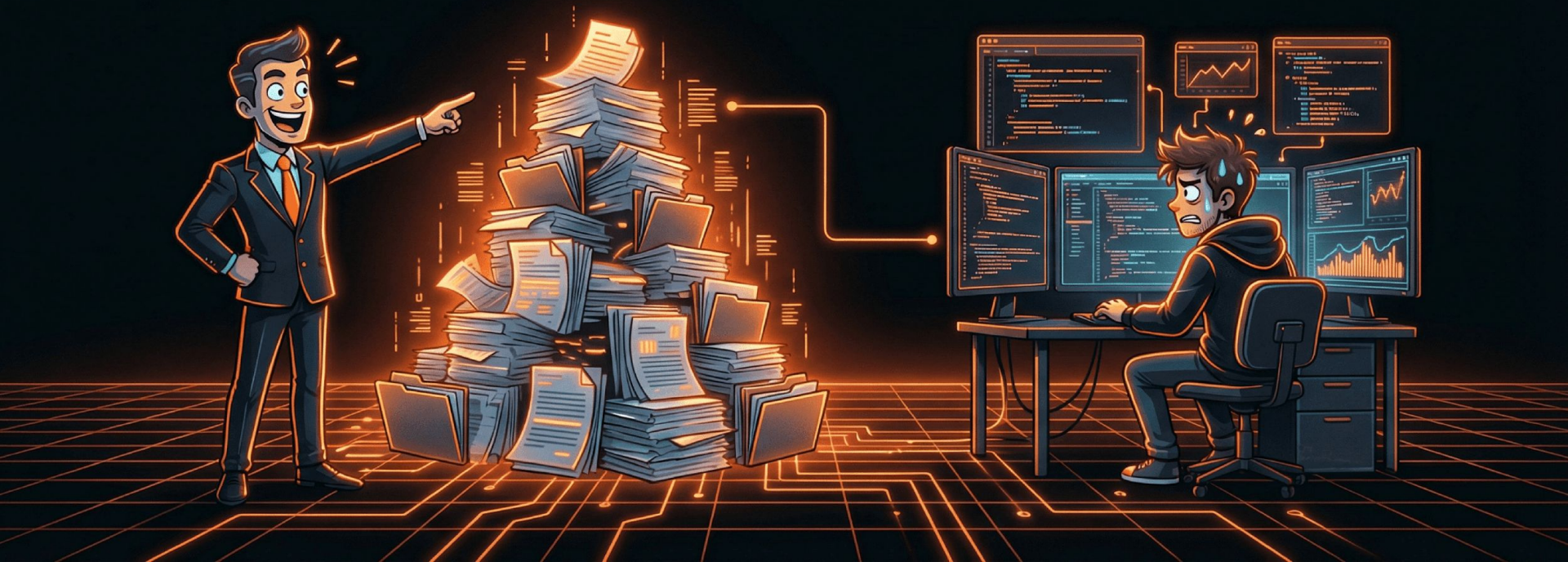
Lille GAB #4

Extraction de données avec l'IA en 2026

Louis Choquel
Co-fondateur & CTO de Pipelex

Extrais les infos de nos 100 000 documents

Le défi de l'ingénieur IA moderne



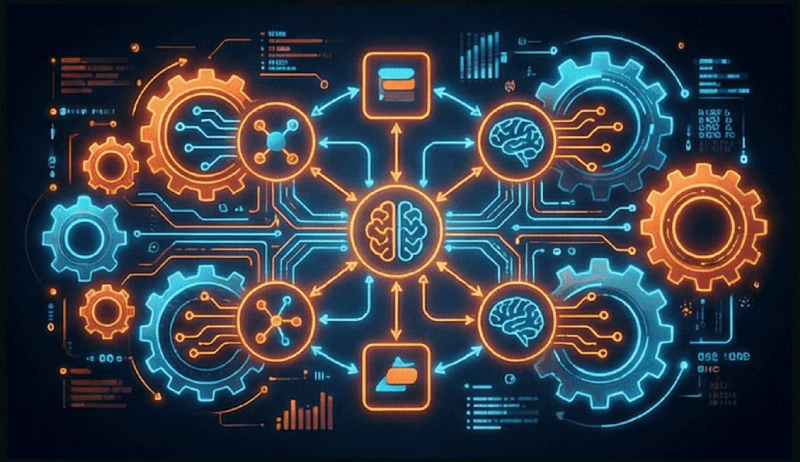
L'Évolution du Mindset

De la Question à l'Automatisation

2024

Je vais poser des questions à mes PDFs

2026



Automatiser · Donner des outils aux agents

Buy SAAS → Build Yourself

Avant de passer ça à mon LLM...
2024 vs 2026

2024

Extraire le texte du PDF

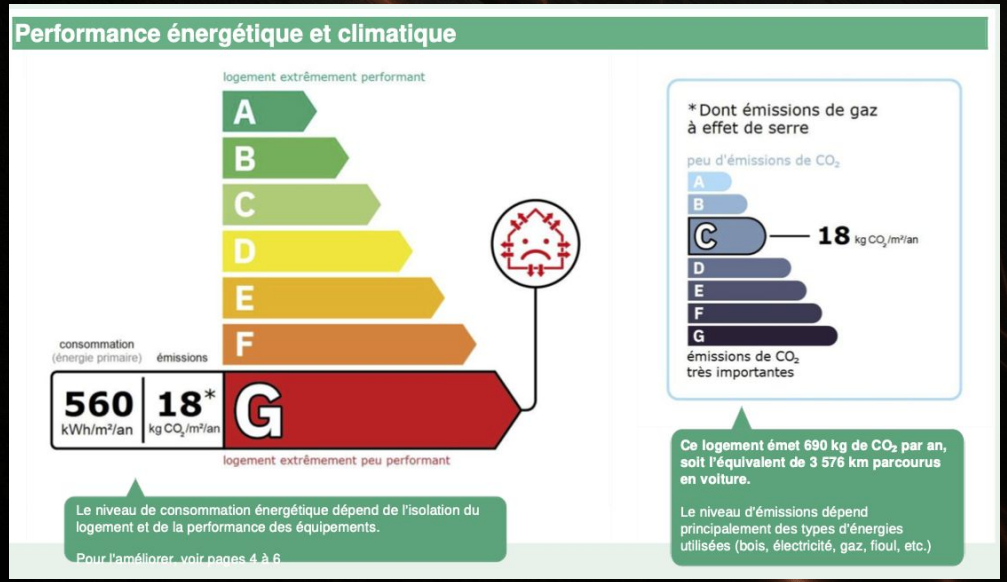
2026

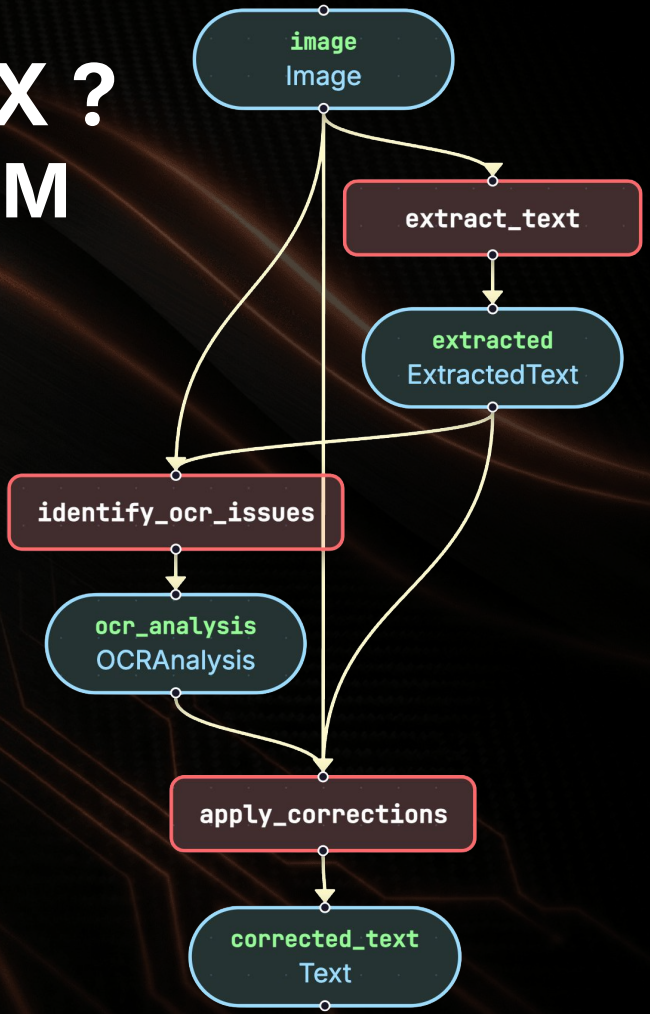
OCR vs VLM

Comparaison des Approches

OCR	VLM
$1-2 / 1K pages	$5-10 / 1K pages
Extrait texte	Comprend
Tables ✅	Tables ⚠️
Diagrammes ❌	Diagrammes ✅
Ne raisonne pas	Raisonne ✅

Besoin de la fiabilité MAX ?
🤯 utilise les deux: OCR + VLM





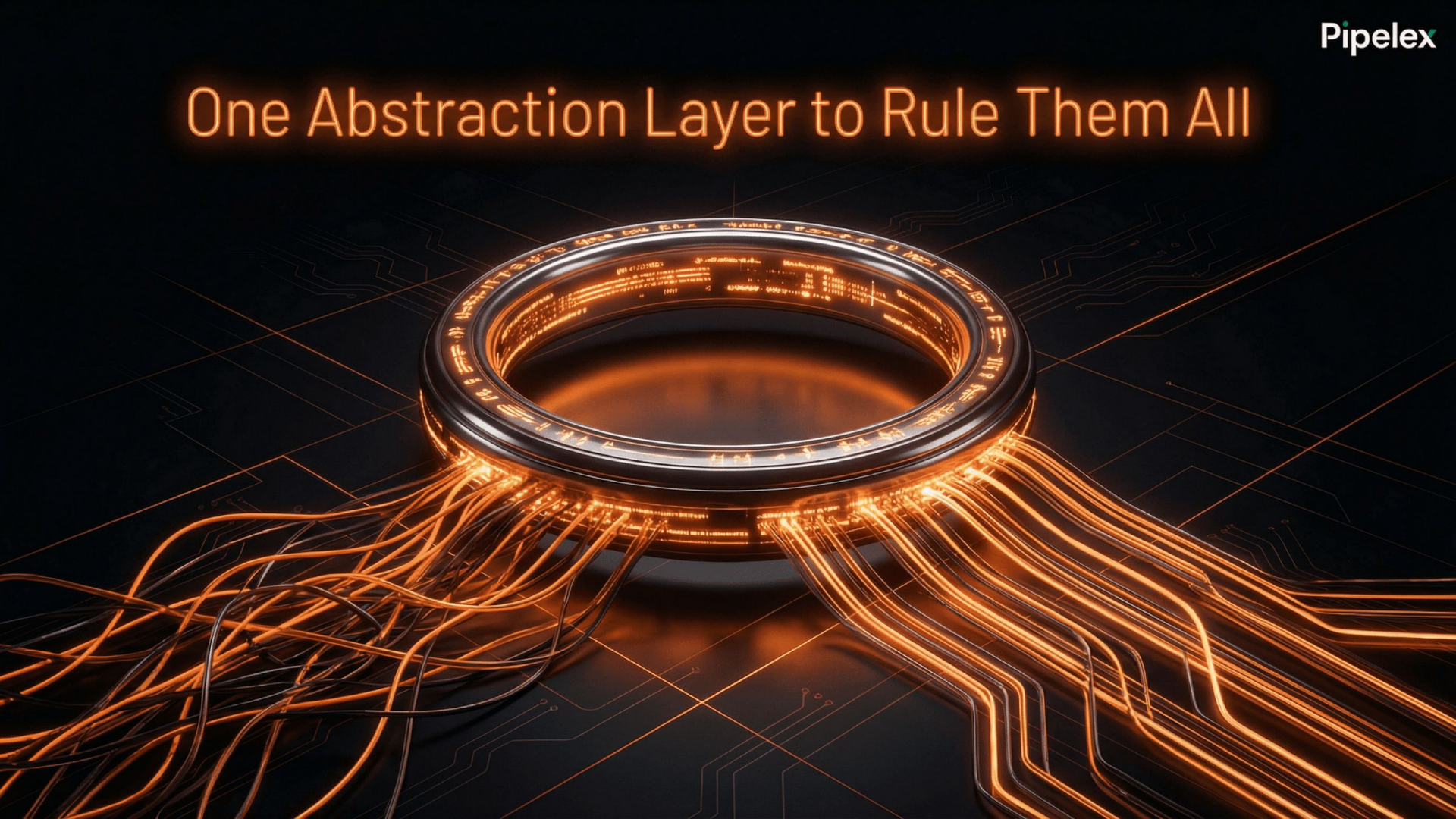
Pipelex

La Jungle des APIs

Le Chaos des Intégrations

Pipelex

One Abstraction Layer to Rule Them All



Welcome to PIPELEX GATEWAY

⭐ github.com/Pipelex/pipelex

Pour vous :

59$ Crédits

Code: **LILLE59**

Il n’y a plus qu’à poser des questions à l’IA

2024 : Répondre à une Question = RAG Search

Quand le vecteur ne comprend pas l'intention

Ce que l'utilisateur veut

« Recettes sans viande »

Plats végétariens
Rien qui ait eu un pouls

?

Ce que la recherche vectorielle renvoie

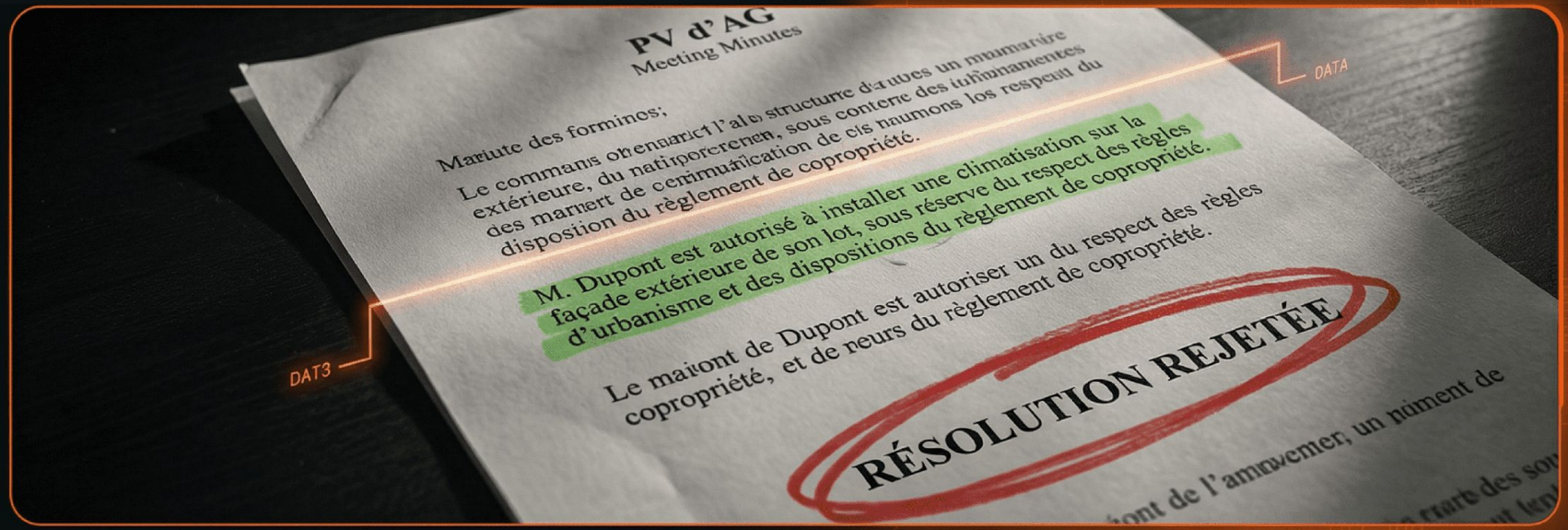
Boeuf bourguignon
Entrecôte grillée
Ragoût de viande

Exemple de piège à LLM :

qui est le propriétaire ?

Le PV d’AG : Texte vs Structure

Quand le LLM voit le texte mais rate le sens



Manager ≠ Owner

Confusion d'Entités

Question: “Pénalités pour le propriétaire ?”

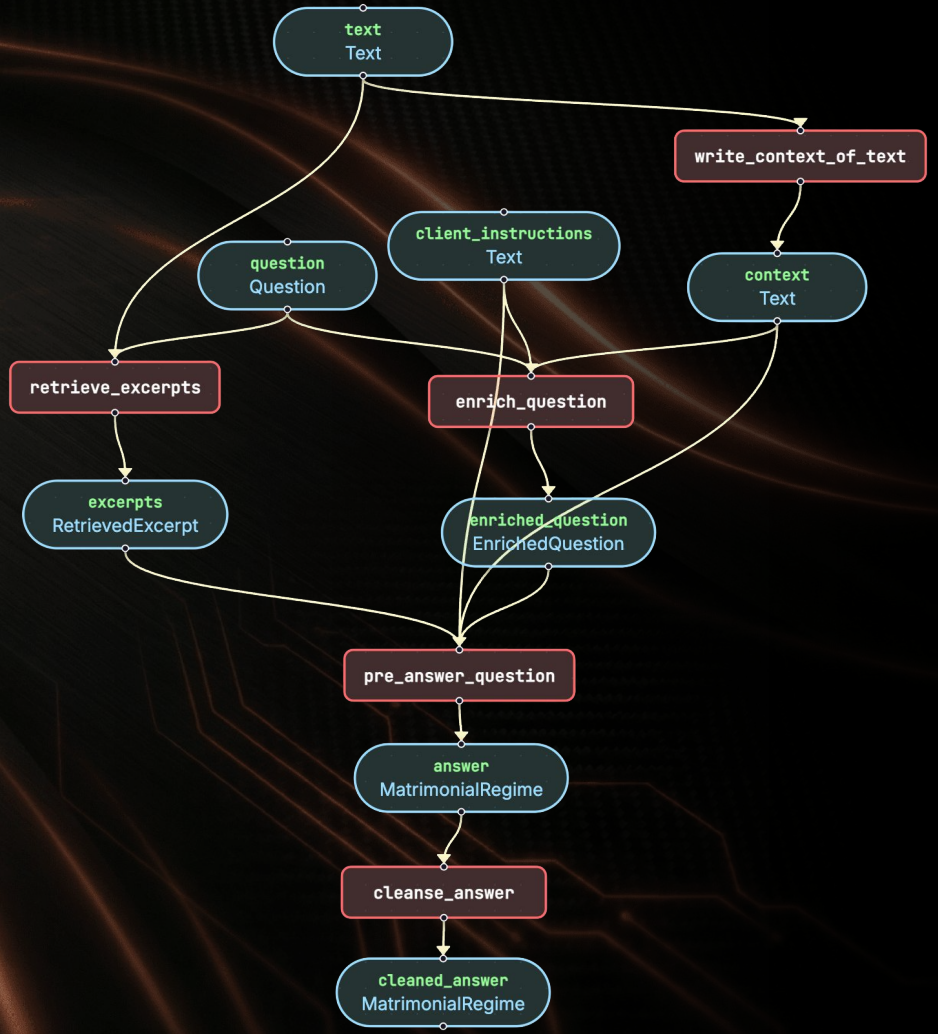
La Méthode pour Éviter les Pièges

Le Pattern Gagnant

Étape par étape

Retrieval: Gemini-Flash

Réponse: Claude 4.5



Génération Structurée

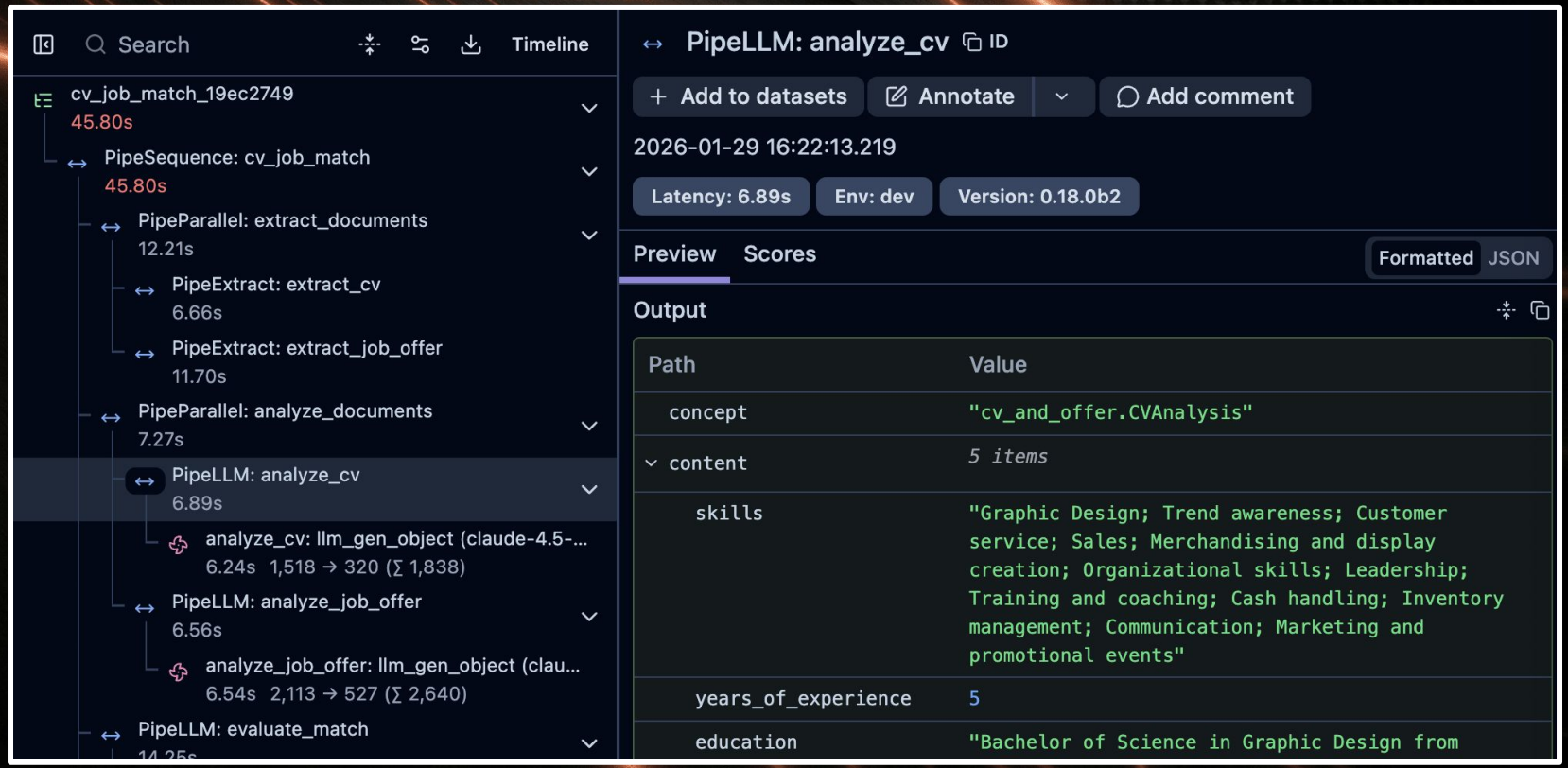
Output of pipe analyze_cv → **CVAnalysis**

Attribute	Value
skills	Graphic Design; Trend Organizational skills; Communication; Marketi
years_of_experience	5.0
education	Bachelor of Science in Cumulative GPA 3.93, D
previous_roles	Sales Associate at Ame present); Sales Associ Victoria's Secret (Jan
key_achievements	Dean's List with 3.93 Month award twice at P Ambassador at college

{ 5 Items
  skills: "Graphic Design; Trend awareness; Customer service; Sales; Merchandising and display creation; Organizational skills; Leadership; Training and coaching; Cash handling; Inventory management; Communication; Marketing and promotional events"
  years_of_experience: 5
  education: "Bachelor of Science in Graphic Design from University of Minnesota, College of Design (May 2011). Cumulative GPA 3.93, Dean's List. Twin Cities Iron Range Scholarship recipient."
  previous_roles: "Sales Associate at American Eagle (July 2009 – present); Spa Consultant at Planet Beach (Aug. 2008 – present); Sales Associate at Heartbreaker (May 2008 – Aug. 2008); Fashion Representative at Victoria's Secret (Jan. 2006 – Feb. 2009); Brand Ambassador at Target Corporation (August 2009)"
  key_achievements: "Dean's List with 3.93 GPA; Twin Cities Iron Range Scholarship recipient; Received Employee of the Month award twice at Planet Beach; Represented Periscope Marketing and Target Inc. as Brand Ambassador at college event."
}

Observabilité & Evals

LangFuse • PostHog • OpenTelemetry



Code Impératif vs. Déclaratif

Impératif : tu détailles COMMENT faire | Déclaratif : tu décris CE QUE tu veux

	Impératif	Déclaratif
Tu écris	du Glue code	**de la Business Logic**
% tokens sur la business logic	**~20%**	**~95%**
Le métier lit et peut itérer	✗	☑
Les IA progressent	Faut tout refaire	Ta méthode reste stable
Vibe-coding	**5-shot, 35 min**	**One-shot, 90 sec**

Code Impératif vs. Déclaratif

L'épreuve du feu

Use case : CV + Offre d'emploi → Analyse du match → 5 questions pour l'entretien

	Python (Impératif)	**Pipelex** (Déclaratif)
Lignes	640	**205**
Caractères	25 603	9 700
Ratio	1×	**3× moins**

"Et dans les 640 lignes Python, combien sont du glue code ?"

Code Impératif vs. Déclaratif

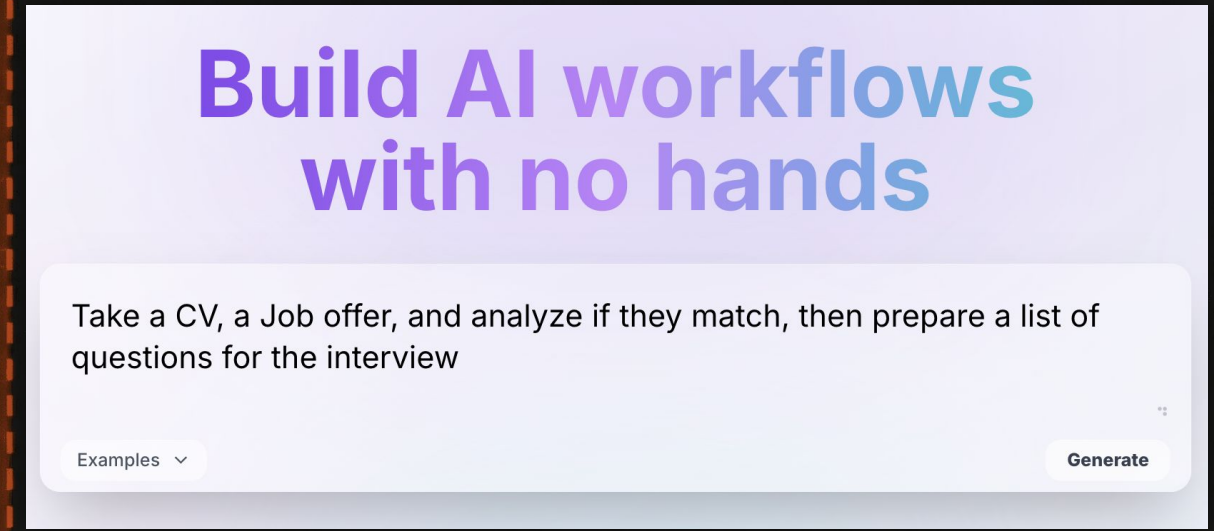
```python
async def extract_text_with_ocr(pdf_bytes: bytes, file_url: str) -> str:
    """Extract text from scanned PDF using OpenAI's native PDF support."
    logger.info("PDF appears to be scanned - using OpenAI Vision with na
    client = AsyncOpenAI()
    pdf_base64 = base64.b64encode(pdf_bytes).decode('utf-8')
    response = await client.chat.completions.create(
        model="gpt-4o",
        messages=[{
            "role": "user",
            "content": [
                {
                    "type": "text",
                    "text": "Extract all text from this PDF document. Pr
                },
                {
                    "type": "image_url",
                    "image_url": {"url": file_url}
                }
            ]
        }],
        max_tokens=4096  # Increased for multi-page PDFs
    )
    extracted_text = response.choices[0].message.content
    logger.info(f"OCR extraction complete - extracted {len(extracted_tex
    return extracted_text
```

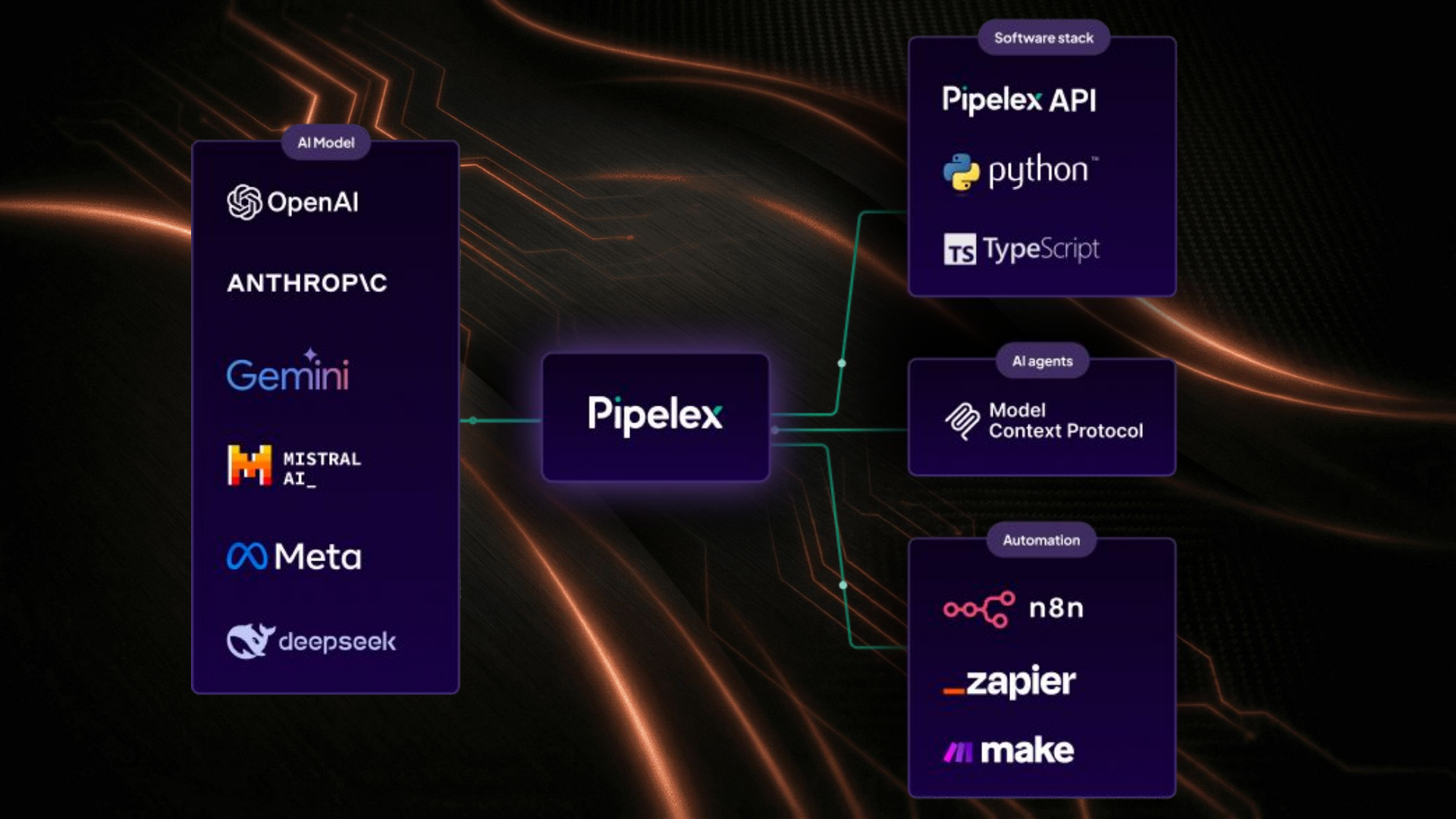
```
[pipe.extract_cv]
type = "PipeExtract"
description = "Extracts text content from the C
inputs = { cv_pdf = "Document" }
output = "Page[]"
model = "azure-document-intelligence"

[pipe.analyze_cv]
type = "PipeLLM"
description = "Analyzes the CV to extract key i
inputs = { cv_pages = "Page" }
output = "CVAnalysis"
model = "gemini-3.0-pro"
system_prompt = "You are an expert HR analyst s
prompt = """
Analyze the following CV and extract the candid
@cv_pages
"""
```

PipeBuilder

Décris ce que tu veux → Workflow en 2 minutes




AI Model
OpenAI
ANTHROP\C
Gemini
MISTRAL AI_
Meta
deepseek
Pipelex
Software stack
Pipelex API
python
TypeScript
AI agents
Model Context Protocol
Automation
n8n
zapier
make

Declarative language
for AI workflows

Dev-tool / Agent-tool

Open-Source

Pipelex

⭐ Star us: github.com/Pipelex/pipelex