PRODUCT PORTFOLIO · CASE STUDY

Babbel Speak

From Hackathon Prototype to AI-Powered Speaking Feature at Scale

How I led the end-to-end product development of Babbel's first LLM-based conversational learning experience — from a rough GPT-3 demo to a home-screen feature with 200+ expert-curated scenarios, a dedicated Speak Tab on iOS & Android, and measurable retention uplift across millions of learners.

~35%
users return to feature within first week

200+
curated conversation scenarios shipped

8
real-life scenario categories

~2 yrs
prototype to home-screen

Franz Ardito
Principal Product Manager · Babbel GmbH
franzardito@pm.me · linkedin.com/in/franzaudio

Problem & Product Vision

THE STRUCTURAL GAP

Language learning apps — including Babbel — had long over-indexed on passive consumption: grammar drills, listening exercises, vocabulary flashcards. Users built knowledge but lacked the most critical skill: the ability to actually *speak*. The anxiety of producing language in real-time was the single biggest barrier to learner confidence and long-term retention.

> *"Users don't avoid speaking because they don't want to — they avoid it because it's deeply uncomfortable. Our job was to make the first conversation feel safe enough to start."*

THE OPPORTUNITY

By 2022, emerging LLM capabilities (GPT-3/3.5) opened a new design space: dynamic, non-scripted conversation practice that adapts to learner input in real time — something that scripted flows could never deliver. This was a rare convergence of a user need that had always existed and new technology capable of addressing it at scale.

Before: Scripted Conversations

Pre-defined dialogue trees, branching scripts maintained by content teams, predictable but rigid. Users felt safe but couldn't practice authentic language production. High completion rate, low speaking output.

After: AI Conversations (ConvoPro)

LLM-driven open-ended dialogue in structured scenarios. Users produce genuine sentences, receive real-time contextual feedback. Lower completion rate initially — but significantly higher active speaking output and stronger retention signal.

PRODUCT VISION

The long-term bet was a fundamental shift in Babbel's product paradigm:

> *From **"learning about a language"** → to **"actively using a language"** — with speaking as the primary modality, not an afterthought.*

This was not just a feature — it was a strategic repositioning of Babbel's core learning loop around active production. Lessons would increasingly serve as preparation for speaking, not as ends in themselves.

MY ROLE

I owned Babbel Speak end-to-end as Principal PM from the early LLM prototyping phase through to the home-screen launch and beyond. This meant leading a cross-functional squad (Engineering, Design, Content, QA, Data Science), owning the roadmap, navigating ambiguity inherent to LLM product development, and balancing engagement uplift against non-linear AI infrastructure costs.

LLM
GPT-3 → GPT-3.5 → custom LLM integration

iOS + Android
simultaneous platform launch

5 PMs
mentored alongside product delivery

Discovery Process & Key Initiatives

FROM HACKATHON TO HOME SCREEN

2022 Q3

Hackathon prototype. First GPT-3 demo exploring unscripted conversation practice. Proof of concept that LLMs could hold a coherent language-learning dialogue. Internal excitement high — but latency, cost, and hallucination risk were unresolved.

2022–2023

Early discovery & LLM evaluation. As Lead PM, ran structured prototyping cycles with GPT-3.5. Tested scenario structures, feedback modalities, and pedagogical fit with linguistics team. Established cost vs. engagement framework. Made the strategic call: AI over scripted for long-term retention.

2023 Q4

ConvoPro alpha launch. First closed test of the AI conversation experience. Validated that users produce meaningfully more words per session vs. legacy formats. Surfaced critical UX friction (beginner confidence, audio controls) that became the onboarding roadmap.

2024

Speak Tab + full rollout. Promoted to Principal PM. Launched dedicated Speak Tab (iOS & Android) centralising all speaking experiences. Shipped 200+ expert-curated scenarios across 8 real-life categories. Integrated Lesson→Speak connection to create a continuous learning loop.

2025

Retention signal confirmed. ~35% of exposed users return to Speak within D1–D7. AI conversations outperform scripted on speaking output and downstream retention. Enriched feedback (pronunciation, inline corrections) enters the roadmap as next frontier.

KEY PRODUCT INITIATIVES

INITIATIVE 01

Welcome Scenarios

Guided first conversation with L1 support and progressive L2 transition. Reduced first-session drop-off, increased words spoken in session 1, and seeded early personalisation signals.

INITIATIVE 02

AI Conversation Core Loop

Open-ended LLM dialogue within structured scenarios. Speak → receive feedback → continue. Higher speaking output than scripted, stronger D7 retention signal.

INITIATIVE 03

Lesson → Speak Bridge

Trigger Speak sessions after lesson completion to immediately activate newly learned vocabulary. Cross-feature cohort showed uplift in both return-to-Speak and return-to-Learn.

INITIATIVE 04

UX Harmonisation

Unified interaction patterns (replay, slow playback, translation) across all speaking surfaces despite different backend stacks. Reduced beginner friction, improved session completion.

INITIATIVE 05

Speak Tab (Surface)

Dedicated home-screen tab making speaking a first-class citizen alongside Learn. Increased discoverability, established speaking as a core daily behaviour not a secondary feature.

INITIATIVE 06

Enriched Feedback (Next)

Missed word detection, inline corrections, pronunciation feedback. Deterministic systems preferred to reduce LLM cost and latency variance. Currently in roadmap.

Metrics, Learnings & Strategic Tradeoffs

METRICS FRAMEWORK

With LLM products, standard funnel metrics are necessary but insufficient. We anchored on behaviours that signal genuine language production and habit formation — not completion rate, which would have led us to over-invest in scripted flows.

Metric Layer	Metric	Signal
North Star	Words spoken per session · Time spent speaking	Direct measure of active language production — the core behaviour we are building
Retention	Return to Speak D1–D7 · Return to Learn D1–D7	Habit formation; target D7 return >35% (achieved ~35%)
Engagement	Conversations completed · Sessions per user · Speaking frequency	Depth and breadth of usage within a session
Cost Efficiency	AI cost per session · Engagement uplift per dollar	LLM + voice introduces non-linear cost scaling — ROI must be tracked against retention benefit

KEY LEARNINGS

Speaking drives retention

Users who engage in speaking return more frequently and build stronger habits. Speaking output is a leading indicator of long-term subscription retention — more predictive than lesson completion.

Beginner friction is disproportionately costly

Small UX improvements — replay audio, slow playback, L1 support at onboarding — had outsized impact on activation and confidence. Never underestimate the emotional barrier of the first sentence.

AI vs. scripted: choose your metric

Scripted flows have higher completion rates. AI conversations have deeper speaking engagement. The correct choice depends entirely on which outcome you are optimising for.

Personalization is a retention multiplier

Users who experience a system that reflects their level, interests, and goals return significantly more. Adaptive scaffolding is not a nice-to-have — it is infrastructure for stickiness.

Small-batch iteration with LLMs

Standard sprint cycles are poorly suited to LLM product development. We adopted a small-batch iteration model optimised for faster hypothesis testing and tolerance for output unpredictability.

Cost is a product decision, not just engineering

Every new feedback feature, every extra LLM call, has a direct cost-per-session implication. PMs owning AI products must carry cost awareness as a first-class product constraint.

STRATEGIC TRADEOFFS NAVIGATED

Engagement vs. Cost

LLM + voice creates non-linear cost scaling. Every engagement gain must be weighed against cost-per-session. We built an explicit ROI framework before each initiative.

`LLM cost` `ROI framework`

Feedback Depth vs. Cognitive Load

Too much feedback overwhelms beginners; too little feels low-value. We designed progressive disclosure: implicit feedback first, explicit corrections unlocked over time.

`UX design` `pedagogy`

AI vs. Legacy Coexistence

Speak Tab had to surface both AI Conversations and legacy Guided Conversations without cannibalising lesson engagement. Required careful surface orchestration.

`product strategy` `cannibalisation`

> *Babbel Speak demonstrated that an LLM-powered speaking feature can move the retention needle in a consumer EdTech product — when product discovery, pedagogical rigour, and cost discipline work in concert.*

Franz Ardito · franzardito@pm.me · linkedin.com/in/franzaudio