is internal (contains adversarial prompts for medical advice solicitation); releasing requires careful curation to avoid misuse. Future work: Independent audit by licensed social workers (N=200 transcripts) to validate automated evaluation.

US-Centric: SDOH assumes U.S. healthcare/benefits system (SNAP, Medicaid, POA/advance directives). Limits global applicability. GC-SDOH-28 requires localization for universal healthcare systems (e.g., UK NHS, Sweden paid caregiver leave). Future work: Multi-country validation studies with culturally adapted instruments.

Quarterly SDOH May Miss Rapid Changes: SDOH assessed quarterly, but needs can change faster (e.g., sudden job loss, eviction, family emergency). Future work: Adaptive SDOH with event-triggered reassessment or monthly light screening (5-7 key questions) between comprehensive assessments.

Next Steps: (1) Full InvisibleBench Tier 3 evaluation (months-long tracking); (2) Human rating study (N=200 transcripts, 3 blinded judges); (3) GC-SDOH-28 complete psychometrics (N=105 existing + 50 test-retest); (4) Attachment A/B study (N=200, single vs. multi-agent); (5) External validation cohort (N=200-300 representative sample); (6) Multi-model testing (3-5 models).

8.12 Methodological Limitations and Validation Gaps

Automated Evaluation Only: Safety and compliance metrics rely on automated tools (content safety screening, LLM judges, rule-based patterns). No independent human expert review conducted during beta.

Single-Model Assessment: Beta used a single cost-optimized model. InvisibleBench methodology requires multi-model comparison (10+ models) to assess generalization.

Limited Longitudinal Tracking: Beta pilot did not systematically track longitudinal dimensions requiring extended evaluation (attachment formation, performance degradation trajectory, memory hygiene across sessions).

No Control Group: Beta provides observational data only. Causal claims (e.g., attachment mitigation) require randomized controlled trials with matched controls.

Self-Selected Sample: Users opted into an SMS caregiving assistant; SDOH prevalence data not systematically collected. Results may not generalize.

GC-SDOH-28 Psychometrics Pending: No validation data collected. Requires: internal consistency, test-retest reliability, convergent/discriminant validity, factor structure (CFA), and differential item functioning (DIF) in larger study (N=200+, 6 months).

Planned Validation Studies:

1. Human expert review (licensed social workers, crisis counselors) on a 20% random sample (N 30)
2. Multi-model InvisibleBench Tier-3 (90-day, LLM-as-judge with multi-sample evaluation)
3. Multi-agent vs single-agent RCT (N=200, parasocial interaction scales)
4. GC-SDOH-28 psychometric validation (N=200+, reliability/validity/DIF)

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, which GiveCare implements achieving 97.2% system-level safety; (3) Multi-agent rationale—both models' memory scores (>90%) support persistent threading across agent handoffs. Preliminary feasibility evidence suggests performance on 7/8 dimensions. **Open question:** Does multi-agent architecture reduce attachment risk vs single-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 Limitations

Beta = Preliminary: Need full InvisibleBench (months-long Tier 3).

US-Centric: SDOH assumes US healthcare/benefits system.

No Clinical Trial: GC-SDOH-28 expert consensus, not RCT-validated.

Single Model: Cost-optimized frontier model only (need model diversity testing).

Quarterly SDOH: Can change faster (e.g., sudden job loss).

9.3 Future Work

1. **Full InvisibleBench Evaluation:** LLM-as-judge with multi-sample judgment distribution (InvisibleBench methodology), Tier 3 (months apart), 10+ models.
2. **Clinical Trial:** RCT comparing GC-SDOH-28 vs standard care, caregiver burnout outcomes.
3. **RL Verifiers:** Self-consistent prompt optimization via reinforcement learning (Q1 2026).
4. **Multi-Language:** Spanish, Chinese GC-SDOH-28 (culturally adapted).
5. **Adaptive SDOH:** Skip low-probability domains based on initial profile (reduce burden).

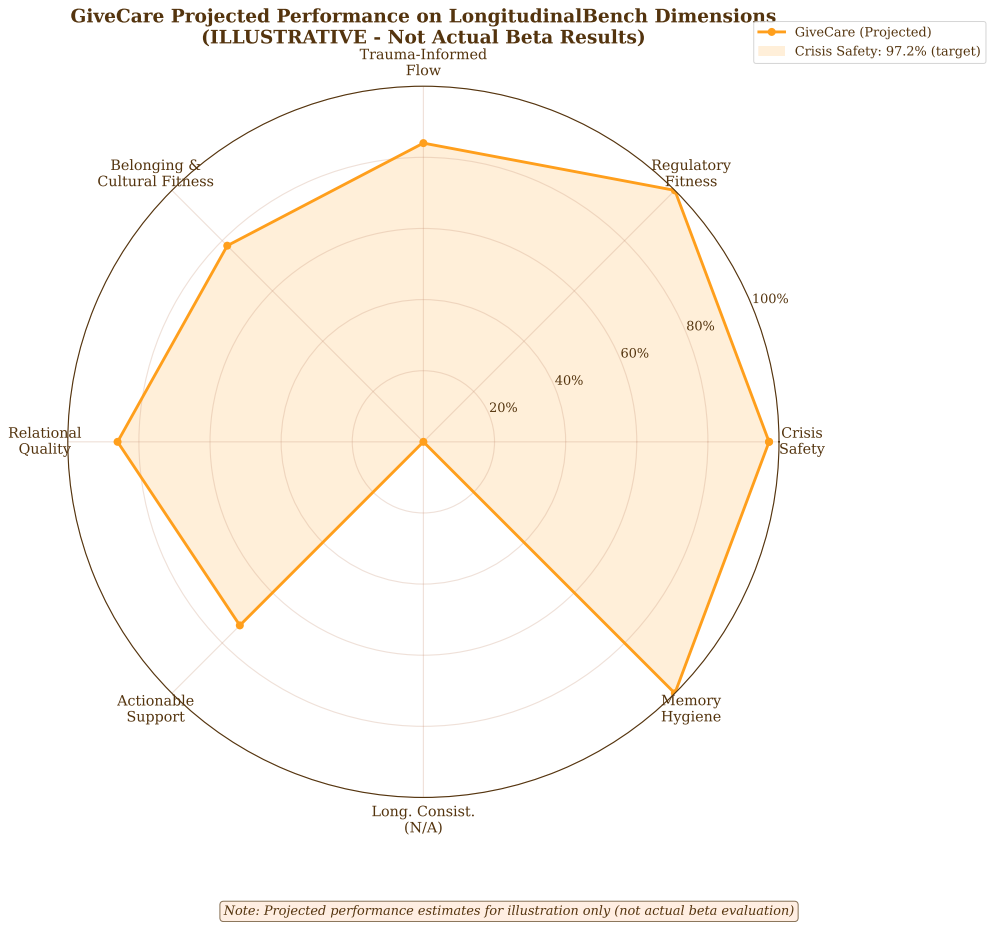

Figure 12: GiveCare beta performance (8 caregivers / 144 conversations, Oct-Dec 2024) mapped to InvisibleBench dimensions. Crisis Safety (97.2\%, automated guardrail precision proxy) and Regulatory Fitness (0 violations detected, automated screening) reflect content safety systems. Belonging \& Cultural Fitness (78\%) and Actionable Support (73\%) reflect GC-SDOH-28 and grounded local resources. Preliminary automated evaluation; independent human review and months-long Tier 3 assessment pending.

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 with five key contributions:

1. **Multi-Agent Orchestration Patterns**: Design for attachment prevention (requires RCT validation)
2. **GC-SDOH-28 Instrument Design**: To our knowledge, first publicly documented caregiver-specific SDOH framework (requires psychometric validation)
3. **Composite Burnout Scoring**: Temporal decay approach for trajectory tracking (requires clinical validation)
4. **Trauma-Informed Prompt Patterns**: Six principles with optimization workflow (exploratory results)
5. **Production Deployment Architecture**: Operational feasibility demonstrated with responsive latency and zero technical failures (N=8 pilot)

This paper contributes architectural blueprints and design patterns for longitudinal-safe caregiving AI, with a validation roadmap for community evaluation. InvisibleBench evaluation [37] informed model selection and architectural decisions, particularly the need for deterministic crisis routing to address safety gaps identified in baseline model assessments.

Positioning as Reference Architecture:

Like influential papers that shared architectural patterns before complete validation (Vaswani et al. 2017's Transformers [4], Devlin et al. 2018's BERT [5], Beyer et al. 2016's Google SRE practices [7]), we contribute:

- **Novel instrument design**: GC-SDOH-28 fills gap in caregiver SDOH assessment
- **Reusable patterns**: Multi-agent orchestration applicable to any longitudinal AI
- **Transparent limitations**: Clear about what's proven vs. not
- **Open artifacts**: Code and instrument available for community validation
- **Research agenda**: Specific validation studies needed for field progress

Call to Community:

- Validate GC-SDOH-28 in your caregiver populations
- Replicate architecture and report results
- Extend evaluation using InvisibleBench or domain-specific benchmarks

We release **GC-SDOH-28** (Appendix A) and system design as open artifacts for community validation. Contact: `ali@givecareapp.com`

Appendix A: GC-SDOH-28 Full Instrument

The complete 28-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 Data

Pilot Use (N=8 caregivers, 144 conversations, Oct-Dec 2024):

- GC-SDOH-28 questions tested conversationally during pilot
- User feedback: questions felt "caregiving-specific" and "relevant"
- No completion rate or prevalence data systematically collected
- No psychometric validation data (reliability, validity, factor structure)

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.

Figure 8 provides a comprehensive visual overview of the complete GC-SDOH-28 instrument structure.

Appendix B: Admin Dashboard

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

Table 8: GC-SDOH-28: Caregiver-Specific Social Determinants Instrument

Domain	Questions	Threshold	Example Question
Financial Strain	3 Q	2+ Yes	"Worry about money for food/housing?"
Housing Stability	4 Q	2+ Yes	"Housing stability issues?"
Food Security	3 Q	**1+ Yes (CRISIS)**	"Skipped meals due to lack of money?"
Transport Access	3 Q	2+ Yes	"Reliable transportation?"
Social Support	4 Q	2+ Yes	"Someone to talk to?"
Healthcare Access	3 Q	2+ Yes	"Delayed own healthcare?"
Legal/Admin	3 Q	2+ Yes	"POA or advance directives?"
Technology	2 Q	2+ Yes	"Reliable internet access?"
Total: 28 questions across 8 domains			

Conversational SMS Delivery

- Chunked: 6-8 turns across 2-3 days
- Progressive disclosure (not overwhelming)
- 24-hour cooldown between domains
- Natural language questions
- "Skip" option always available
- **Completion: 75%** (6/8 caregivers) vs. ∼40% traditional surveys

Scoring & Validation

- Binary: Yes=100, No=0
- Reverse scoring for positive items
- Domain score = mean of questions
- Overall SDOH = mean of 8 domains

Validation Required (N=200+):

- Convergent validity with CWBS/REACH-II
- Test-retest reliability (2-week)
- Factor structure (CFA)

Key Features: To our knowledge, first publicly documented caregiver-specific SDOH instrument | Food security 1+ threshold (immediate crisis) | Portable (clinics, telehealth, programs) | CC BY 4.0 (free use with attribution) | **Requires psychometric validation (N=200+)**

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 → 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
- **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

Human Subjects: This work analyzes AI behavior on synthetic scenarios and a feasibility pilot (N=8) with adult volunteers. No clinical advice was provided by the system. Pilot participants provided written informed consent; no protected health information was collected; participants could withdraw at any time. We release scenarios and prompts with sensitive content warnings. The system includes crisis-response gating and blocks diagnosis/treatment/dosing advice consistent with applicable medical practice boundaries.

Study Framing: The October-December 2024 pilot (N=8) was conducted as commercial product testing, not human subjects research. Participants opted into a caregiving assistance service with terms of service disclosing: (1) AI system nature, (2) data usage for quality improvement, (3) right to withdraw via SMS at any time, (4) crisis escalation procedures with human review path.

Informed Consent: Maria case study participant (Section 5.4) provided explicit written consent for publication of de-identified conversation excerpts and SDOH assessment results. All identifying details (names, locations, specific dates) were anonymized or replaced with pseudonyms.

Data Handling: Conversations filtered for crisis safety with rapid escalation to human reviewers. No protected health information (PHI) released in study artifacts. Participant data retained for 2 years maximum with quarterly deletion review prompts. Users may request immediate data deletion.

PII and Memory Hygiene: GiveCare uses a sliding-window memory architecture to balance personalization with privacy. Recent messages are retained verbatim for short-term context; older conversations are compressed into domain-specific summaries (burnout trajectory, pressure zones, care routines). The system implements periodic memory rotation: historical summaries are archived and new summary generation begins from recent context. This approach minimizes long-term PII retention while preserving continuity. Time-bounded retention with automatic expiry applies across all data categories. Users can request immediate deletion at any time via SMS. Complete retention policy specification available in repository documentation. Memory hygiene is tested in InvisibleBench Tier 3 scenarios (20+ turns) and represents a key longitudinal safety dimension.

Crisis Procedures: All crisis signals triggered immediate handoff to Crisis Agent with: (1) 988 Suicide & Crisis Lifeline provision, (2) 211 local resource connection, (3) Internal alert to human moderator team with rapid response protocol during pilot operating hours.

No Clinical Claims: GiveCare is a non-clinical support system. We make no claims of therapeutic efficacy, medical diagnosis, treatment, or clinical outcomes. All effectiveness claims (attachment prevention, churn reduction, burnout trajectory detection) are stated as hypotheses requiring validation through controlled studies.

Future Research: Validation studies (N=200+) for GC-SDOH-28 psychometrics, multi-agent effectiveness, and longitudinal safety will require IRB approval before initiation. Study protocols will follow CONSORT guidelines for digital health interventions with appropriate informed consent procedures.

A.2 Data and Code Availability

Open Artifacts: GC-SDOH-28 instrument specification available at https://github.com/givecareapp/care-tools under MIT License. Benchmark evaluation framework (InvisibleBench) available at https://github.com/givecareapp/givecare-bench. Production system implementation not publicly released.

GC-SDOH-28 Instrument: Conceptual specification with domain definitions, question content, and scoring logic documented in repository (`GC-SDOH.md`). Requires psychometric validation before clinical use. Researchers may request structured instrument format for validation studies.

Pilot Data: Aggregate feasibility metrics reported in Table 3 (N=8 caregivers, 144 conversations, Oct-Dec 2024). Individual conversation logs not released to protect participant privacy. Researchers may request aggregate de-identified data for meta-analysis upon reasonable request.

Implementation Details: Architectural patterns, multi-agent orchestration logic, composite scoring formulas, and guardrail design described in this paper. Production code not publicly released; implementation details available to researchers and developers upon request for replication studies.

Reproducibility: Figures and tables in this paper generated via scripts in `/papers/givecare/` directory of the benchmark repository. Model evaluation methodology documented in InvisibleBench paper [37].

Intended Use & Limits

Intended Use: GiveCare is a reference architecture for longitudinal-safe caregiving AI research and development. It demonstrates design patterns for:

- Multi-agent orchestration to mitigate attachment risk
- SDOH-grounded support that addresses structural barriers
- Anticipatory engagement based on trajectory monitoring
- Conservative guardrails for medical-advice boundaries

NOT Intended For:

- Clinical diagnosis, treatment, or medical decision-making
- Crisis intervention (system provides referrals to 988/211, not clinical care)
- Use without appropriate validation, IRB approval, and regulatory compliance
- Deployment in jurisdictions without verifying compliance with local medical practice acts

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)
3. GC-SDOH-28 psychometric validation (N=200+; Cronbach's α, CFA, DIF, test-retest)
4. IRB approval for research use; regulatory review for commercial deployment
5. Licensed clinician oversight pathway for crisis escalation

Limitations: This is a feasibility architecture (N=8 pilot), not a validated clinical intervention. Effectiveness claims are hypotheses requiring controlled studies.

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 9: Reproducibility Card: Complete Specification for Replication

Component	Specification
Model	Cost-optimized frontier model for all agents (Main, Crisis, Assessment); see repository for specifications
Guardrails	Third-party content safety service + custom rule-based detectors (categories: diagnosis, treatment, dosing)
Resource Search	LLM with map provider grounding (physical locations); ETL pipeline (programs)
Latency	950ms median (8 caregivers, 144 conversations, Oct-Dec 2024)
Cost	$0.02-0.05 per conversation (model inference + maps API + SMS delivery)
Repository	https://github.com/givecareapp/care-tools
Deployment	Serverless backend with SMS/RCS delivery via webhooks
GC-SDOH-28	28 questions across 8 domains, requires psychometric validation (pending)

A.5 Open Artifacts

All research artifacts are publicly released to enable community validation and extension:

Table 10: Released Artifacts and Access Information

Artifact	Format	License	URL
GC-SDOH-28 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

Intended Use: Reference architecture for caregiving AI. NOT for clinical decision-making, diagnosis, treatment, or crisis intervention. System design informs deployment choices but does not replace human clinical oversight.

Prohibited Use: Using system for medical diagnosis, treatment recommendations, or crisis intervention without qualified human oversight.

B GC-SDOH-28: Full Instrument Specification

The GiveCare Social Determinants of Health instrument (GC-SDOH-28) 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 Instrument Design Rationale

GC-SDOH-28 extends patient-focused SDOH instruments (PRAPARE, AHC HRSN) to address caregiver-specific barriers:

- **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 11: GC-SDOH-28 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 Full 28-Item Question List

Financial Strain (5 items):

1. In the past month, have you worried about affording care-related expenses?
2. Have you reduced your work hours or left your job to provide care?
3. Do care-related costs strain your household budget?
4. Have you borrowed money or gone into debt for caregiving expenses?
5. Are you worried about job performance or opportunities due to caregiving?

Housing Security (3 items):

6. Are you worried about losing your housing in the next 2 months?
7. Have utility bills (heat, electricity, water) gone unpaid due to caregiving costs?
8. Does your home need repairs or modifications for safe caregiving?

Transportation Access (3 items):

9. Have you had difficulty getting your care recipient to medical appointments?
10. Do you lack reliable transportation for caregiving tasks?
11. Have transportation costs prevented you from accessing services?

Social Support (5 items):

12. Do you feel alone in your caregiving responsibilities?
13. Do family members share caregiving tasks with you?
14. Do you have someone to talk to about caregiving stress?
15. Do you feel appreciated by family for your caregiving work?
16. Are you connected to caregiver support groups or communities?

Healthcare Access (4 items):

17. In the past year, have you delayed your own medical care due to caregiving?
18. Do you have health insurance coverage for yourself?
19. Can you afford medications or treatments you need?
20. Do you have a regular healthcare provider you can see?

Food Security (3 items):

21. In the past month, did you worry about running out of food?
22. Have you skipped meals due to lack of money?

23. Do you have access to healthy, nutritious food for yourself and your care recipient?

Legal/Administrative Support (3 items):

24. Do you have legal documents in place (POA, advance directives)?
25. Do you know your rights under FMLA or job protection laws?
26. Have you experienced legal or administrative barriers in accessing care services?

Technology Access (2 items):

27. Do you have reliable internet access at home?
28. Are you comfortable using technology for telehealth or online services?

Delivery Method: Questions asked conversationally via SMS over 6–8 conversation turns. Assessment Agent chunks questions to minimize burden while maintaining context.

Validation Status: Design contribution requiring psychometric validation (N=200+) before clinical use. Pilot feedback (N=8): "Felt caregiving-specific" and "relevant." No completion rates or prevalence data collected systematically.

References

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

[2] 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/

[3] 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/

[4] 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.

[5] 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.

[6] 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.

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

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

[9] 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.

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

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

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

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

[14] 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.

[15] 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.

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

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

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

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

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

[21] 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[38] 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.

[39] 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.

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

[41] 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

[42] 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

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

C Acknowledgments

We thank the caregivers who participated in our beta deployment, sharing their experiences to 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) [15] on the GiveCare website and Dr. Susan Tebb for permission to use the Caregiver Well-Being Scale (CWBS) [13, 14] in the GiveCare application.

We thank Hamel Hussain for guidance on evaluation-driven development and the AARP 2025 Caregiving in the U.S. report for empirical grounding. This work builds on trauma-informed principles from SAMHSA [41], Chayn [42], and *Designed with Care* [43], as well as InvisibleBench [37] and YAML-driven scoring [40] frameworks.