

Got it — you’re not asking for just “tools” but for **deep, free, high-signal knowledge bases and datasets** that could feed a “founder-assessment brain” with enough breadth and depth to make credible evaluations across:

- 1. **AI codebase quality & tech feasibility**
- 2. **Business idea viability**
- 3. **Investment & market potential**

If we think of your “brain” as a **multi-domain knowledge ingestion engine**, the goal is to give it **structured, up-to-date, diverse, and cross-referenced data** from trusted sources.

1. AI Codebase & Technical Feasibility

These give you **technical benchmarks, code quality standards, bug datasets, and open-source code for comparison.**

Source	Type	Why It’s Valuable
CodeNet (IBM)	14M+ code samples in 55 languages	Train models to detect bad patterns, optimize code, and assess maintainability
BigQuery Public GitHub Dataset	Parsed GitHub repos & metadata	Lets you analyze code popularity, stars, issue activity, commit frequency
Snyk Vulnerability DB (snyk.io/vuln)	Security flaws & severity ratings	Adds security-risk awareness to assessments
SonarSource Rules & Quality Profiles (rules.sonarsource.com)	Code smells, maintainability, complexity rules	Standardizes code quality scoring
Stack Overflow Developer Survey (insights.stackoverflow.com)	Annual skill demand, tech usage trends	Helps brain contextualize language/framework adoption

Awesome Engineering Standards (github.com/standard)

Curated repos of engineering best practices

Benchmarks against accepted practices

2. Business Idea Viability

These provide **market data, consumer trends, startup performance metrics, and competitive landscapes**.

Source	Type	Why It's Valuable
World Bank Open Data (data.worldbank.org)	Economic, industry, and population stats	Macro-level market size and feasibility
OECD Data (oecd.org/statistics)	Industry output, R&D spend, innovation indexes	Helps judge innovation climate
UN Comtrade (comtrade.un.org)	Global import/export flows	Useful for supply chain & trade feasibility
Google Trends API	Keyword popularity over time	Gauges consumer interest growth/decline
GEM – Global Entrepreneurship Monitor (gemconsortium.org)	Entrepreneurial activity & attitudes by country	Benchmarks startup culture readiness
Statista Free Reports (statista.com)	Market research snippets	Context for industry traction
Wayback Machine (Archive.org)	Historical site & product positioning	Competitive analysis over time

3. Investment & Market Potential

These bring **financial health, funding history, valuation benchmarks, and investor sentiment**.

Source	Type	Why It's Valuable
--------	------	-------------------

Crunchbase Basic Dataset (data.crunchbase.com)	Funding rounds, valuations, investor profiles	Historical success patterns
CB Insights Research Briefs (free editions)	Industry & funding trend reports	Identifies hot sectors and deal volumes
AngelList API (limited)	Startup profiles & investor activity	Current investment appetite in niches
SEC EDGAR (sec.gov/edgar)	Filings for public companies	Benchmarking exits and IPO pathways
PitchBook Blog & Free Data (pitchbook.com/news)	Deal flow trends	Helps align with investor expectations
Alpha Vantage (alphavantage.co)	Financial markets data	Macro investor sentiment indicators
OECD Venture Capital Indicators	Cross-country VC activity data	Global fundraising comparisons

4. Meta-Knowledge for the Brain (Cross-Domain Intelligence)

To make credible *cross-cutting* founder assessments, your brain will need:

- **Benchmarks:** Pull in **industry averages** (growth rates, burn rates, CAC/LTV) from sources like **OpenVC**, **Y Combinator public talks**, and **First Round Review** archives.
 - **Failure Analysis:** Ingest **CB Insights "Why Startups Fail" dataset** + postmortems from Failory.
 - **Innovation Scoring Models:** Use frameworks like **TRIZ**, **Blue Ocean Strategy** summaries, and **Lean Startup** metrics.
 - **Regulatory Data:** For compliance-heavy industries, integrate from **FDA Open Data**, **EU Open Data Portal**, etc.
-

If we fed all of this into an **indexed vector store** (e.g. Weaviate, Milvus, or PostgreSQL pgvector) with **cross-linking** between technical, market, and investment data, your “brain” could:

- Score a founder’s *AI tech stack* against industry best practice and security standards
- Validate *market readiness* using macro/microeconomic signals
- Estimate *fundability* based on similar past startups’ funding success & exit paths

If you want, I can **map this into a knowledge ingestion pipeline** showing **what gets pulled, how it’s stored, and how it’s linked** so you have the “ultimate founder assessment brain” blueprint.

That way, it’s not just a list—it’s a structured, query-ready brain.

Do you want me to do that next?

You said:

Yes!

ChatGPT said:

Alright — here’s the **blueprint for your “Ultimate Founder Assessment Brain”**:

We’ll structure it so your brain can **pull, store, link, and query** data across **AI code quality, business idea viability, and investment potential**.

1. Data Architecture Overview

Goal: Ingest *free, credible knowledge bases* into a **linked, queryable memory** so your brain can run multi-domain assessments.

[Data Sources] → [ETL/Collector Layer] → [Vector Store + Relational DB] → [Brain/LLM Interface]

2. Data Sources by Domain

A. AI Codebase Quality

- **CodeNet** → algorithmic solutions, language benchmarks
- **GitHub BigQuery Dataset** → repo stats, contributor activity
- **SonarSource Rules DB** → code smells, complexity, maintainability
- **Snyk Vuln DB** → known security issues + CVSS scores
- **Stack Overflow Survey** → tech adoption, skill relevance

Key Fields to Store:

language, complexity_score, security_vulnerabilities, repo_activity, maintainability_index

B. Business Idea Viability

- **World Bank / OECD** → GDP, internet penetration, R&D spend
- **Google Trends API** → market interest over time
- **GEM** → entrepreneurial readiness index by country
- **UN Comtrade** → supply chain/trade feasibility
- **Statista Free Reports** → sector growth projections
- **Wayback Machine** → competitor history

Key Fields to Store:

market_size, trend_score, innovation_index, competitive_density, supply_chain_risk

C. Investment Potential

- **Crunchbase Basic API** → funding rounds, investors, valuations
- **AngelList API** → current startup funding appetite

- **SEC EDGAR** → IPO/exit financials
- **CB Insights Free Reports** → hot sectors, funding volumes
- **Alpha Vantage API** → market sentiment indicators
- **OECD VC Indicators** → regional capital flows

Key Fields to Store:

`funding_round_size, valuation_trend, investor_density, sector_heat, exit_probability`

3. Data Pipeline

1. Collection Layer

- **APIs** (Alpha Vantage, Crunchbase, Google Trends, AngelList, etc.)
- **Bulk Downloads** (CodeNet, GEM, OECD)
- **Web Scrapers** (Wayback Machine competitor snapshots, Statista free sections)

2. Transformation Layer (ETL)

- Normalize date formats, currencies, language names, and units
- Run NLP tagging on text data (e.g., business descriptions → keywords)
- Calculate derived metrics (e.g., “market growth rate” = CAGR from OECD data)

3. Storage Layer

- **Vector Database (Weaviate / Milvus / pgvector)**
 - Stores semantic embeddings for *searching across domains* (e.g., “Find startups with code security scores > 80 and market CAGR > 15%”)
- **Relational DB (PostgreSQL/MySQL)**
 - Stores structured facts & metrics for filtering, aggregation

4. Linking the Data

We create **entity graphs**:

- **Startup** → linked to **CodebaseQualityProfile**, **MarketProfile**, **FundingProfile**
- **MarketProfile** → linked to **CountryData**, **TrendData**
- **FundingProfile** → linked to **InvestorProfiles** & **HistoricalRoundData**

Example link query your brain could run:

“For a startup in fintech using Python, compare its repo security score to industry median, market growth over 5 years, and average funding for similar stage companies.”

5. Brain / LLM Interface

- Use an **LLM orchestration layer** (LangChain, LlamaIndex)
- LLM queries both:
 - **Vector DB** → to retrieve related knowledge/documents for context
 - **Relational DB** → to pull hard metrics for scoring
- Implement **assessment templates**:
 - **Technical Feasibility Score** (0–100)
 - **Market Opportunity Score** (0–100)
 - **Investment Attractiveness Score** (0–100)
- Combine into a **Final Founder Score** weighted by your preference

6. Example Flow

1. Founder submits:
 - GitHub repo URL
 - One-sentence business idea
 - Target market/location
2. System:
 - **Pulls repo stats** from GitHub BigQuery → runs Sonar rules & Snyk scan → outputs **Tech Score**
 - **Checks market** via Google Trends + OECD sector data → outputs **Market Score**
 - **Matches funding comps** via Crunchbase & AngelList → outputs **Investment Score**
3. Brain merges → returns **Assessment Report + Recommendations**

7. Scaling the Brain

- Start with **static ingestion** of free datasets (bulk CSV, JSON)
- Add **live API hooks** for trending/real-time data
- Add **fine-tuned scoring models** based on historical founder success patterns



