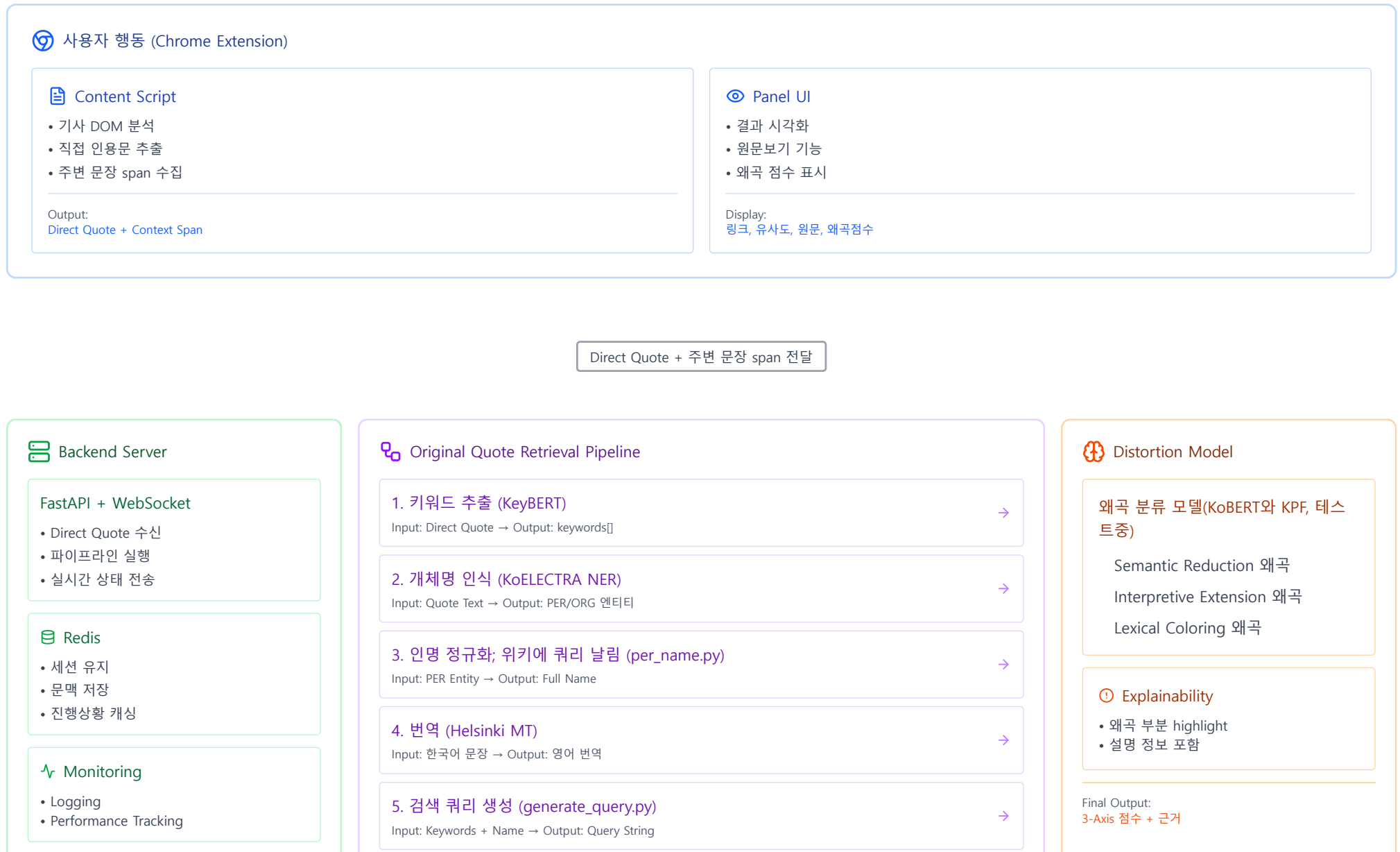


인용문 왜곡 탐지 시스템 아키텍처



6. Google Search API 호출

Input: Query String → Output: Search Results[]



7. Span 유사도 평가 & Alignment

Input: Results[] → Output: Best Match + Score

Method: SBERT, cosine 유사도

Backend Data Layer (Faster Retrieval)

Local Speech Database

- 대표적인 정치인 연설 녹취록 저장
- 예: Donald Trump speeches

Fields:

- id, speaker, date
- span, embedding vector
- source metadata

Type:
PostgreSQL / MongoDB

Vector Search Engine

- Semantic similarity search
- 빠른 span 매칭
- Pre-hit 체크 수행

Input/Output:

Input: span embedding

Output: candidate spans + similarity scores

Technology:
FAISS / Elasticsearch

Redis Cache Layer-필요시 넣을 예정

- 매칭된 span 캐싱
- 외부 검색 결과 저장
- 반복 쿼리 가속화

Features:

- TTL-based eviction
- Frequently matched results

Acceleration:
10x faster for cached queries

Data Flow & Logic

Preprocessing (Pipeline Steps 1-4):

1. Keywords
(KeyBERT)

2. NER
(PER/ORG)

3. Normalize
(Names/Dates)

4. Translate
(Helsinki)

Output: Normalized Names + Dates + Keywords

1. Receive Extracted Features

Names, Dates, Keywords



2. Check Cache

Query by extracted features

Quote Distortion Detection Diagram

