AI-ENABLED PROFILING PIPELINE TRANSFER PRICING · TAX ADVISORY

Audit-ready company profiles, built, compared and screened at scale

EmergeFlow built an automated, AI-enabled multilingual profiling pipeline that turns spreadsheet batches of website URLs into rich, cited, translated company profiles — then compares each against a defined **tested party** using a Gemini LLM, accepting matches and rejecting non-matches with a documented reason. Over one year it processed roughly **150,000 websites**, replacing thousands of hours of manual research.

01 CLIENT & NEED

The client provides comprehensive **transfer-pricing services** and international taxation advisory — benchmarking studies, cross-border royalty structuring, and audit-ready documentation — across IT/ITeS, financial services, automotive, manufacturing, and pharmaceuticals. They needed an automated, resilient, scalable system to **extract, translate, summarize, and screen** corporate data at a volume reaching **~150,000 sites a year**.

- **Manual, fragmented research.** Analysts navigated to each site by hand, took screenshots, ran foreign domains through external translators, and copy-pasted sentences to build profiles.
- **Subjective comparability screening.** Each profile was judged against a defined tested party by hand — slow, inconsistent, with no audit trail for why a company was accepted or rejected.
- **Untraceable claims.** Tracking the exact source URL behind every statement was tedious and prone to human error.
- **Unsustainable at scale.** Volumes reaching ~150,000 sites a year made manual research, deduplication, and screening impossible to keep up with.
- **Brittle scraping.** Dynamic content, anti-bot protections, and single-page architectures broke standard scrapers, leaving significant data gaps.

BEFORE IMPLEMENTATION

Analysts manually navigated sites, captured screenshots, translated foreign domains with external tools, and hand-built profiles sentence by sentence — then judged each one against the tested party by hand, with no record of why it was accepted or rejected. The approach left data gaps and could not scale toward 150,000 sites a year.

AFTER IMPLEMENTATION

An AI pipeline runs end to end: intelligent ingestion with automated screenshots; a resilient LLM fallback when sites block crawlers; a multilingual engine producing strictly extractive, sentence-cited summaries with English translations; a Gemini comparison stage that scores every profile against the tested party and accepts or rejects it with a documented reason; and a MongoDB dedup and reporting layer exporting ordered Excel — sustained across ~150,000 sites a year.

02 RESULTS

Small difference. Large impact.

0 halluc.
Extractive by design
Summaries are quoted, not generated — every sentence carries its exact source URL, so nothing is invented.

Fall back
Crawler-proof ingestion
When a site blocks bots, a search-grounded LLM reconstructs the profile — closing the data gaps scrapers leave.

Gemini score
Explainable screening
Each profile is matched to the tested party and accepted, or rejected with a written, defensible reason.

150K /yr
Dedup'd throughput
Cross-batch dedup skips already-fresh records; output stays in the source file's exact order.