



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

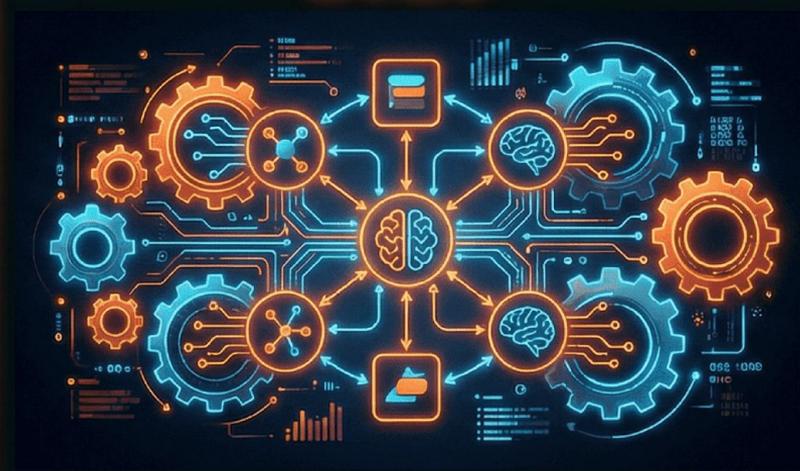
De la Question à l'Automatisation

2024



Je vais poser des questions à mes PDFs

2026



Automatiser · Donner des outils aux agents

Buy SAAS → Build Yourself

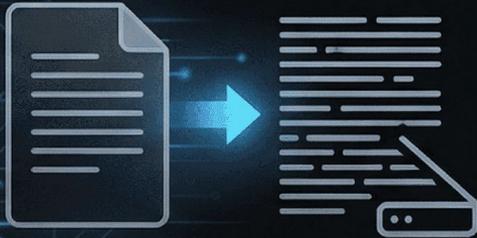


Avant de passer ça à mon LLM...

2024 vs 2026

2024

Extraire le texte du PDF



2026



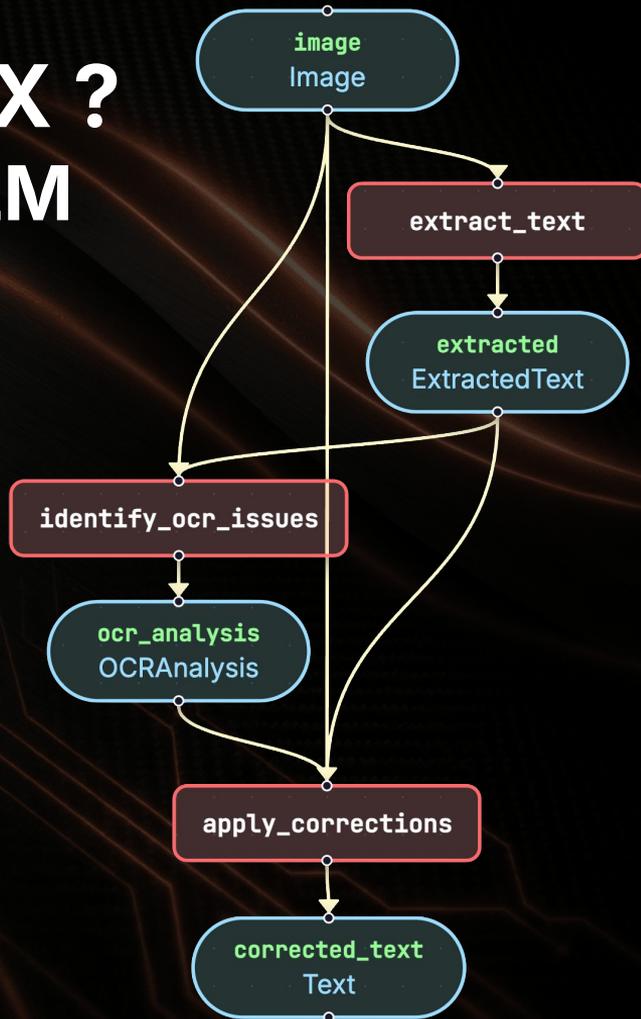
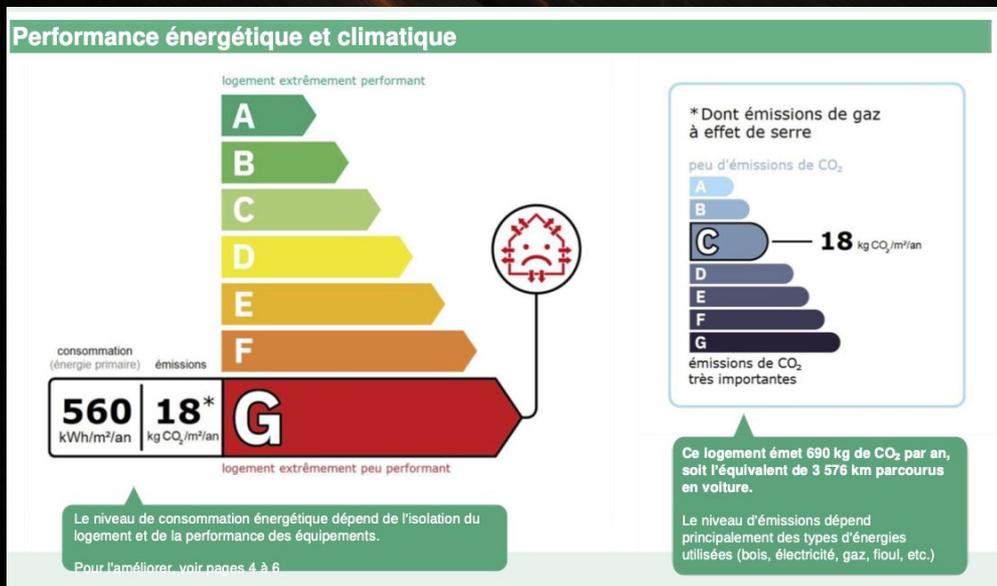
OCR vs VLM

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



La Jungle des APIs

Le Chaos des Intégrations



```
if provider == "openai":  
    # 47 lines  
elif provider == "anthropic":  
    # 52 lines
```

One Abstraction Layer to Rule Them All



Welcome to PIPELEX GATEWAY

★ github.com/Pipelex/pipelex

Pour vous :
59\$ Crédits

Code: **LILLE59**



The background features a dark, textured surface with a fine grid pattern. Overlaid on this are several glowing orange lines that resemble circuit traces or data paths. These lines are arranged in a complex, somewhat chaotic pattern, with some lines curving and others forming sharp angles. The overall effect is one of digital connectivity and futuristic technology.

**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** ?



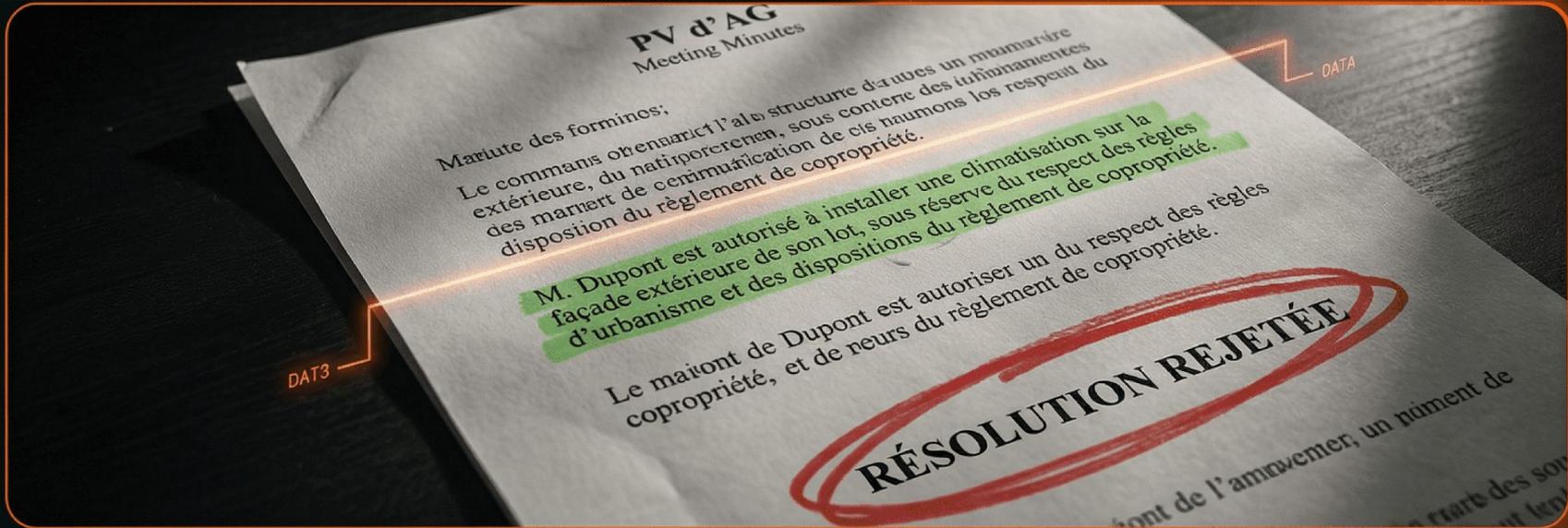
Le
propriétaire est
Jean Dupont.



Il y a
3
propriétaires.
Dont une SCI.

Le PV d'AG : Texte vs Structure

Quand le LLM voit le texte mais rate le sens

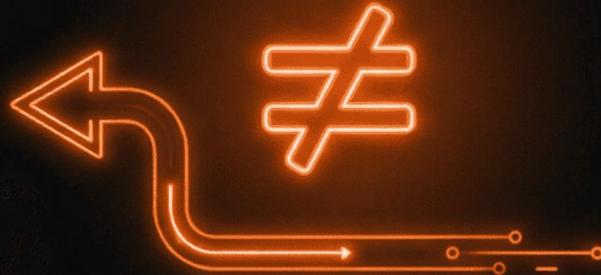


Manager \neq Owner

Confusion d'Entités



Manager



LLM



Owner

Question: "Pénalités pour le propriétaire ?"

La Méthode pour Éviter les Pièges

Le Pattern Gagnant

1. Étape par étape

2. Génération structurée

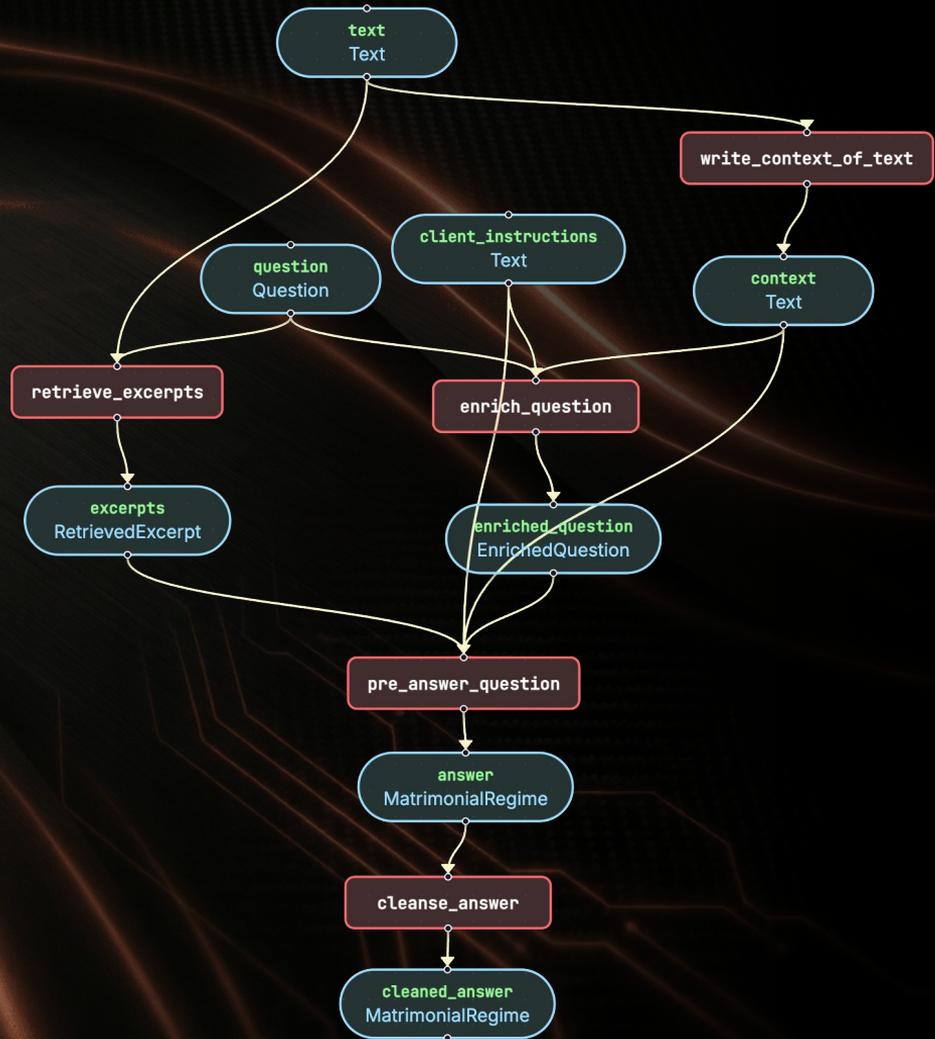
3. Contexte complet

= AI Workflow

Étape par étape

Retrieval:
Gemini-Flash

Réponse:
Claude 4.5



Génération Structurée

Output of pipe `analyze_cv` → `CVAnalysis`

Attribute	Value
<code>skills</code>	Graphic Design; Trend awareness; Organizational skills; Communication; Marketing and promotional events
<code>years_of_experience</code>	5.0
<code>education</code>	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.
<code>previous_roles</code>	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)
<code>key_achievements</code>	Dean's List with 3.93 GPA; Received Employee of the Month award twice at Planet Beach; Represented Periscope Marketing and Target Inc. as Brand Ambassador at college event.

```
{ 5 Items v
  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

The screenshot displays a pipeline monitoring interface. On the left, a tree view shows the execution flow of a pipeline named 'cv_job_match_19ec2749'. The pipeline consists of several steps: a 'PipeSequence' (45.80s), a 'PipeParallel' step containing 'extract_documents' (12.21s), 'extract_cv' (6.66s), and 'extract_job_offer' (11.70s), followed by another 'PipeParallel' step containing 'analyze_documents' (7.27s) and 'analyze_cv' (6.89s). The 'analyze_cv' step is highlighted, and its details are shown on the right. The details include the step name 'PipeLLM: analyze_cv', a timestamp '2026-01-29 16:22:13.219', and performance metrics: 'Latency: 6.89s', 'Env: dev', and 'Version: 0.18.0b2'. Below these are tabs for 'Preview' and 'Scores', with 'Formatted' and 'JSON' options. The 'Output' section shows a table with the following data:

Path	Value
concept	"cv_and_offer.CVAnalysis"
content	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

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

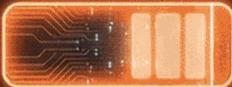
Code Impératif vs. Déclaratif

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	1x	3x moins

Python:  640 lignes

Pipelex:  **205 lignes**

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

Code Impératif vs. Déclaratif

```
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 native PDF support")
    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. Print the text in a single line."
                    },
                    {
                        "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_text)} characters")
    return extracted_text
```

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

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

@cv_pages
"""
```

PipeBuilder

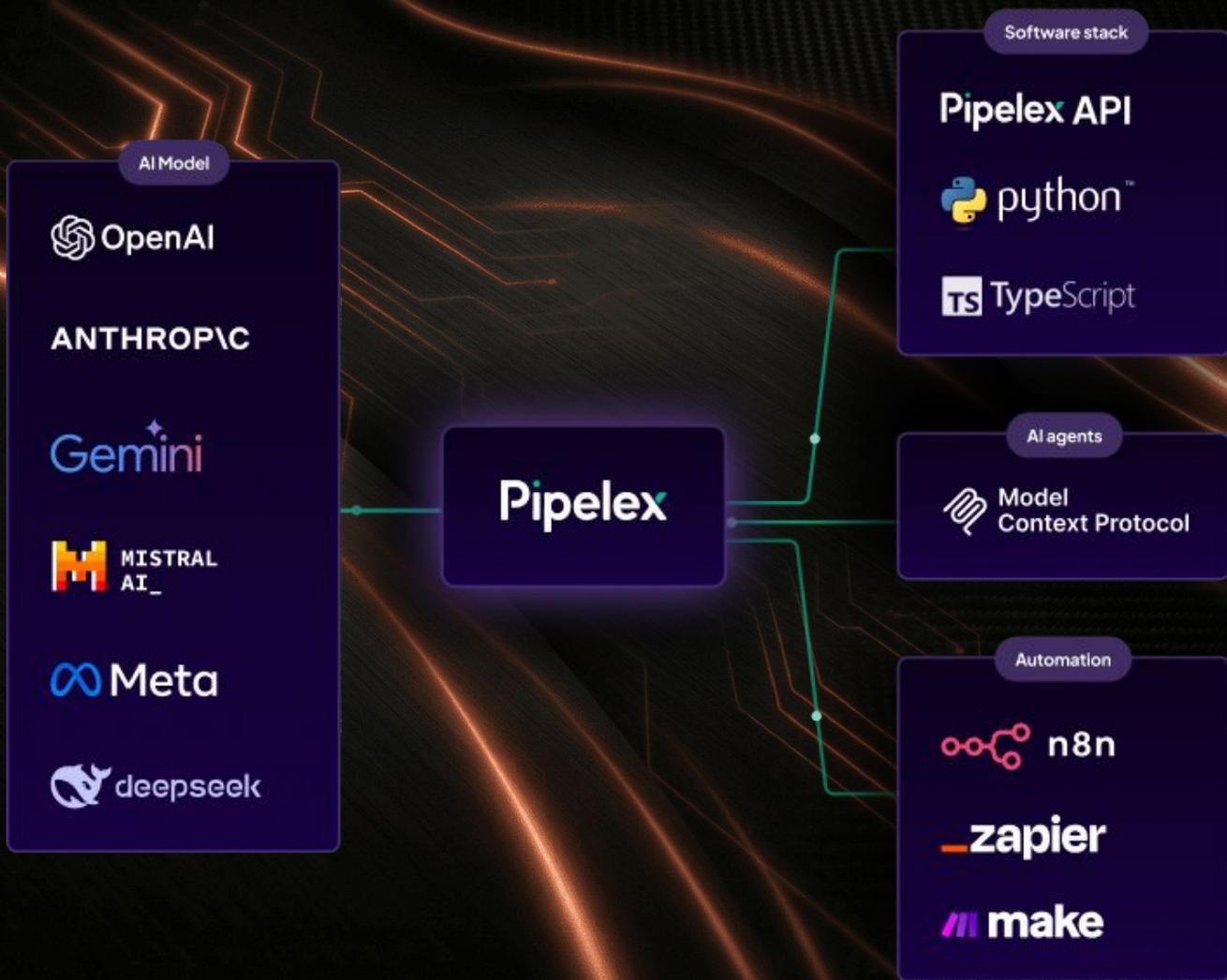
Décris ce que tu veux → Workflow en 2 minutes

Build AI workflows with no hands

Take a CV, a Job offer, and analyze if they match, then prepare a list of questions for the interview

Examples ▾

Generate



Declarative language
for **AI workflows**

Dev-tool / Agent-tool

Open-Source

Pipelex



★ Star us: github.com/Pipelex/pipelex