
swAIpe: A Microlearning Retention Platform for Cognitive AI Defense

Tang Anh Quy
Ai Soph

Nguyen Thi Cam Luong
Ai Soph

Yan Yee Wong
University of Edinburgh

With
Apart Research

Abstract

AI is rapidly amplifying both beneficial capabilities and scalable harms, while the public often underestimates the risks and overestimates resilience to manipulation (Köbis et al., 2021). We present a **mobile-first micro-learning web application** that treats strengthening the public’s thinking and awareness as a core line of defense, not an afterthought (Sebastian Bay et al., 2019). Instead of passive ‘AI safety explainers’, the app channels doom-scrolling into **short, retrieval-based learning loops**, each backed by **automated fact-checking**, that train users to **recognize, question, and counter AI-driven manipulation** and misinformation (Chu-Ke & Dong, 2024).

Creators feed source links or topics into an integrated AI copilot that scripts, validates, and renders 30–60 second Remotion clips, each focused on a single concept (e.g., hallucinations, prompt injection, or cognitive bias exploits). A fact-validation engine is embedded in the publishing pipeline, verifying key claims against authoritative sources and retrieval-augmented context before clips can go live.

Learners consume these clips through a TikTok-style mobile feed where every swipe triggers an active-recall prompt, spaced-repetition scheduling, and metrics aimed at improving knowledge durability (Carpenter et al., 2022) rather than watch time.

Our core defense thesis is that **an informed, critically literate public is itself a form of defensive infrastructure**: people who can recognize AI-generated misinformation, understand failure modes, and circulate vetted knowledge exert bottom-up pressure on platforms, policymakers, and institutions (Pilditch et al., n.d.). We present the design and implementation of this mobile-first system—from ingestion → scripting → fact-validation → rendering → retrieval-based playback →

spaced review—alongside key architectural choices, security considerations, limitations, and areas for future expansion.

Keywords: cognitive defense, AI literacy, misinformation, public education, spaced repetition, fact validation

Introduction

Most people now interact with AI through opaque systems-recommender feeds, search summaries, generative media, and “smart” assistants. Empirically, they:

- Underestimate how easily AI can generate convincing false content
- Receive little repeated exposure to concepts like hallucinations, data poisoning, or persuasion-driven cognitive biases

This creates a cognitive asymmetry: attackers can cheaply scale AI-driven manipulation while the public gets little structured defence training. Traditional formats—long courses, PDFs, one-off explainers—don’t match attention patterns shaped by TikTok, Reels, and Shorts.

This form factor—rapid, swipe-based, ultra-short videos tuned for instant engagement—has been shown to strongly capture user attention (Zannettou et al., 2024). Repurposing that same attention-optimized interface offers an opportunity to deliver high-frequency, lightweight learning that strengthens cognitive resilience.

Our project asks: **Can we repurpose those addictive interaction patterns (vertical scroll, bite-sized clips) into a defensive micro-learning system that measurably increases AI literacy and cognitive resilience?**

Our **def/acc** path: Cognitive Resilience

Within the def/acc frame, we focus on **cognitive defense** as a complement to “hard” technical measures like biosurveillance or red-teaming. Generative models have expanded the **attack surface to the public mind**: influence operations, synthetic experts, and AI-shaped discourse exploit a population that often cannot distinguish credible from dubious outputs. Strengthening public literacy closes this gap while creating a **policy feedback loop**, as informed citizens are more likely to demand sane AI governance and support defensive infrastructure.

Our approach uses **trusted human creators**—initially our team and vetted AI safety partners—to produce validated, pedagogically sound micro-lessons that propagate through existing social distribution channels. All content undergoes mandatory human review, ensuring accuracy and instructional quality as the creator network expands.

This project contributes:

1. **A microlearning defense mobile-first web application** that repurposes vertical-video attention patterns into retrieval practice, spaced repetition, and mastery tracking for cognitive-security concepts.
2. **A fact-validation engine** in the publishing pipeline, reducing the risk of inaccurate or misleading educational content.
3. **Retention-focused metrics and telemetry**, emphasizing recall and interval fidelity over traditional engagement metrics.
4. **A theory of change** (Vogel, 2012) linking individual learning to improved public discourse, policy pressure, and societal resilience.

Together, these elements form an **accessible, technology-driven cognitive defense layer** that can be applied to domains such as AI literacy, biosecurity basics, or phishing awareness—while using the same underlying defensive learning engine.

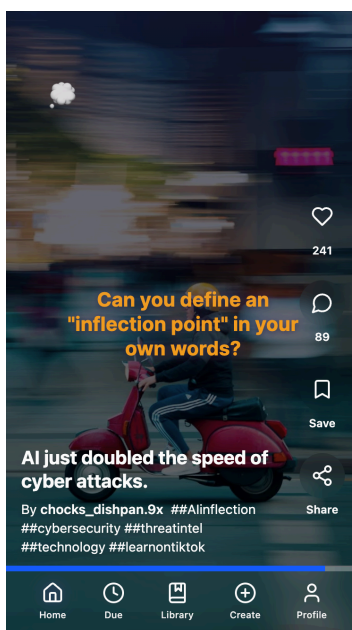


Figure 1: A screenshot of our swAIpe web app

Methods & Results

We implemented and designed a web application architecture with:

- **Client:** A mobile-first experience for learners, with a web-based creator interface. The documented UX centers on a vertical home feed, a “Due” tab for scheduled reviews, a Library, a Create flow, and a Profile area.
- **Backend:** Node/TypeScript/Python services (via Supabase functions and cloud container services) orchestrating clip generation, content validation, learning scheduling and analytics.

- **Rendering:** A Remotion-backed engine generating 9:16 vertical clips (1080×1920, 30 FPS) with brand-safe overlays and subtitles.
- **Data:** Postgres + **pgvector** for embeddings and semantic operations; object storage for media assets.

The pipeline is organized into four main stages: **Ingestion, Content Analysis and Clip Generation, Concept Verification (Fact Validation Engine)** and **Learning Engine**.

Below takes us through the pipeline, with screenshots of the interactions and the methodology behind its implementation.

1. Ingestion

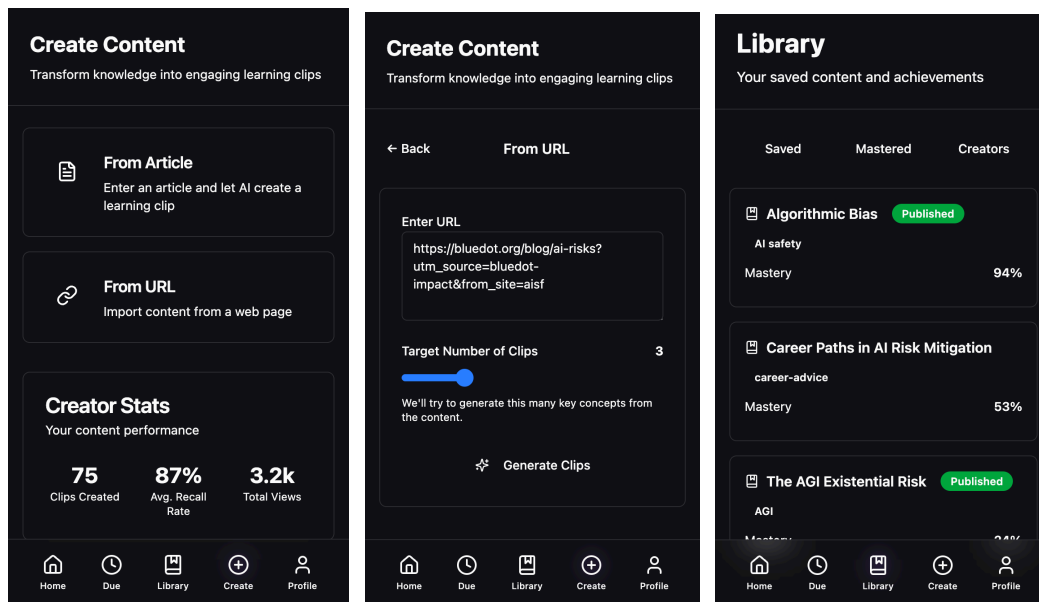


Figure 2: Screenshots of the app's content-creation interface and creator dashboard

When a creator browses digital content online and would like to create content from what they have read, they can submit either through:

- Text (for example, an essay, a blog post)
- URLs (articles, blog posts, reputable reports).

to the app.

2. Content Analysis and Clip Generation

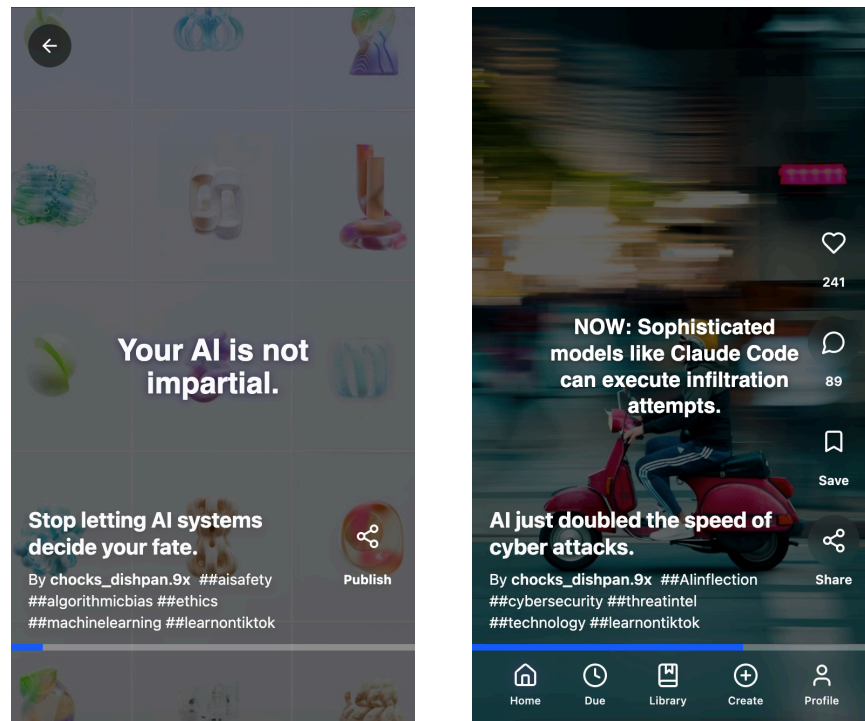


Figure 3: Sample Remotion clips and scripts generated for microlearning content

An LLM then reads the given content and generates question answers, thought-provoking statements in a clip format based on the learning point given:

1. Template-Based Script Generation and Rendering

We define a catalog of pedagogical templates tuned for micro-learning:

- **Define + Hook, Flash-Quiz, Example Variation, Recap Bundle, Mistake → Correction, Analogy Story**, and others.
- Each template encodes:
 - Intended learning outcome.
 - Timing allocations per segment (for example, 0–3 seconds hook, 3–8 seconds explanation, 8–18 seconds example).
 - Retrieval moments where the learner is prompted to recall or apply.

2. Rendering of clips

- Remotion components build the visual clip of less than 60s. See Figure 2 for screenshots.
- Background images come from public sources like [Pexels.com](https://www.pexels.com/), [Pixabay.com](https://www.pixabay.com/) instead of AI generation to respect artist rights.

3. Concept Verification (Fact Validation Engine)

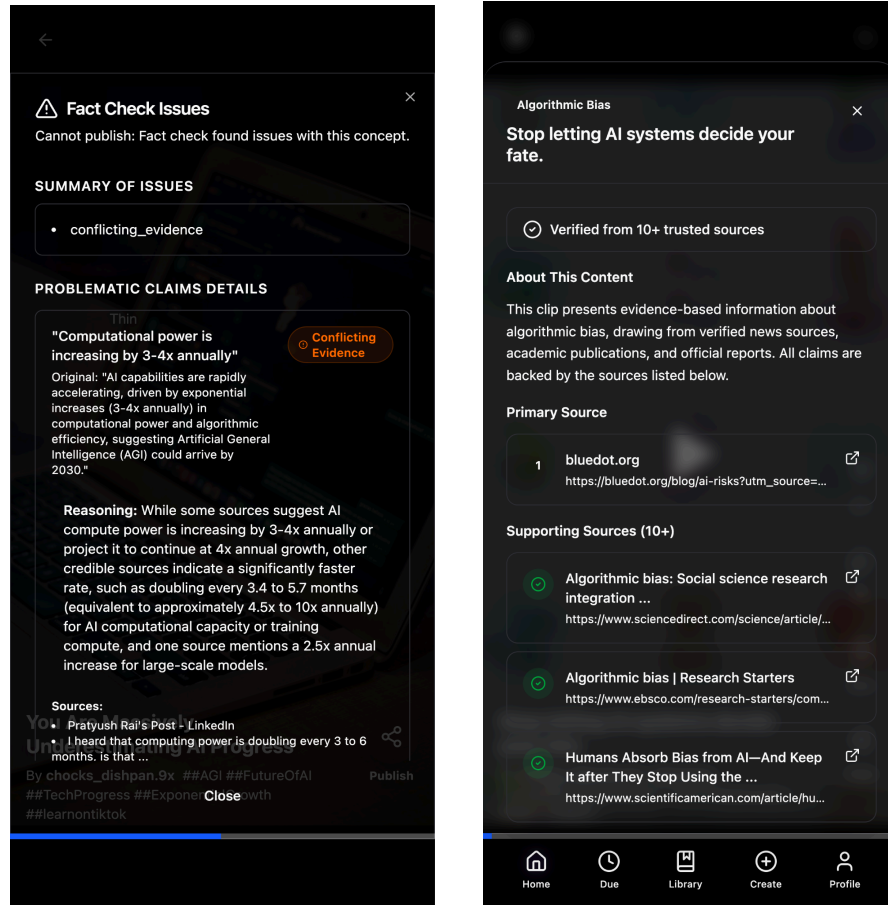


Figure 4: Refuted and supported learning points via fact validation engine

Each of the learning points above are also fed through a fact validation engine, which is a key defence mechanism emphasised in this submission to prevent misinformation spread. We use ClaimeAI (Bharath, 2025/2025), a recently open sourced fact-checker that's based off the Search-Augmented Factuality Evaluator (SAFE) methodology in "Long-form factuality in large language models" (Wei et al., 2024).

1. Claim Extraction

A learning point is then dissected into different claims. For example:

- Sentence: Major religions in Southeast Asia (SEA) include Islam, Buddhism and Hinduism.
- 3 claims are inferred from the sentence above:
 1. Islam is a major religion in SEA.
 2. Buddhism is a major religion in SEA.
 3. Hinduism is a major religion in SEA

2. Claim Verification

For each claim of a learning point, we:

- a. Run a web search to find comprehensive evidence for fact-checking each claim. An LLM is used to craft search queries that target authoritative sources (i.e, news, government, academic papers)
- b. Check whether the evidences directly addresses the claim, is from authoritative, reliable, and credible sources using LLM.
- c. Claims fall into categories:
 - i. **Supported** (confirmed by multiple independent sources).
 - ii. **Refuted** (authoritative sources explicitly contradict the claim).
 - iii. **Insufficient Information** (limited evidence, key information missing).
 - iv. **Conflicting evidence** (multiple credible sources present opposing views).
- d. In the previous example, the claim “Hinduism is a major religion in SEA” would have been *Refuted*, as only approximately 1.1% of the SEA population are Hindus (Miner, 2023).

3. Guardrails in Publishing

Clips with unverified or contradicted claims:

- Either cannot be published without revision, or must be reframed with explicit uncertainty (“Some experts argue that...”).
- We log claim text, evidence snippets, model version, and final decision for auditing.
- For higher-risk domains (for example, health or finance) and low-confidence claims, future iterations will include a human-in-the-loop review step where designated reviewers must explicitly approve or revise learning points before publication.

This engine directly addresses AI-enabled misinformation risk: we do not assume that AI-generated drafts are truthful; we subject them to systematic verification.

4. Learning Engine: Spaced Repetition and Difficulty Modeling

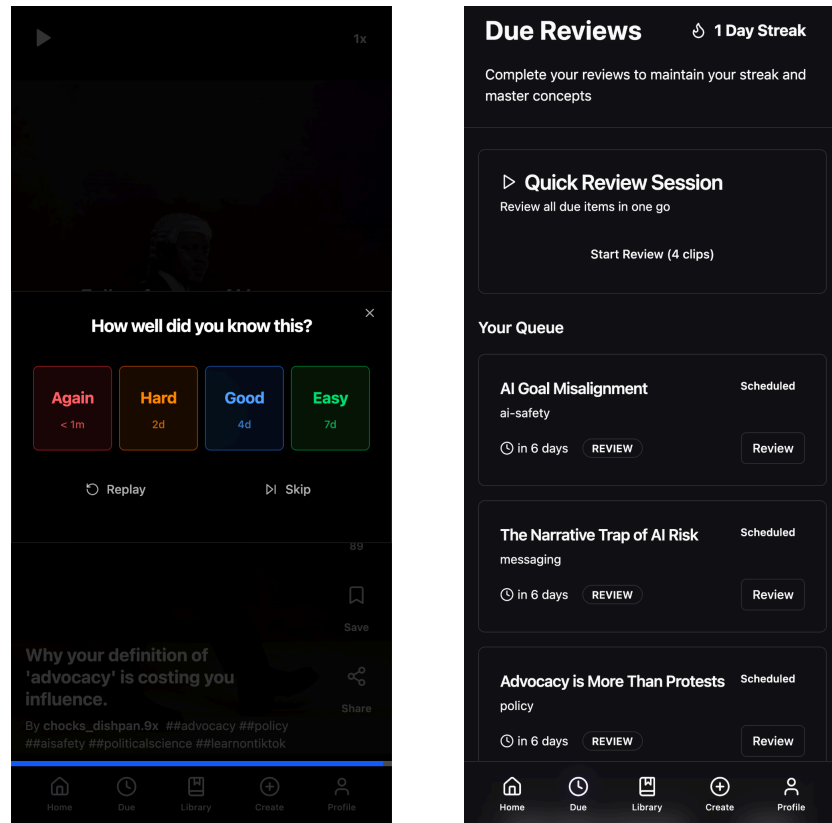


Figure 5: Difficulty assessment and spaced repetition review schedule

The learning engine treats each concept as a schedulable item with state:

- Stability, ease, and retrievability parameters.
- Current due date and review history.

We use:

- **FSRS-4.0** for spaced repetition scheduling, updating stability based on observed difficulty.
- Difficulty labels derived from interaction signals:
 - **Easy / Good / Hard / Again** based on recall correctness, timing, replays, and self-reported confidence.

Edge cases (for example, early skips with no recall attempt, off-topic comments) are handled conservatively to avoid overestimating mastery.

A background knowledge tracing model can refine ease parameters, suggest next concepts, and identify clips with anomalously high “again” rates—potentially signaling confusing or misleading content.

Conclusion

Our project addresses these AI harms below to the information ecosystem:

- Large-scale misinformation and influence operations.
- Overtrust in AI outputs (for example, hallucinated legal or medical advice).
- Erosion of shared understanding (people struggle to tell real expert consensus from AI-generated noise).

We do this by:

1. **Raising AI literacy at scale:** teaching how models fail, how manipulation works, and how to spot generated content.
2. **Using evidence-based learning design:** retrieval practice, one-concept-per-clip, and spaced repetition, rather than passive explainers.
3. **Embedding fact validation into the content pipeline:** preventing incorrect or misleading “educational” content from being widely propagated inside the app.
4. **Leveraging creators as multipliers:** enabling educators and subject-matter experts to create and export validated clips that propagate defensive knowledge across existing feeds.

In def/acc terms, we are strengthening the cognitive shield: making it harder for AI-enabled attackers to exploit public ignorance or cognitive shortcuts.

References

- Bharath. (2025). *BharathxD/ClaimeAI* [TypeScript].
<https://github.com/BharathxD/ClaimeAI> (Original work published 2025)
- Carpenter, S. K., Pan, S. C., & Butler, A. C. (2022). The science of effective learning with spacing and retrieval practice. *Nature Reviews Psychology*, 1(9), 496–511. <https://doi.org/10.1038/s44159-022-00089-1>
- Chu-Ke, C., & Dong, Y. (2024). Misinformation and Literacies in the Era of Generative Artificial Intelligence: A Brief Overview and a Call for Future Research. *Emerging Media*, 2(1), 70–85. <https://doi.org/10.1177/27523543241240285>
- Köbis, N. C., Doležalová, B., & Soraperra, I. (2021). Fooled twice: People cannot detect deepfakes but think they can. *iScience*, 24(11), 103364. <https://doi.org/10.1016/j.isci.2021.103364>
- Miner, J. E., Kelsey Jo Starr, Manolo Corichi and William. (2023, September 12). Religious landscape and change. *Pew Research Center*.
<https://www.pewresearch.org/religion/2023/09/12/religious-landscape-and-change/>
- Pilditch, T. D., Roozenbeek, J., Madsen, J. K., & van der Linden, S. (n.d.). Psychological inoculation can reduce susceptibility to misinformation in large rational agent networks. *Royal Society Open Science*, 9(8), 211953. <https://doi.org/10.1098/rsos.211953>
- Sebastian Bay, Giorgio Bertolin, Nora Biteniece, Edward H, Christie, Anton Dek, & Rolf E. (2019). *RESPONDING TO COGNITIVE SECURITY CHALLENGE*. NATO Strategic Communications Centre of Excellence.
- Vogel, I. (2012). *Review of the use of ‘Theory of Change’ in international development*.
- Wei, J., Yang, C., Song, X., Lu, Y., Hu, N., Huang, J., Tran, D., Peng, D., Liu, R., Huang, D., Du, C., & Le, Q. V. (2024). *Long-form factuality in large language models* (No. arXiv:2403.18802). arXiv. <https://doi.org/10.48550/arXiv.2403.18802>
- Zannettou, S., Nemeth, O.-N., Ayalon, O., Goetzen, A., Gummadi, K. P., Redmiles, E. M., & Franziska Roesner. (2024). *Analyzing User Engagement with TikTok’s Short Format Video Recommendations using Data Donations* (No. arXiv:2301.04945). arXiv. <https://doi.org/10.48550/arXiv.2301.04945>

Appendix A: Limitations and Risks

We explicitly acknowledge limitations and risks:

1. **Fact validation**
 - Fast-moving domains (for example, frontier models) may have limited consensus; we must represent uncertainty rather than overstate confidence.
 - May exhibit certain bias towards some sources to back up claims.
 - May miss nuance in complex claims, losing overall context.
 - Tavily AI might not find the best sources when searching for sources on the internet.
2. **Model reliance**
 - Claim extraction and evidence synthesis rely on LLMs; these components may themselves hallucinate or misjudge relevance.
 - We mitigate by combining multiple checks, human review queues, and conservative publishing rules, but residual risk remains.
3. **Unequal access and attention**
 - The platform mainly reaches users already engaged with short-form content; highly vulnerable or disconnected populations may be missed.
4. **Adversarial use**
 - In a worst case, an attacker could try to use similar pipelines to create more convincing persuasive content; open-sourcing must include clear safeguards and responsible-use guidelines.

Appendix B: Future Work

- **Partnerships with AI safety organizations:** collaborate with leading AI safety labs and nonprofits to source vetted, up-to-date content, co-design learning tracks, and coordinate broad promotion of high-quality AI safety education.
- **Institutional tracks and cohort features:** dedicated “tracks” for civil servants, journalists, educators, and community leaders.
- **Cross-integration with security tools:** connecting microlearning modules to corporate security dashboards, triggering targeted lessons after risky behavior is observed.
- **Richer validation sources:** integration with structured policy databases, official standards (for example, NIST AI Risk Management Framework), and curated AI safety corpora.
- **Ingestion of additional input formats:** expand beyond articles and URLs to support transcripts (SRT, podcast transcripts) and documents (PDF, DOCX), broadening the range of high-quality sources creators can pull from.
- **Localization and cultural adaptation:** Support multi-language content with local examples of AI misuse, local media norms, and region-specific policy references.