# PHÂN TÍCH BÀI TOÁN THEO TƯ DUY AI

## Mô tả bài toán thực tế

Trong bối cảnh thị trường tuyển dụng ngày càng cạnh tranh, nhà tuyển dụng cần đánh giá ứng viên một cách nhanh chóng, hiệu quả và chính xác. Tuy nhiên, việc đặt câu hỏi phỏng vấn phù hợp với từng hồ sơ ứng viên lại đòi hỏi thời gian, sự chuẩn bị kỹ lưỡng và hiểu biết chuyên sâu về vị trí tuyển dụng.

**Mục tiêu: (ko đi sâu vào tech)**

* Tự động phân tích hồ sơ ứng viên và trích xuất thông tin quan trọng
* Sinh ra các câu hỏi phỏng vấn cá nhân hóa dựa trên nội dung CV và yêu cầu công việc
* Đảm bảo tính công bằng, khách quan trong quá trình đánh giá
* Tiết kiệm thời gian và nâng cao hiệu quả tuyển dụng
* Cải thiện trải nghiệm phỏng vấn cho cả nhà tuyển dụng và ứng viên
* Xây dựng hệ thống tự động gợi ý câu hỏi phỏng vấn dựa trên phân tích hồ sơ ứng viên.
* Tiết kiệm thời gian cho nhà tuyển dụng.
* Nâng cao chất lượng phỏng vấn với các câu hỏi cá nhân hóa.
* Đề xuất câu hỏi đánh giá tiềm năng, không chỉ dựa trên kinh nghiệm hiện có.

**Đối tượng người dùng chính: (2-3 loại)**

Công ty tuyển dụng, nền tảng tuyển dụng (VD: TopCV, VietnamWorks)

**•** Nhà tuyển dụng (HR): Những người chịu trách nhiệm sàng lọc và phỏng vấn ứng viên

**•** Quản lý tuyển dụng: Những người điều phối quá trình tuyển dụng

**•** Hội đồng phỏng vấn: Các chuyên gia kỹ thuật tham gia đánh giá ứng viên

## Phân tích bài toán theo AI Thinking

### Mục tiêu AI cần đạt

* Trích xuất thông tin tự động: Phân tích CV để nhận diện kỹ năng, kinh nghiệm, học vấn, dự án đã thực hiện
* Đối chiếu với yêu cầu công việc: So sánh thông tin ứng viên với job description để xác định điểm mạnh/yếu
* Sinh câu hỏi thông minh: Tạo ra các câu hỏi phỏng vấn có mục tiêu rõ ràng, đánh giá đúng năng lực cần thiết
* Cá nhân hóa cao: Đảm bảo mỗi bộ câu hỏi phù hợp với profile riêng biệt của từng ứng viên
* Loại bỏ thiên lệch: Đảm bảo câu hỏi không mang tính phân biệt đối xử về giới tính, tuổi tác, xuất thân
* Tích hợp dễ dàng: Giao diện thân thiện, dễ sử dụng cho nhà tuyển dụng
* **Hiểu Hồ Sơ Ứng Viên:** Trích xuất thông tin quan trọng từ CV và portfolio.
* **Xác Định Điểm Mạnh/Yếu:** Phát hiện các kỹ năng nổi bật và khoảng trống kinh nghiệm.
* **Gợi Ý Câu Hỏi Phù Hợp:** Tạo câu hỏi đánh giá năng lực ứng viên dựa trên yêu cầu công việc và hồ sơ cá nhân.
* **Cá Nhân Hóa:** Đề xuất câu hỏi khác nhau cho cùng vị trí tùy theo background ứng viên

### Phương pháp AI phù hợp

**Xử lý ngôn ngữ tự nhiên (NLP):**

* Trích xuất thực thể từ CV (Named Entity Recognition)
* Phân loại kỹ năng, lĩnh vực chuyên môn

**Hệ thống gợi ý (Recommendation System):**

* Dựa trên rule-based (luật) hoặc content-based filtering
* Có thể áp dụng GPT hoặc mô hình sinh (generative models) để đề xuất câu hỏi
* **Xử Lý Ngôn Ngữ Tự Nhiên (NLP):**

Phân tích CV với Named Entity Recognition (NER) để trích xuất thông tin như tên, học vấn, và kinh nghiệm.

Phân tích mô tả công việc để xác định kỹ năng cần thiết.

* **Hệ Thống Gợi Ý (Recommendation System):**
* **Content-based Filtering:** Gợi ý câu hỏi dựa trên sự tương đồng giữa hồ sơ ứng viên và yêu cầu công việc.
* **Knowledge-based:** Sử dụng bộ câu hỏi mẫu đã được phân loại theo kỹ năng và mức độ.
* **Mô Hình Sinh Văn Bản (Text Generation):**
* Sử dụng các mô hình ngôn ngữ lớn như GPT-3 để sinh câu hỏi phỏng vấn phù hợp dựa trên thông tin ứng viên và yêu cầu công việc.

***Natural Language Processing (NLP):***

* Sử dụng các kỹ thuật NLP tiên tiến để:
* Trích xuất thực thể có tên (Named Entity Recognition): Nhận diện tên công ty, công nghệ, kỹ năng
* Phân tích cảm tính (Sentiment Analysis): Đánh giá thái độ và động lực làm việc
* Phân loại văn bản (Text Classification): Categorize kinh nghiệm theo lĩnh vực
* Trích xuất từ khóa (Keyword Extraction): Xác định những kỹ năng và công nghệ quan trọng

***Machine Learning & Deep Learning: (cần cân nhắc filter lại)***

* Áp dụng các mô hình học máy để:
* Mô hình phân loại (Classification): Xếp loại ứng viên theo mức độ phù hợp
* Recommendation System: Gợi ý câu hỏi dựa trên similar profiles
* Sequence-to-Sequence Models: Sinh câu hỏi từ thông tin đầu vào
* Transfer Learning: Sử dụng pre-trained models như BERT, GPT để hiểu ngữ cảnh tốt hơn (dùng 3rd party để hỗ trợ việc training AI) -> rút ngắn thời gian train…

***Rule-based System:***

* Kết hợp với hệ thống luật để đảm bảo:
* Tuân thủ quy định pháp lý về tuyển dụng
* Áp dụng best practices trong phỏng vấn
* Kiểm soát chất lượng câu hỏi được sinh ra
* Đảm bảo tính nhất quán trong quy trình đánh giá

### **Dữ liệu cần có, nguồn gốc, rủi ro**

**Dữ Liệu Cần Có:**

Dataset CV ứng viên (định dạng PDF, Word).

Bộ mô tả công việc cho các vị trí phổ biến.

Bộ câu hỏi phỏng vấn mẫu được gán nhãn theo loại kỹ năng và mức độ.

Dữ liệu phản hồi từ nhà tuyển dụng về chất lượng câu hỏi.

**Nguồn Gốc:**

Public dataset từ Kaggle, GitHub. (dataset abc lấy từ .. mô tả thông tin của dataset đó?)

Thu thập từ website tuyển dụng.

Tạo dataset giả lập nếu thiếu dữ liệu thật.

**Rủi Ro:**

Thiếu dữ liệu đa dạng.

Thông tin cá nhân trong CV.

Bias trong dữ liệu.

Chất lượng dữ liệu không đồng đều.

### Công cụ, mô hình dự kiến sử dụng

Xử lý dữ liệu:

* Thư viện NLP:
  + SpaCy: Phân tích cú pháp, NER (trích xuất tên, skill, công ty...) - Ví dụ: Nhận diện: "Python" → SKILL, "Google" → ORG
  + NLTK: Tokenization, stemming, stopwords removal - Ví dụ: Chuẩn hóa văn bản trước khi đưa vào model
  + StanfordNLP:
    - POS Tagging, Dependency Parsing (cho ngôn ngữ phức tạp như tiếng Việt) - Ví dụ: Xác định ngữ cảnh của từ trong CV
    - OCR: Tesseract (cho CV scan)
  + PDF/Word parser:
    - Apache PDFBox: Trích xuất text từ CV PDF
    - python-docx: Đọc thư xin việc định dạng Word

Mô hình AI:

* Trích xuất thông tin:
* BERT-NER: Transformer-based - Độ chính xác cao, hỗ trợ đa ngôn ngữ
* SpaCy NER: CNN + CRF - Nhẹ, tốc độ xử lý nhanh
* Phân loại văn bản:
* FastText: Phân loại nhanh JD theo ngành nghề
* BERT Classification: Phân loại chi tiết (vị trí, seniority...)
* Hệ thống gợi ý:
* Cosine similarity: So khớp CV và JD dựa trên TF-IDF vectors
* TF-IDF + KNN: Tìm câu hỏi tương tự trong database
* Sinh câu hỏi:
* GPT-3/4: OpenAI API - Sinh câu hỏi tự nhiên, không cần training
* BART, T5 (fine-tuned): Hugging Face - Fine-tune được với dataset riêng

### Bối cảnh xã hội / đạo đức nghề nghiệp

* **Quyền riêng tư**: Dữ liệu CV chứa thông tin cá nhân nhạy cảm → cần đảm bảo bảo mật, không lưu trữ lâu dài nếu không được phép.
* **Thiên lệch (bias):** Gợi ý câu hỏi không được mang định kiến giới tính, sắc tộc, tuổi tác.
* **Tính công bằng:** Hệ thống cần đảm bảo công bằng giữa các ứng viên, không thiên vị dựa trên nền tảng học vấn hoặc tên trường.

**Đạo Đức:** Minh bạch trong cách hệ thống đưa ra đề xuất và cho phép nhà tuyển dụng chỉnh sửa câu hỏi.

* Quyền riêng tư:
* Ẩn danh hóa thông tin cá nhân (tên, số điện thoại, email...)
* Chỉ phân tích nội dung liên quan đến chuyên môn
* Tuân thủ GDPR và các quy định bảo vệ dữ liệu
* Thiên lệch (Bias):
* Kiểm tra bias trong đề xuất câu hỏi (không thiên vị giới tính, độ tuổi...)
* Đa dạng hóa dataset để tránh tập trung vào một nhóm ứng viên
* Giám sát bằng thuật toán phát hiện bias (AIF360)
* Đạo đức:
* Minh bạch về cách hệ thống đưa ra đề xuất
* Cho phép nhà tuyển dụng chỉnh sửa câu hỏi đề xuất
* Không sử dụng thông tin nhạy cảm để phân biệt đối xử

# THIẾT KẾ MÔ HÌNH & PIPELINE

## Phân tích bài toán theo AI Thinking

**Input:**

* Thông tin vị trí tuyển dụng (mô tả công việc, yêu cầu kỹ năng, mức độ kinh nghiệm cần thiết).
* File CV (PDF, DOC hoặc dạng text)
* Độ khó mong muốn của câu hỏi, ngữ cảnh của buổi pv

**Output:**

Danh sách 5–10 câu hỏi phỏng vấn được gợi ý, có thể chia theo nhóm:

* Câu hỏi kỹ thuật/chuyên môn
* Câu hỏi tình huống
* Câu hỏi kiểm tra kiến thức nền
* Câu hỏi kiểm tra tư duy, sáng tạo

Lý do đề xuất câu hỏi và độ phù hợp của từng câu hỏi.

## 1. Input – Output cụ thể

|  |  |
| --- | --- |
| Thành phần | Nội dung chi tiết |
| **INPUT** |  |
| Hồ sơ ứng viên | • CV/Resume (PDF, DOC, DOCX) • LinkedIn profile URL • Portfolio links • Cover letter • Certificates và achievements |
| Mô tả công việc | • Job title và level (Junior/Mid/Senior) • Required skills và technologies • Years of experience needed • Company culture information • Team structure và working style |
| Cấu hình phỏng vấn | • Interview type (Technical/Behavioral/Cultural fit) • Time allocation (30min/45min/1hour) • Number of questions needed • Difficulty level preference • Focus areas (algorithms, system design, etc.) |
| Dữ liệu bổ trợ | • Company question bank • Industry standards • Previous interview feedback • Role-specific templates |
| **OUTPUT** |  |
| Câu hỏi phỏng vấn | • Danh sách 10-15 câu hỏi cá nhân hóa • Phân loại theo category (technical, behavioral, situational) • Estimated time per question • Follow-up questions suggestions |
| Báo cáo phân tích | • CV analysis summary • Strengths và gaps identification • Fit score với job requirements • Risk factors và red flags |
| Gợi ý chiến lược | • Interview flow recommendations • Areas to probe deeper • Potential concerns to address • Backup questions for different scenarios |
| Export options | • PDF report với branded template • Word document cho editing • Integration với ATS systems • Email template với questions |