

# Real-time safety screening for AI companion apps

Companion apps are one bad conversation from a lawsuit or app-store ban - minors matched with romantic content, users spiraling in late-night chats - and keyword moderation misses the slow relationship-arc escalation that defines the medium.

Real-time safety screening for AI companion apps should be tested as a narrow first-win workflow for Founder or head of trust & safety at an AI companion app.

HIGH DIFFICULTY

USAGE-BASED API PRICING PER CONVERSATION-HOUR SCREENED, WITH COMPLIANCE REPORTING AS AN ENTERPRISE TIER.

55/100

VALIDATION VERDICT / RESEARCH

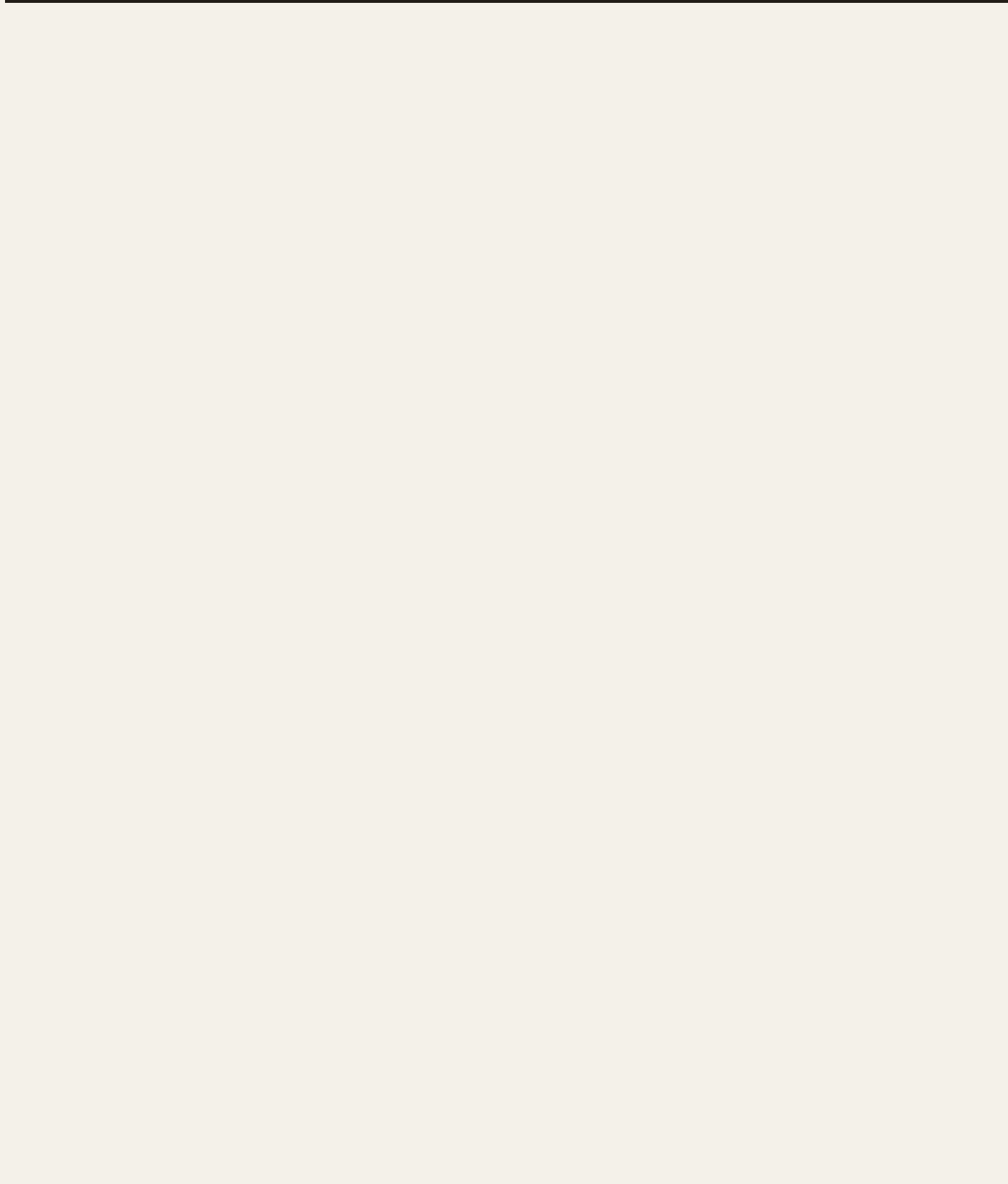
Validation is a weighted rubric, not a guarantee. Use the next validation step before building.

|            |                            |
|------------|----------------------------|
| Confidence | 54%                        |
| Lifecycle  | Heating                    |
| Timing     | 54/100                     |
| Rubric     | INAV-VALIDATION-2026-06-04 |

HEATING

Watch window

|                        |        |
|------------------------|--------|
| Demand signal          | 5.5/10 |
| Problem severity       | 6.3/10 |
| Willingness to pay     | 5/10   |
| Competitive saturation | 6.3/10 |
| Feasibility            | 4/10   |



#### VERDICT

## Research • 55/100

Real-time safety screening for AI companion apps should be tested as a narrow first-win workflow for Founder or head of trust & safety at an AI companion app.

#### THIS WEEK'S TEST

Integrate with two mid-size companion apps under NDA, replay their historical incidents through the screen, and measure catch-rate on the cases their own moderation missed.

#### KILL IT IF

Fewer than five qualified buyers agree to discuss the workflow after targeted outreach.

#### — READER DEMAND SIGNAL

## Would you build this? Would you pay?

One tap each — anonymous, one vote per question per day. Tallies update as readers weigh in.

### Would you build this?

Be the first to weigh in

### Would you pay for this?

Be the first to weigh in

■ ■ ■

TRUST & SAFETY INFRASTRUCTURE FOR CONSUMER AI

SIGNAL MODEL

Real-time safety screening for AI companion apps

Real-time safety screening for AI companion apps should be tested as a narrow first-win workflow for Founder or head of trust & safety at an AI companion app.

| Index | Signal Value (Relative) |
|-------|-------------------------|
| 1     | Baseline                |
| 2     | Baseline                |
| 3     | Slightly Above Baseline |
| 4     | Slightly Below Baseline |
| 5     | Slightly Above Baseline |
| 6     | Slightly Above Baseline |
| 7     | Peak                    |
| 8     | Drop                    |

VALIDATION

55/100

Research

CONFIDENCE

54%

Editorial confidence

SCORE AVG

5.8/10

Scorecard average

PROOF

5.8/10

Proof signal average

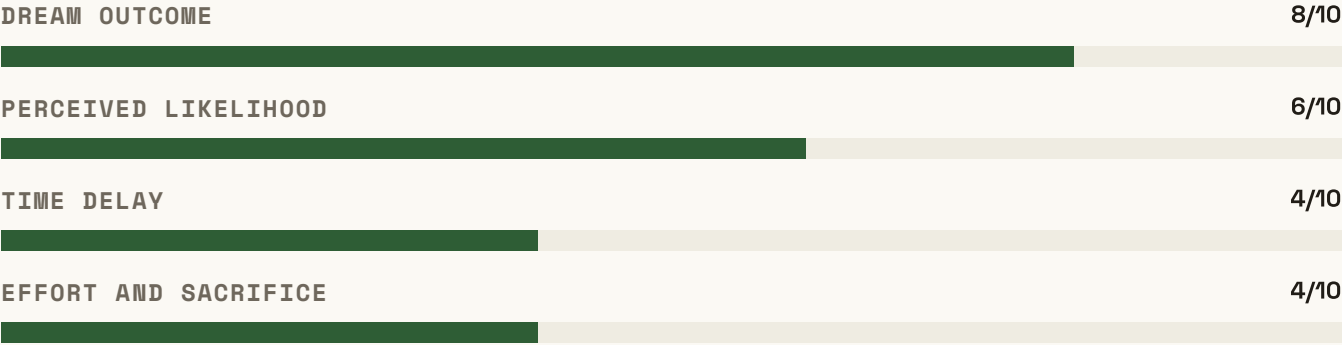
SCORE RADAR

Decision balance



VALUE EQUATION

Offer strength



MARKET MAP

Category king candidate

CUSTOMER VALUE



High value plus high uniqueness deserves deeper research; lower uniqueness requires a clear distribution advantage.

VALIDATION FUNNEL

From pain to product.

|   |                                                                                                                   |        |
|---|-------------------------------------------------------------------------------------------------------------------|--------|
| 1 | <b>Buyer pain</b><br>Founder or head of trust & safety at an AI companion app                                     | 5.4/10 |
| 2 | <b>Concierge proof</b><br>Integrate with two mid-size companion apps under NDA, replay their historical inc...    | 5.8/10 |
| 3 | <b>Paid wedge</b><br>Concierge review or paid template                                                            | 5.5/10 |
| 4 | <b>Repeatable product</b><br>Usage-based API pricing per conversation-hour screened, with compliance reporting... | 5.9/10 |

EVIDENCE HEATMAP

Signal intensity.

|                                                   |                                              |
|---------------------------------------------------|----------------------------------------------|
| <b>WHY NOW</b><br>5/10<br>Demand visibility       | <b>WHY NOW</b><br>4/10<br>Tooling readiness  |
| <b>WHY NOW</b><br>4/10<br>Budget clarity          | <b>WHY NOW</b><br>7/10<br>Competitive window |
| <b>PAIN</b><br>5/10<br>Repeated workflow friction | <b>MONEY</b><br>4/10<br>Budget hypothesis    |

URGENCY

6/10

Switching pressure

DISTRIBUTION

8/10

Reachable buyer language

# Window opening (54/100): demand is rising while saturation is still manageable.

Deterministic stage assignment from re-check status, demand signals, complaint echo, and competitive saturation.

54/100

HEATING

10 trend-discovery signals match this idea.

1 matched company signal raise saturation.

Demand

75/100

Not old enough for a 30-day re-check yet.

Saturation

30/100

1 funded signal across 1 matched competitor signal.

Complaint echo

22/100

Matched adoption substrate is up 576%.

— VALIDATION ENGINE

Evidence-backed idea-validation score.

The score uses a versioned 2026 rubric across demand, problem severity, willingness to pay, competitive saturation, and feasibility.

55/100

Research

Research is the current validation verdict: problem severity is the strongest signal, while feasibility is the main evidence gap to close before scaling the build.

Rubric version: INAV-VALIDATION-2026-06-04 / generated August 15, 2026

Demand signal

5.5/10

24% WEIGHT

Demand looks thin because the report has 2 source-backed signal(s), an editorial confidence of 54/100, and a defined buyer in Trust & safety infrastructure for consumer AI.

- High-profile lawsuits allege companion chatbots contributed to self-harm by minors, and platforms have pulled apps over safety screenshots going viral.
- Target buyer: Founder or head of trust & safety at an AI companion app

Problem severity

6.3/10

22% WEIGHT

Problem severity is thin when the buyer pain, customer value, and dream-outcome scores are combined.

- Companion apps are one bad conversation from a lawsuit or app-store ban - minors matched with romantic content, users spiraling in late-night chats - and keyword moderation misses the slow relationship-arc escalation that defines the medium.
- High-profile lawsuits allege companion chatbots contributed to self-harm by minors, and platforms have pulled apps over safety screenshots going viral.

## Willingness to pay

5/10

20% WEIGHT

Willingness to pay is weak; the model has a monetization hypothesis, but it must still be proven through paid pilots or explicit pricing objections.

- Usage-based API pricing per conversation-hour screened, with compliance reporting as an enterprise tier.
- Integrate with two mid-size companion apps under NDA, replay their historical incidents through the screen, and measure catch-rate on the cases their own moderation missed.

## Competitive saturation

6.3/10

18% WEIGHT

No source-backed direct match is recorded yet, so saturation risk is treated as unknown rather than proof of novelty.

- Existing-product check has no named direct match.
- Competitive score rewards a narrow wedge, not absence of research.

## Feasibility

4/10

16% WEIGHT

Feasibility is weak for a high build if the MVP is limited to the first measurable workflow.

- Integrate with two mid-size companion apps under NDA, replay their historical incidents through the screen, and measure catch-rate on the cases their own moderation missed.
- Foundation-model providers could bundle companion-grade safety into their APIs.

## Next validation step

Integrate with two mid-size companion apps under NDA, replay their historical incidents through the screen, and measure catch-rate on the cases their own moderation missed.

# Seven days to a build / kill decision.

Derived from this report's own validation test, channels, offers, and kill criteria.

Each day has a threshold, so the week ends in a decision instead of a feeling.

## DAY 1

### Build the buyer list

List 50-100 named founder or head of trust & safety at an ai companion app prospects from Community pain posts and Direct outreach — names, not categories.

**Threshold:** 50+ named, reachable buyers on the list.

## DAY 2

### Join the watering holes

Join and observe Reddit / forums, Launch communities, Review and alternative pages. Collect the exact words buyers use for this pain.

**Threshold:** 10+ verbatim pain quotes captured.

## DAY 3

### Send first outreach

Send the cold outreach template (below) to 15 buyers from the day-1 list, personalized with one detail each.

**Threshold:** 15 sent; 3+ replies of any kind.

## DAY 4

### Run buyer interviews

Hold 15-minute calls using the interview script (below). Listen for current workarounds and what they cost.

**Threshold:** 3+ completed interviews.

## DAY 5

### Run the report's validation test

Integrate with two mid-size companion apps under NDA, replay their historical incidents through the screen, and measure catch-rate on the cases their own moder...

**Threshold:** Problem resonance: 5+ calls or 10+ detailed replies.

## DAY 6

### Make the smoke offer

Offer "Concierge review or paid template" at \$19-\$99 to every interviewed buyer. Manual delivery is fine — payment is the signal.

**Threshold:** 1+ pre-commitment (payment, signed LOI, or scheduled paid pilot).

## DAY 7

### Decide against the kill criteria

Score the week against this report's kill criteria, then take the stated next validation step: Integrate with two mid-size companion apps under NDA, replay their historical incidents through the screen, and measure catch-rate on the cases their own moder...

**Threshold:** A written build / keep-testing / kill decision.

### Pass signal

Pass: thresholds on days 3, 4, and 6 are met — proceed to the next validation step with real buyer language in hand.

### Fail signal

Kill or rethink if the week confirms: Fewer than five qualified buyers agree to discuss the workflow after targeted outreach.

# Decision scorecard.

The report is structured to force a yes, no, or test decision instead of leaving the reader with a loose brainstorm.

## Opportunity

5/10

PROMISING

Real-time safety screening for AI companion apps has an editorial confidence score of 54/100 before live buyer validation.

## Problem

5/10

PROMISING

Companion apps are one bad conversation from a lawsuit or app-store ban - minors matched with romantic content, users spiraling in late-night chats - and keyword moderation misses the slow relationship-arc escalation that defines the medium.

## Feasibility

4/10

NEEDS PROOF

A high build can work if the MVP stays limited to the first repeated workflow.

## Why now

9/10

EXCEPTIONAL

Litigation and state attorney-general scrutiny of companion apps intensified through 2025-2026, and app stores now demand demonstrable safeguards, creating compelled demand for a compliance layer.

# Business fit and offer ladder.

## Revenue potential

\$250K-\$2M ARR potential if the wedge proves budget urgency and becomes a recurring workflow.

## Execution difficulty

Execution is high; the main constraint is staying narrow enough for a first proof loop.

## Go-to-market

Start with manual concierge output, direct outreach, and community proof before paid acquisition.

## Founder fit

Best for an AI-assisted solo founder who can interview the buyer and ship a focused first version quickly.

### 1. Lead magnet

## Real-time Safety Screening For Ai Companion Apps checklist

**Free**

Helps Founder or head of trust & safety at an AI companion app audit the painful workflow before buying software.

Capture qualified leads and learn the buyer's exact language.

### 2. Frontend offer

## Concierge review or paid template

**\$19-\$99**

Delivers the first useful output manually before automation is trusted.

Validate urgency, workflow fit, and willingness to pay.

### 3. Core offer

## Real-time safety screening for AI companion apps focused SaaS

**\$49-\$499/month**

Turns the recurring manual workflow into a repeatable product loop.

Create the recurring revenue product after the narrow wedge survives tests.

4. Continuity

**Monitoring, benchmarks, and monthly reporting**

**\$99-\$1,000/year add-on**

Keeps the buyer engaged with ongoing proof, saved time, or reduced risk.

Increase retention and make the product part of a routine.

5. Backend offer

**Done-with-you setup, agency, or team rollout**

**Custom**

Adds implementation help, integrations, and workflow migration.

Capture higher-value accounts once the productized wedge is proven.

## Price-anchored revenue scenarios.

Derived from this report's "Core offer" offer-ladder stage (\$49-\$499/month). These are price-anchored scenarios, not market-size claims.

### Proof

**\$490-\$4,990 MRR****10 CUSTOMERS**

Ten paying customers proves willingness to pay and funds continued validation.

### Wedge

**\$2,450-\$24,950 MRR****50 CUSTOMERS**

Fifty customers in one niche makes the workflow the default in that circle and feeds referrals.

### Vertical leader

**\$12,250-\$124,750 MRR****250 CUSTOMERS**

A few hundred accounts in one vertical is a real business before any horizontal expansion.

### Break-even

At \$49-\$499/month, 1 customers cover the stated Local-first MVP budget: \$0-\$10K before paid acquisition. budget within a month; fewer if they land at the top of the range.

### Sizing the buyer universe

Size the buyer universe in one day: count founder or head of trust & safety at an ai companion app reachable through the report's channels (directories, associations, communities) until the list stops growing — the test only needs the first 100 names, not a TAM estimate.

### Pricing benchmark

No public look-alike products were recorded in this report, so price against the manual workaround's time cost, not against software.

# Mixed reflexivity — execution over secrecy.

Does publishing this opportunity strengthen it or crowd it? low confidence.

**Publish or protect**

No dominant reflexivity signal. Publish the analysis, but note that execution speed and distribution matter more than secrecy here; revisit if the space shows saturation.

**Signals**

No dominant network or scarcity marker — a mixed case where execution speed and distribution matter more than secrecy.

# Why now and proof signals.

## Why now

5/10

### Demand visibility

High-profile lawsuits allege companion chatbots contributed to self-harm by minors, and platforms have pulled apps over safety screenshots going viral.

Build only if the complaint repeats across interviews, posts, or existing workflow artifacts.

[commonsensemedia.org](https://commonsensemedia.org)

4/10

### Tooling readiness

AI-assisted product work and managed infrastructure reduce the first-version cost.

The first release should automate one high-friction step rather than become a broad platform.

[en.wikipedia.org](https://en.wikipedia.org)

4/10

### Budget clarity

Usage-based API pricing per conversation-hour screened, with compliance reporting as an enterprise tier.

Ask for money during validation before building the full workflow.

[commonsensemedia.org](https://commonsensemedia.org)

7/10

### Competitive window

The wedge is specific enough to test without claiming the whole market.

Position around one buyer and one measurable first-win outcome.

[commonsensemedia.org](https://commonsensemedia.org)

## Proof signals

5/10

### Pain: Repeated workflow friction

High-profile lawsuits allege companion chatbots contributed to self-harm by minors, and platforms have pulled apps over safety screenshots going viral.

[commonsensemedia.org](https://commonsensemedia.org)

4/10

## Money: Budget hypothesis

Founder or head of trust & safety at an AI companion app is the first group to test because the monetization path is: Usage-based API pricing per conversation-hour screened, with compliance reporting as an enterprise tier. [commonsensemedia.org](https://commonsensemedia.org)

6/10

## Urgency: Switching pressure

Urgency becomes real only if the current workaround costs time, risk, money, or reputation every week. [en.wikipedia.org](https://en.wikipedia.org)

8/10

## Distribution: Reachable buyer language

The first channel should be whichever source lane already contains the buyer's vocabulary.

[commonsensemedia.org](https://commonsensemedia.org)

— POSITIONING

# Market gaps and execution plan.

## Underserved segments

- Founder or head of trust & safety at an AI companion app who still run the workflow in spreadsheets, generic docs, email, or chat threads.
- Small teams in Trust & safety infrastructure for consumer AI that feel the pain weekly but are too narrow for broad incumbents.
- New adopters who need guided proof before committing to a larger platform.

## Feature gaps

- A narrow workflow that reaches value without configuration-heavy onboarding.
- A buyer-facing proof artifact that shows time saved, risk reduced, or communication improved.
- A handoff path from manual concierge service to repeatable software.

## Differentiation levers

- Use specificity as the wedge: one buyer, one workflow, one measurable result.
- Show proof earlier than broad competitors with before-and-after examples and small pilot data.
- Keep implementation lighter than incumbent suites or generic AI assistants.

## Execution snapshot

|                                                                                                                                                                     |                                                            |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------|
| Type                                                                                                                                                                | Consumer app product                                       |
| Timeline                                                                                                                                                            | 8-12 weeks                                                 |
| Budget                                                                                                                                                              | Local-first MVP budget: \$0-\$10K before paid acquisition. |
| Initial offer                                                                                                                                                       | Concierge review or paid template                          |
| Build only the first-win workflow for "Real-time safety screening for AI companion apps" and keep research, setup, and exceptions manual until the wedge is proven. |                                                            |

**Weekly**

## Community pain posts

Use communities and forums where Founder or head of trust & safety at an AI companion app already describe the painful workflow.

Problem teardown, interview ask, and short demo clip / 5 qualified calls or 10 detailed replies in 7 days

**Daily during validation**

## Direct outreach

Direct conversations are the fastest way to verify budget ownership and switching cost.

Concierge pilot offer with a manually prepared sample / 3 paid pilots, LOIs, or budget-owner follow-ups

**Bi-weekly**

## Searchable comparison content

Alternative and comparison pages reveal objections, pricing language, and buying intent.

Before-and-after page or alternatives memo for the exact workflow / Organic clicks, booked demos, or waitlist joins from comparison intent

**Once MVP is clickable**

## Launch directory

Launches test whether the promise is legible to people outside the first interview set.

Single-purpose demo and first-win story / 25% demo completion or 10 waitlist joins

|                                                                                                                                                                                                                                                                                                                                  |                                                                                                                                                              |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------|
| COMPETITIVE LANDSCAPE                                                                                                                                                                                                                                                                                                            |                                                                                                                                                              |
| Alternatives, incumbents, and whitespace.                                                                                                                                                                                                                                                                                        |                                                                                                                                                              |
| This section names likely workarounds and public players so the report can argue where the wedge is still open.                                                                                                                                                                                                                  |                                                                                                                                                              |
| Real-time safety screening for AI companion apps should be positioned against generic AI assistants, no-code workarounds, and any vertical incumbent that already owns Trust & safety infrastructure for consumer AI. The opening is a narrower first-win workflow for Founder or head of trust & safety at an AI companion app. |                                                                                                                                                              |
| WORKAROUND                                                                                                                                                                                                                                                                                                                       | <div>Notion</div> <div>Workspace and documentation</div> <div>Competes when buyers can solve the pain with templates, checklists, and shared pages.</div>    |
| DIRECT                                                                                                                                                                                                                                                                                                                           | <div>Clio</div> <div>Legal practice management</div> <div>Relevant to legal operations, records, intake, and compliance workflows.</div>                     |
| WORKAROUND                                                                                                                                                                                                                                                                                                                       | <div>Airtable</div> <div>No-code database</div> <div>Competes when the first version can be modeled as a lightweight database and workflow view.</div>       |
| ADJACENT                                                                                                                                                                                                                                                                                                                         | <div>Microsoft 365 Copilot</div> <div>Office workflow assistant</div> <div>Competes inside Word, Excel, PowerPoint, Outlook, and enterprise workflows.</div> |
|                                                                                                                                                                                                                                                                                                                                  |                                                                                                                                                              |

## WORKAROUND

### Asana

Project management

Competes where the buyer can express the workflow as tasks, owners, and due dates.

### Whitespace

- A narrow workflow that reaches value without configuration-heavy onboarding.
- A buyer-facing proof artifact that shows time saved, risk reduced, or communication improved.
- A handoff path from manual concierge service to repeatable software.
- Use specificity as the wedge: one buyer, one workflow, one measurable result.
- Show proof earlier than broad competitors with before-and-after examples and small pilot data.
- Keep implementation lighter than incumbent suites or generic AI assistants.
- Own the specific buyer workflow instead of selling a broad AI assistant.

### Positioning moves

- Lead with the exact buyer: Founder or head of trust & safety at an AI companion app.
- Show a proof artifact for: Integrate with two mid-size companion apps under NDA, replay their historical incidents through the screen, and measure catch-rate on the cases their own moderation missed.
- Name the generic-assistant workaround directly and explain what it misses.
- Offer concierge setup before promising a full platform.

Public source

**Notion**

<https://www.notion.com/>

Public source

**Clio**

<https://www.clio.com/>

Public source

**Airtable**

<https://www.airtable.com/>

Public source

Microsoft

<https://www.microsoft.com/en-us/microsoft-365/copilot>

Public source

Asana

<https://asana.com/>

Public source

Report source

<https://www.common sense media.org/>

Public source

Report source

<https://en.wikipedia.org/wiki/Replika>

Who’s already moving in Software & AI

Public companies and funding signals the intelligence graph links to this vertical (related by keyword overlap — sized players, not direct competitors). Source: [/graph.json](#) .

FIELD SERVICE MANAGEMENT

\$625M

ServiceTitan

Operations software for contractors and field-service trades: scheduling, dispatch, quotes, jobs, and crew management.

IPO · 2024-12-12



## Execution-readiness scorecard.

The score turns the report into bottlenecks, accelerators, and a dated first-month launch plan.

44/100

### Research first

Real-time safety screening for AI companion apps scores 44/100 for execution readiness. The recommended next step is Integrate with two mid-size companion apps under NDA, replay their historical incidents through the screen, and measure catch-rate on the cases their own moderation missed.

Execution scorecard is generated from report validation, confidence, feasibility, founder fit, and difficulty.

### Bottlenecks

- Foundation-model providers could bundle companion-grade safety into their APIs.
- Serving this market means absorbing part of its legal and reputational tail risk.
- A broad AI assistant can flatten differentiation unless the wedge is painfully specific.
- The first release can become a generic dashboard if the job is not named tightly.
- Needs real buyer access, not only desk research.

### First milestones

- 2026-08-15: Frame the wedge
- 2026-08-18: Interview 10 people who match the buyer persona.
- 2026-08-22: Ship a clickable demo or concierge workflow that produces the first useful artifact.
- 2026-08-29: Run one paid pilot or collect explicit pricing objections before automating the rest.

## Value equation, matrix, and ACP.

# Fit, roast, and kill criteria.

## 6/10

### Founder fit

A solo or AI-assisted founder with direct access to Founder or head of trust & safety at an AI companion app.

### ADVANTAGES

- Can talk to the buyer before writing much code.
- Can ship a narrow first-win demo quickly.
- Can use local-first research artifacts to keep validation moving without a large team.

### GAPS

- Needs real buyer access, not only desk research.
- Needs proof of budget or repeated urgency.
- Needs a crisp wedge before broad product work starts.

### Roast

Interesting hypothesis, but it needs sharper demand evidence before build time.

### BLIND SPOTS

- Foundation-model providers could bundle companion-grade safety into their APIs.
- A broad AI assistant can flatten differentiation unless the wedge is painfully specific.
- The first release can become a generic dashboard if the job is not named tightly.

### HARD QUESTIONS

- Who wakes up already trying to solve this?
- What do they stop paying for or stop doing when this works?
- What proof would make a skeptical buyer trust it in one screen?
- What is the smallest paid version of this idea?

### Kill criteria

- Fewer than five qualified buyers agree to discuss the workflow after targeted outreach.
- No buyer can name a current cost in time, money, risk, or reputation.
- The first demo does not produce a clear next step, paid pilot, or specific objection.

## Next actions

- Write the one-sentence promise and test it in the strongest channel.
- Create the lead magnet and use it to recruit interviews.
- Build the smallest demo that proves the first win.

# Move from reading to testing.

Local-first handoff cards copy prompts or structured data without requiring an account.

BUILD THIS IDEA

Copy the focused build brief for a coding agent.

COPY

ROAST

Copy the critique lens and blind spots before committing time.

COPY

LANDING PAGE

Copy a landing-page brief based on buyer, pain, and validation.

COPY

BRAND PACKAGE

Copy positioning inputs for naming, messaging, and design direction.

COPY

## AD CREATIVES

Copy campaign angles for buyer-problem validation.

**COPY**

## EXPORT DATA

Copy structured JSON for a research engine, roadmap tracker, or another agent.

**COPY**

## FOUNDER FIT

Copy the founder-fit self-check before entering build mode.

**COPY**

# Outreach template and interview script.

Built from this report's buyer, pain language, and channels. Personalize one detail per message — these are starting points, not spam ammunition.

## Cold outreach message

QUESTION ABOUT REAL WORKFLOW

HOW ARE YOU HANDLING COMPANION APPS ARE ONE BAD CONVERSATION FROM A LAWSUIT OR A...

15 MINUTES ON A TRUST & SAFETY INFRASTRUCTURE FOR CONSUMER AI WORKFLOW?

Hi {{firstName}},

I'm researching how founder or head of trust & safety at an ai companion app handle this today: Companion apps are one bad conversation from a lawsuit or app-store ban - minors matched with romantic content, users spiraling in late-nig...

I'm not selling anything yet — I'm testing whether "Real-time safety screening for AI companion apps" is worth building, and I'd rather learn from people living the workflow than guess.

Would you trade 15 minutes for first access (and a say in what gets built) if it goes ahead?

{{yourName}}

**COPY MESSAGE**

## Buyer interview script

1. Walk me through the last time this happened: Companion apps are one bad conversation from a lawsuit or app-store ban - minors matched with romantic content, users s... What did you actually do?
2. What does that workaround cost you — in hours, money, or risk — in a normal month?
3. What have you already tried or bought to fix it, and why didn't it stick?
4. If "An API that scores each conversation turn in context of the full relationship history for self-harm..." existed, what would have to be true for you to switch in the first week?
5. Who else feels this worse than you do — and would you introduce me?

## WHERE TO SEND IT

- Community pain posts — Problem teardown, interview ask, and short demo clip
- Direct outreach — Concierge pilot offer with a manually prepared sample
- Searchable comparison content — Before-and-after page or alternatives memo for the exact workflow
- Reddit / forums — Post a problem teardown for Trust & safety infrastructure for consumer AI and ask how people solve it today.
- Launch communities — Ship a narrow demo and watch which promise gets clicks.

## Build and review prompts.

### Build prompt

Build a narrow MVP for "Real-time safety screening for AI companion apps" for Founder or head of trust & safety at an AI companion app. Preserve the evidence, build only the first-win workflow, include source links, and treat Integrate with two mid-size companion apps under NDA, replay their historical incidents through the screen, and measure catch-rate on the cases their own moderation missed. as the first acceptance gate.

### Review prompt

Review the "Real-time safety screening for AI companion apps" MVP for over-breadth, unsupported claims, weak buyer proof, privacy risk, and missing validation instrumentation. Do not approve expansion until the kill criteria and success metrics are measurable.

[organization / commonsensemedia.org](https://commonsensemedia.org)

#### **Common Sense Media**

Advocacy organization whose AI-companion risk assessments have shaped the regulatory pressure on the category.

[encyclopedia / en.wikipedia.org](https://en.wikipedia.org)

#### **Replika - Wikipedia**

The best-documented companion app, including the safety controversies that illustrate the liability this layer addresses.

## If this exact wedge isn't yours, these are adjacent.

Derived deterministically from this report's buyers, vertical language, and business model.

### Same problem, different buyer: Budget owner who feels the operational cost of the broken workflow.

The workflow pain in this report is not exclusive to founder or head of trust & safety at an ai companion app. Budget owner who feels the operational cost of the broken workflow. faces the same friction with their own budget and urgency.

**First test:** Re-run day 3 of the sprint (15 outreach messages) against this buyer only, and compare reply rates before changing anything else.

### Same workflow, adjacent vertical: Legal, Risk & Compliance

This report's language already overlaps Legal, Risk & Compliance (law practices). The same first-win workflow usually transfers with new vocabulary and one changed integration.

**First test:** Rewrite the one-line promise for a Legal & Risk buyer and test it in that vertical's channels before building anything new.

Open that vertical's brief

### Same wedge, alternate model: a data or monitoring product (alerts, benchmarks, watchlists)

This report monetizes via "Usage-based API pricing per conversation-hour screened, with compliance reporting as an enterprise tier.". If buyers will not change their workflow, selling them awareness of the problem is the lower-friction wedge.

**First test:** Offer both versions on day 6 of the sprint and let the first pre-commitment choose the model.

# Where this report sits in the intelligence graph.

Links from the ontology layer. Declared links are explicit in the research record; inferred links are keyword overlap and labeled as such. Full graph at /graph.json.

EVIDENCE INDEPENDENCE 87/100

8 source domains, 26 evidence edges. Dominant family: local:data/regulatory-clock.json.  
Audit all provenance .

## Complaint evidence

- Privacy, trust, and data-control anxiety  
— keyword overlap (trust, users)

## Adjacent verticals

- Legal, Risk & Compliance

## Related reports (shared keywords)

- Real-time business-closure alerts for asset buyers — real ai, real workflow, time automation, time validation

— IN THIS VERTICAL

Software, AI & Developer Tooling

Ranked 30 of 35 by validation score among published Software, AI & Developer Tooling reports.

VALIDATE · 79/100

AI workflow reliability monitor for small teams

AI operations

OPEN REPORT

VALIDATE · 78/100

AI operations signal monitor: Amazon CEO’s talks with U.S. officials triggered crackdown on Anthropic models

AI operations

OPEN REPORT

VALIDATE · 78/100

AI operations signal monitor: AMD acquires Taalas to boost inference performance by etching models in silicon

AI operations

OPEN REPORT

SHARED TAGS

Vendor insurance certificate tracker for property managers

Accessibility issue triage board for small websites

Data processing agreement tracker for micro SaaS teams

## — FULL NARRATIVE