

B2B Matching Algorithms and Data Sourcing for Swiss SME Ecosystems

Reframing matching as an information-retrieval problem on Swiss open data, with a pivot to conversational intelligence

Student

Koray Kaya

Initial Situation: Small and medium-sized enterprises make up more than 99 percent of the Swiss economy, yet only about 17 percent of them take part in formal research and development. Project Work 1 (December 2025) addressed this innovation gap with the hypothesis of an automated B2B matching platform for Swiss SMEs. Phase 2 set out to move that hypothesis from a conceptual prototype to an operational system, to test it empirically, and to determine whether the project should continue as a Master thesis.

Approach: The work followed a Lean Startup loop: build, test, and let the results drive the next step. Three Swiss open data sources (Zefix, simap.ch, Swissreg) and the LINDAS linked data infrastructure were integrated; structural company identity was enriched with free-text fields from company websites through a multi-model pipeline with an evidence-based judge layer. Three institutional partnerships supported the work: CSCS Swiss AI (ETH Zürich) for large language and embedding models, Weaviate for vector storage and hybrid search, and Langfuse for LLM observability. The matching task was reframed as an asymmetric information-retrieval problem. A four-stage hybrid search architecture (parallel sparse and dense retrieval, RRF fusion, funnel cutoff, cross-encoder re-ranker) was built on multilingual embeddings, with HyDE query expansion. Two studies followed: Experiment A compared sparse, dense, and hybrid retrieval on 3,153 St. Gallen companies with 30 queries and 544 LLM-scored pairs; Experiment B implemented an end-to-end supplier search system on 247 Swiss B2B companies, testing HyDE and a provenance-based filter strategy.

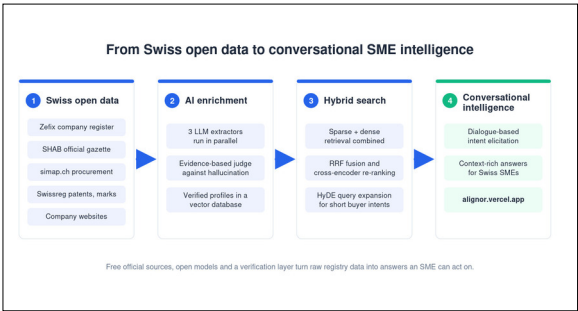
Result: Three core findings emerged. First, retrieval performance is determined by the architecture of the pipeline, not by parameter tuning: equal-weighted hybrid retrieval underperformed dense embeddings on their own, while a correctly placed funnel cutoff lifted hybrid performance to the level of dense retrieval. Second, filtering strategy should follow source reliability: moving certificate fields from a hard filter into the semantic text raised nDCG@10 from 0.444 to 0.794. Third, static matching offers users only one-time value, while sustainable engagement requires the interactive elicitation of buyer intent through conversation.

These findings led to a structured pivot: the project shifted from a static matching pipeline to a conversational Swiss SME intelligence platform, keeping the technical infrastructure and turning the user interface into a dialogue-based intent elicitation process. A ten-screen prototype (alignor.vercel.app) was released, and inbound interest from a pre-seed investor (SIVentures, Leipzig) followed two weeks later, providing constructive external

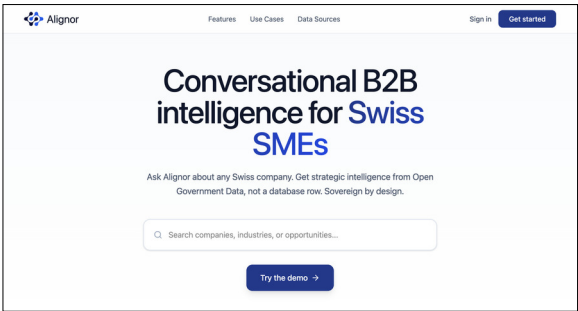
feedback.

The project was repositioned as an AI engineering transition portfolio and will continue as a Master thesis starting in September 2026. The empirical results, two open-source repositories (hybrid-search-benchmark, swiss-b2b-supplier-search), and a weekly Substack journal document the full process.

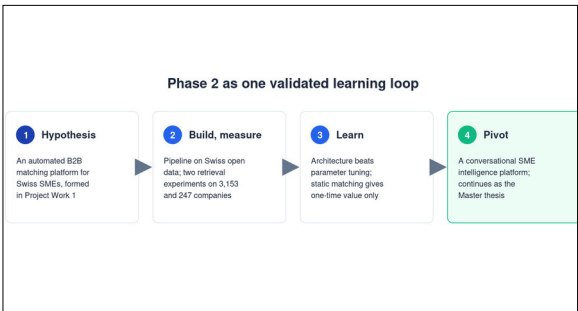
From Swiss open data to conversational SME intelligence: sources, AI enrichment, hybrid search, chat
Own presentation



The conversational prototype released at alignor.vercel.app after the pivot
Own presentation



Phase 2 as one validated learning loop: hypothesis, experiments, findings, and the pivot to a chat platform
Own presentation



Advisor
Samuel Böhm

Subject Area
Business Engineering,
Computer Science