

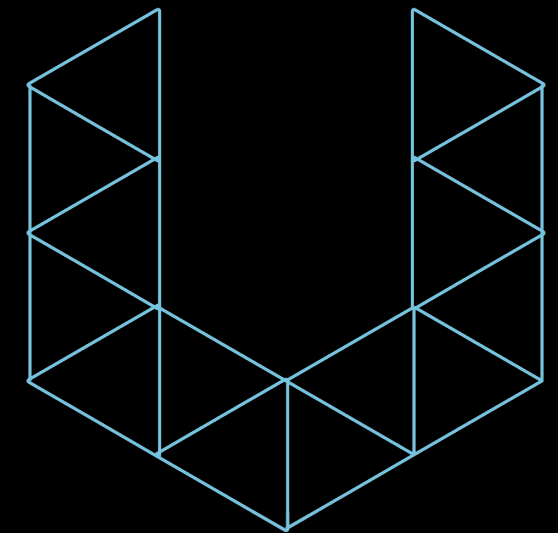


Unstructured.io

Mitigating the Risks of Bias & Generative AI Limitations with Gold-Standard Data

BENGAL Lightning Talk

Jonathan Nadzam, Public Sector



UNSTRUCTURED



LLM Threats, Biases & Vulnerabilities

- For safe & effective use of Large Language Models (LLMs) in IC applications, critical to understand and mitigate against LLM threat vectors and vulnerabilities
- When developing LLM applications for Government use, high-quality data plays a crucial role in combating these threats and weaknesses in various stages of development.

LLM01 Prompt Injection This manipulates a large language model (LLM) through crafty inputs, causing unintended actions by the LLM. Direct injections overwrite system prompts, while indirect ones manipulate inputs from external sources.	LLM02 Insecure Output Handling This vulnerability occurs when an LLM output is accepted without scrutiny, exposing backend systems. Misuse may lead to severe consequences like XSS, CSRF, SSRF, privilege escalation, or remote code execution.	LLM03 Training Data Poisoning Training data poisoning refers to manipulating the data or fine-tuning process to introduce vulnerabilities, backdoors or biases that could compromise the model's security, effectiveness or ethical behavior.	LLM04 Model Denial of Service Attackers cause resource-heavy operations on LLMs, leading to service degradation or high costs. The vulnerability is magnified due to the resource-intensive nature of LLMs and unpredictability of user inputs.	LLM05 Supply Chain Vulnerabilities LLM application lifecycle can be compromised by vulnerable components or services, leading to security attacks. Using third-party datasets, pre-trained models, and plugins add vulnerabilities.
LLM06 Sensitive Information Disclosure LLMs may inadvertently reveal confidential data in its responses, leading to unauthorized data access, privacy violations, and security breaches. Implement data sanitization and strict user policies to mitigate this.	LLM07 Insecure Plugin Design LLM plugins can have insecure inputs and insufficient access control due to lack of application control. Attackers can exploit these vulnerabilities, resulting in severe consequences like remote code execution.	LLM08 Excessive Agency LLM-based systems may undertake actions leading to unintended consequences. The issue arises from excessive functionality, permissions, or autonomy granted to the LLM-based systems.	LLM09 Overreliance Systems or people overly depending on LLMs without oversight may face misinformation, miscommunication, legal issues, and security vulnerabilities due to incorrect or inappropriate content generated by LLMs.	LLM10 Model Theft This involves unauthorized access, copying, or exfiltration of proprietary LLM models. The impact includes economic losses, compromised competitive advantage, and potential access to sensitive information.

OWASP Top 10 for LLM, https://owasp.org/www-project-top-10-for-large-language-model-applications/assets/PDF/OWASP-Top-10-for-LLMs-2023-slides-v1_0.pdf



BENGAL

LLMs need recent, relevant, and validated data to mitigate against weaknesses

Problem 1: Nearly all LLMs have been trained on the same corpus of internet data

Solution 1: Pre-train or fine tune on **proprietary data**

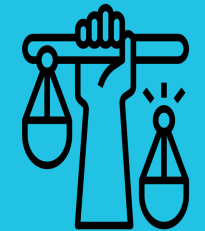
Problem 2: LLMs are “frozen in time”

Solution 2: Make **new data** available to the model in a vector database

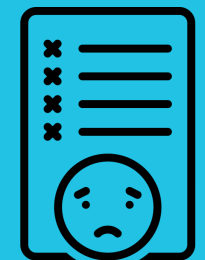
Problem 3: LLMs hallucinate

Solution 3: Force LLMs to focus on **validated data**

MITIGATE



BIAS



HALLUCINATIONS

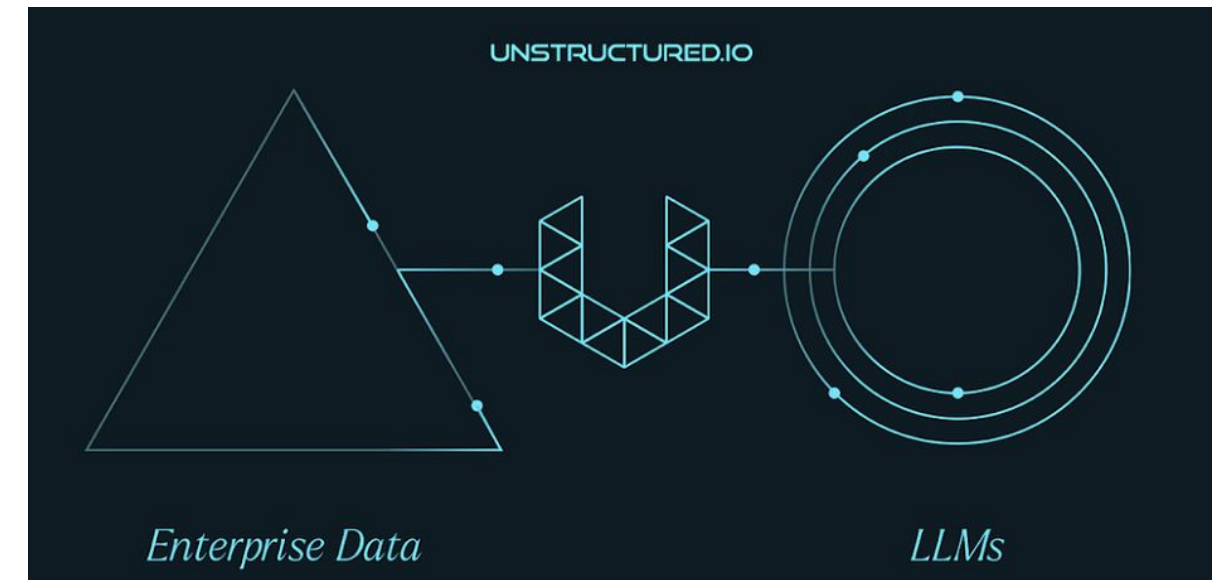


BENGAL

High-Quality, Organizational-Specific Data for Safe LLM Use

With Gold-Standard, Organization Specific Data:

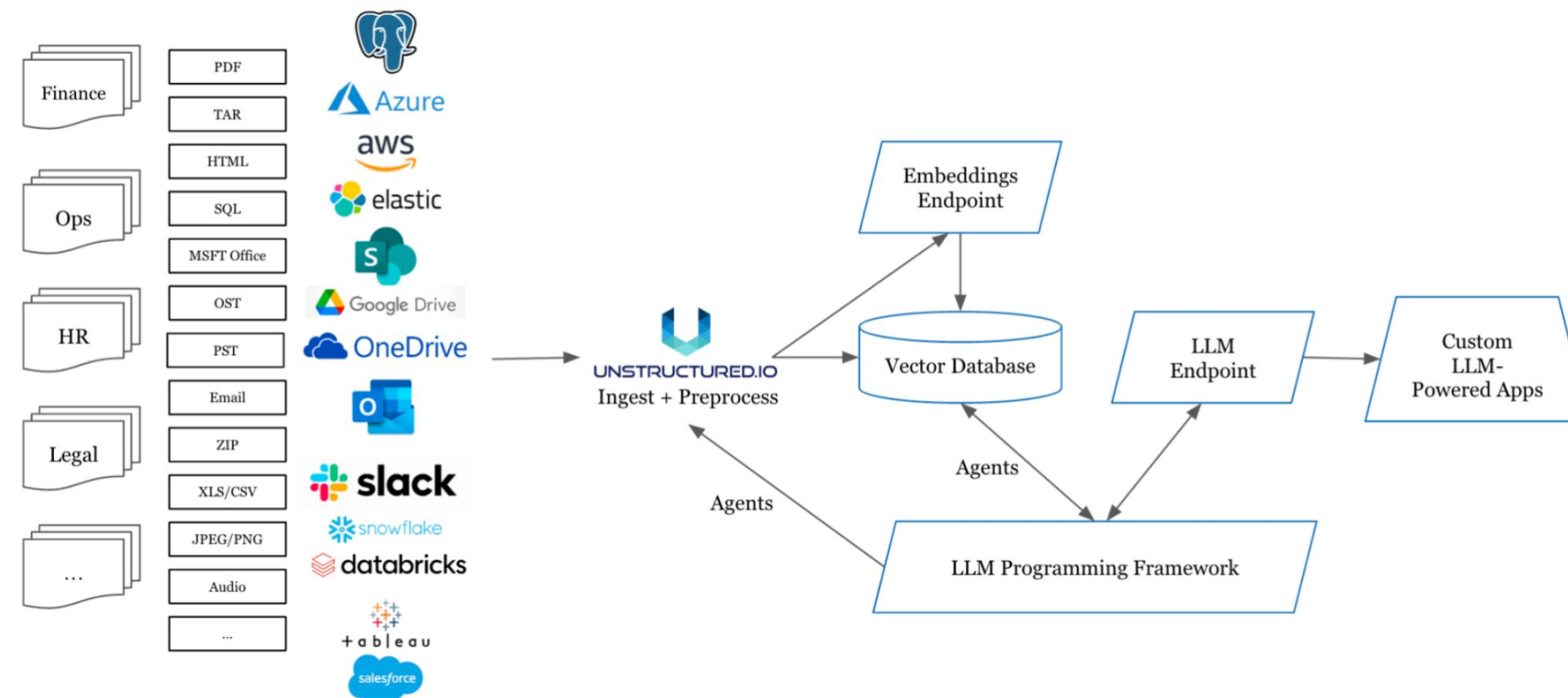
- Reduce Hallucinations/Erroneous Outputs
- Mitigate Bias
- Verify Sourcing
- Control Access
- Reduce Toxic Outputs





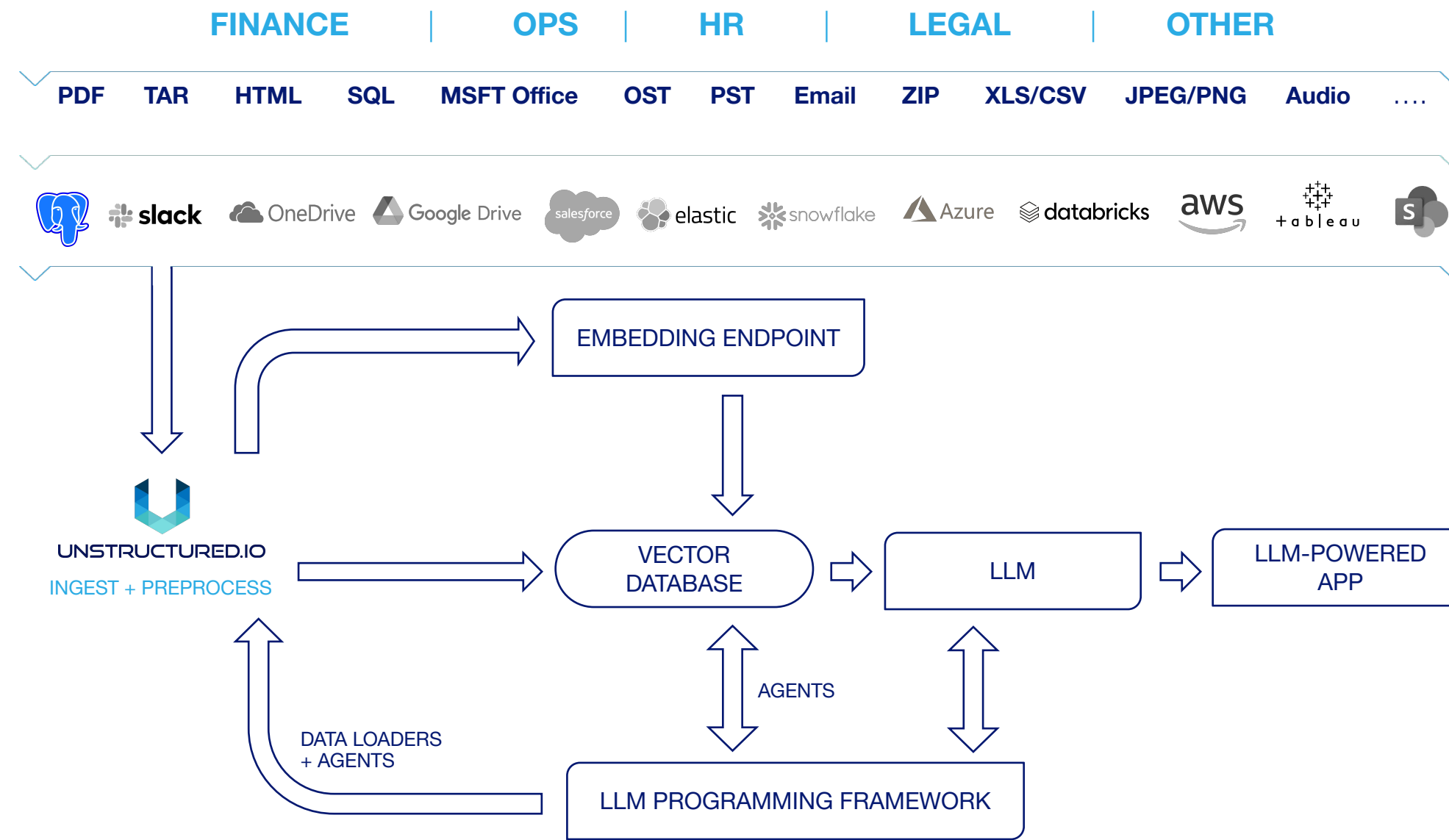
Emerging LLM Tech Stack: Retrieval Augmented Generation

- **Retrieval Augmented Generation (RAG):** an AI framework that synergistically combines the capabilities of LLMs with an integrated information retrieval system, leveraging external databases to deliver more precise, contextually relevant, and up-to-date responses & limit hallucinations





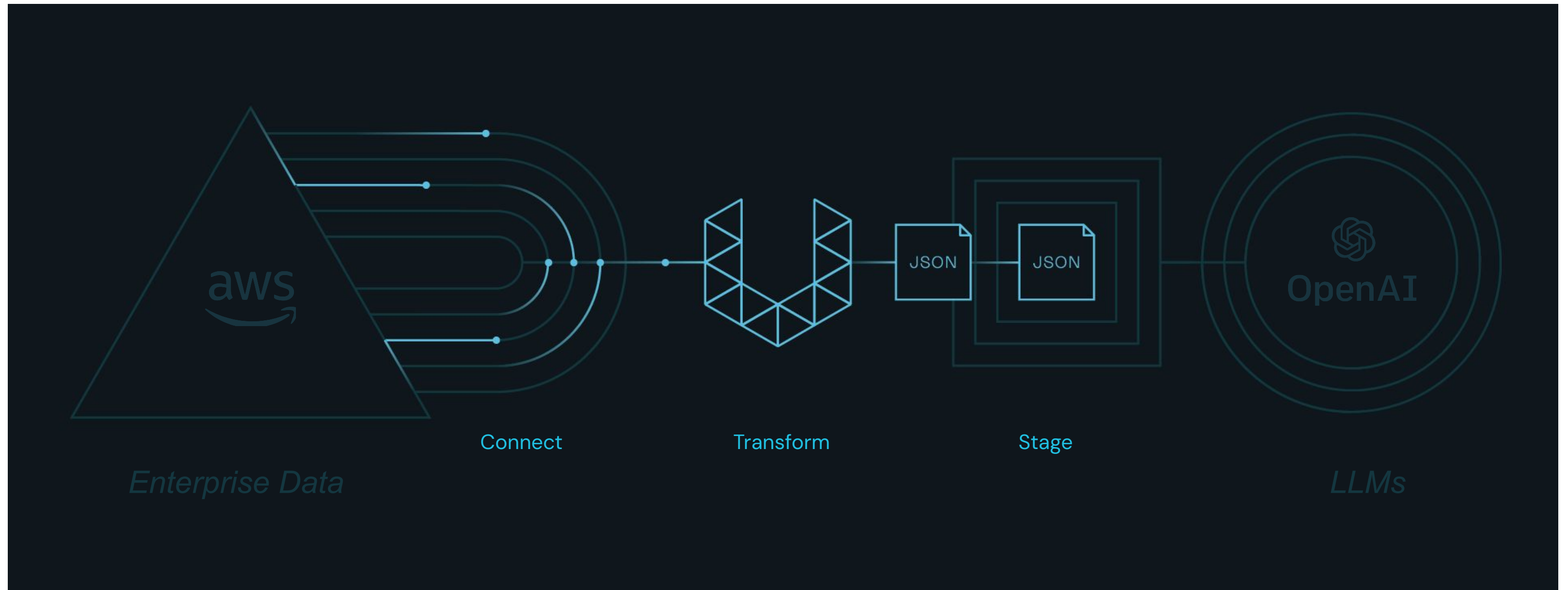
Where Unstructured Fits In





About Us

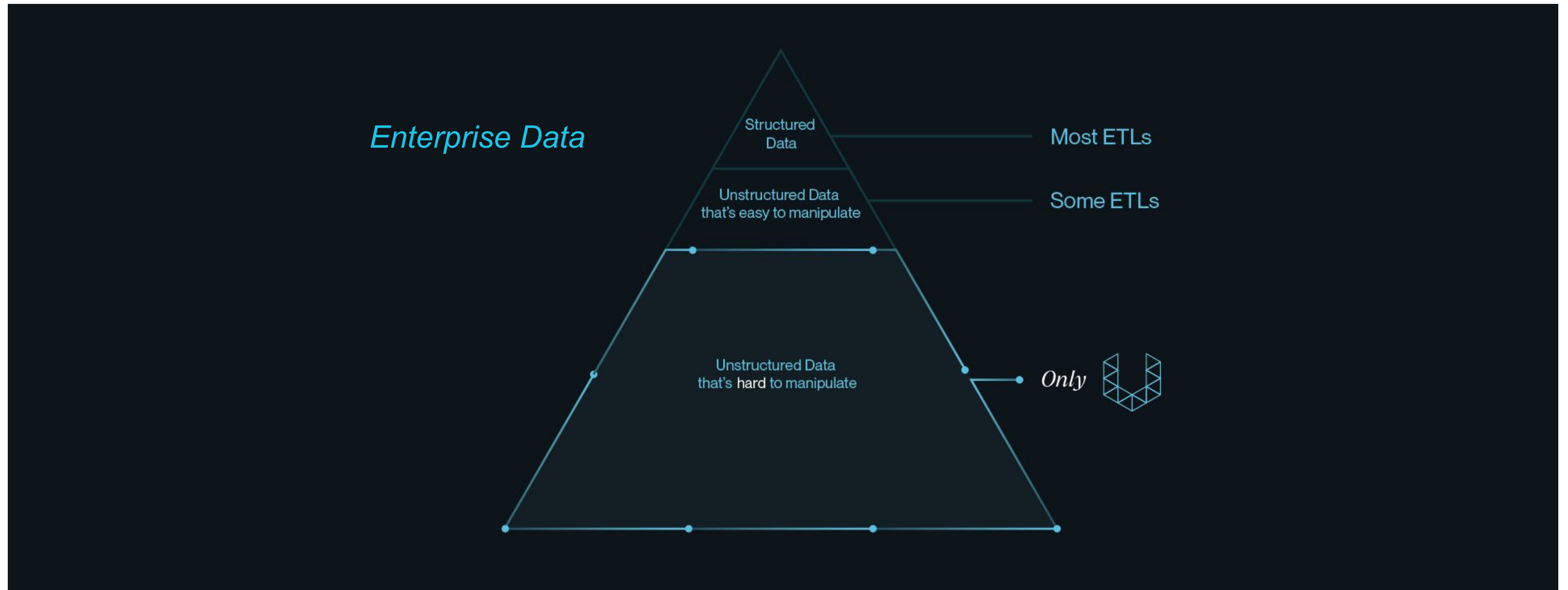
We connect natural language data to LLMs





How We're Different

Any document. Any file type. Any layout.



UNSTRUCTURED



In Detail

Logical Architecture:

