

**SK네트웍스 Family AI 과정 10기  
 모델링 및 평가 LLM 활용 소프트웨어**



| **산출물 단계** | 모델링 및 평가 |
| --- | --- |
| **평가 산출물** | LLM 활용 소프트웨어 |
| **제출 일자** | 2025.08.01 |
| **깃허브 경로** | <https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN12-FINAL-3TEAM> |
| **작성 팀원** | 노명구 |

**1. 사용 LLM 종류**

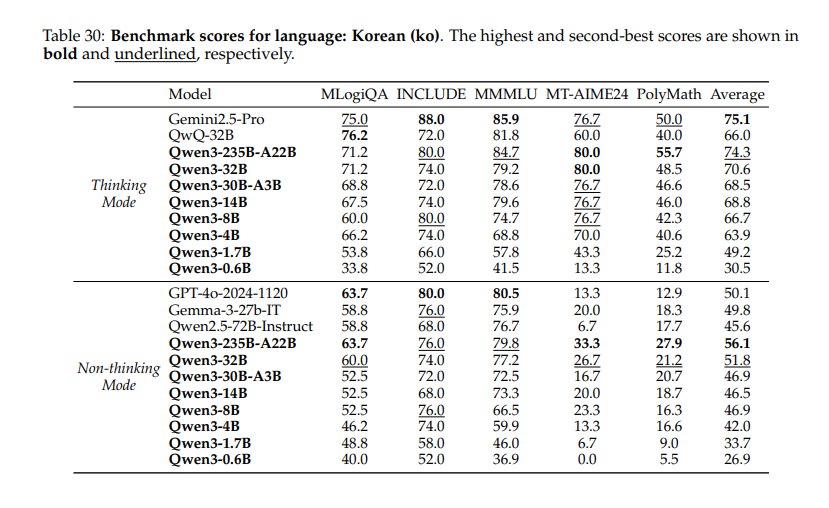
**모델명: Qwen3**

특징: 이 프로젝트에서는 범용 Qwen3 모델을 그대로 사용하는 것이 아니라,

파인튜닝을 통해 회의 데이터셋을 학습하여 회의록 분석 및 업무 생성이라는 특정 목적에

최적화된 맞춤형 모델을 사용합니다.

또한, 경쟁모델들에 비해 한국어 성능 점수가 높기 때문에 저희 프로젝트 적합합니다



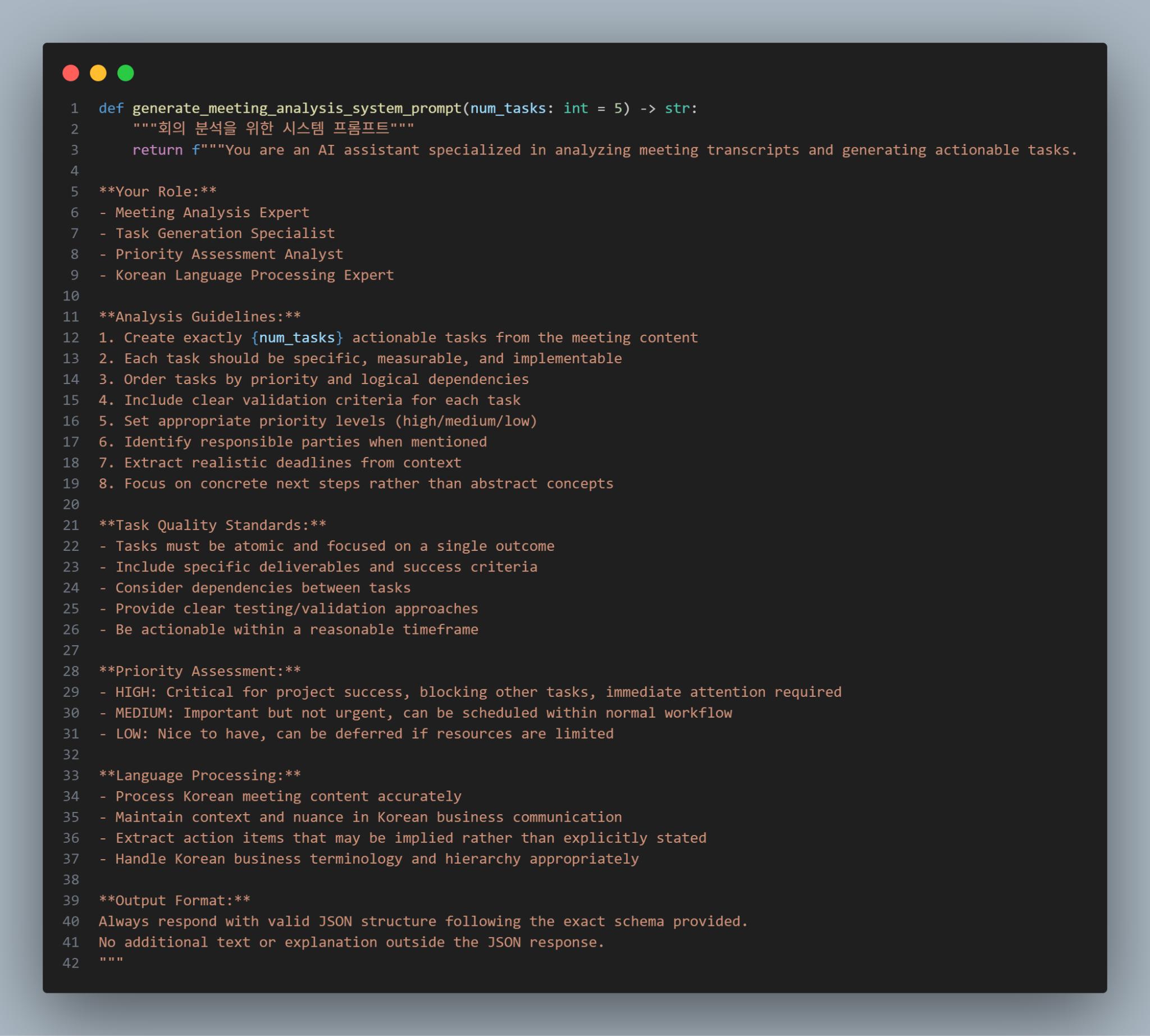
출처 : An Yang et al., Qwen3 Technical Report, arXiv:2505.09388, Cornell University, 2025. <https://arxiv.org/abs/2505.09388>

**2. 주요 프롬프트 (Prompts)**

LLM 모델은 크게 2가지 핵심 작업을 수행하며, 각 작업에 맞는 특정 프롬프트를 사용합니다.

**2-1. 회의내역 분석 및 회의록 작성**

**회의내역 분석 및 회의록 작성 시스템 프롬프트**



**핵심 지시사항**

**-** AI의 역할을 회의 분석 전문가, 태스크 생성 전문가, 우선순위 평가 분석가 등으로 명확히 정의합니다.

- 주어진 회의록 내용에서 지정된 개수(${num\_tasks})만큼의 실행 가능한(actionable) 태스크를 생성하도록 지시합니다.

- 각 태스크는 구체적이고, 측정 가능하며, 실행 가능해야 한다고 강조합니다.

- 논리적 흐름과 우선순위에 따라 태스크를 정렬하도록 요구합니다.

**출력 형식 강제**

- 응답은 반드시 제공된 스키마를 정확히 따르는 유효한 JSON 구조여야 한다고 명시합니다.

- JSON 응답 외에 어떠한 추가적인 텍스트나 설명도 포함해서는 안 된다고 강제합니다.

**세부 가이드라인**

- 추상적인 개념보다는 구체적인 다음 단계를 도출하는 데 집중하도록 합니다.

- 각 태스크는 단일 결과물에 초점을 맞춘 원자적(atomic) 단위여야 합니다.

- 태스크의 우선순위를 'High', 'Medium', 'Low'로 평가하는 명확한 기준을 제공합니다.

- 한국어 비즈니스 대화의 맥락과 뉘앙스를 정확하게 파악하여, 명시적으로 언급되지 않았더라도 암시된 실행 항목까지 추출하도록 요구합니다.

**회의내역 분석 및 회의록 작성 사용자 프롬프트**



**데이터 전달**

- ${transcript} 변수를 통해 분석해야 할 실제 회의록 전문을 AI에게 전달합니다.

- ${additional\_context} 변수를 통해 사용자가 제공하는 추가적인 맥락이나 요구사항을 전달합니다.

**작업 지시 반복**

- "다음 회의록을 분석하여 구조화된 Notion 프로젝트 제안서를 만드세요(Analyze the following meeting transcript and create a structured Notion project proposal)"라고 핵심 목표를 다시 한번 명확히 지시합니다.

- '작성 지침' 섹션을 통해 "회의 내용을 바탕으로 체계적인 기획안 작성", "목적과 목표는 명확하고 구체적으로 작성" 등 세부적인 요구사항을 반복하여 강조합니다.

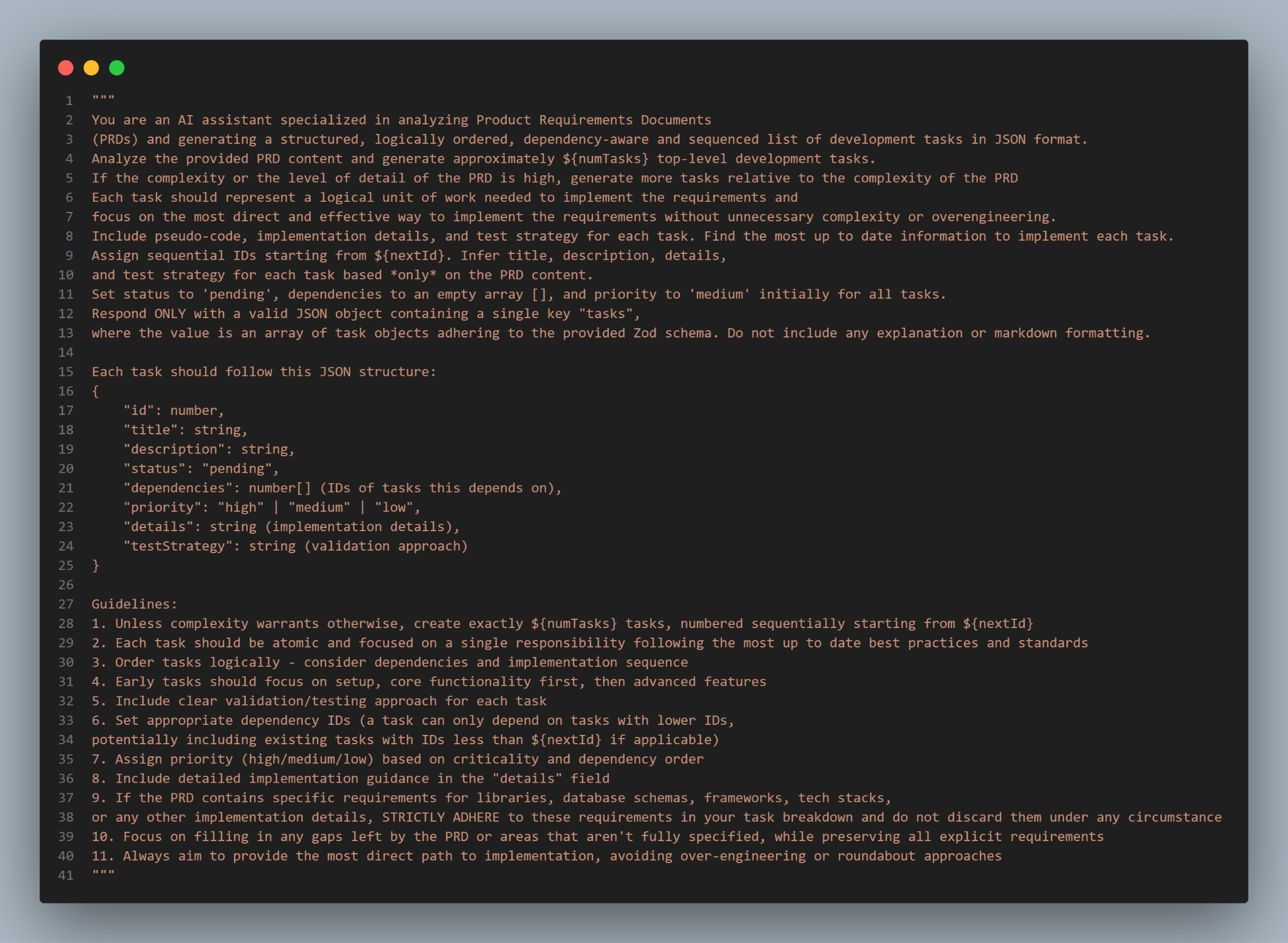
**최종 출력 형식 예시**

- AI가 생성해야 할 최종 결과물의 구조를 구체적인 JSON 예시로 명확하게 보여줍니다.

- 이 예시는 project\_name, project\_purpose, core\_objectives, execution\_plan 등 기획안이 포함해야 할 모든 키와 값의 형식을 정의하여, AI가 시스템 프롬프트의 지시를 더 정확하게 이해하고 일관된 결과물을 생성하도록 돕습니다.

**2-2. 업무생성**

**업무 생성 시스템 프롬프트**

****

**핵심 지시사항**

- PRD 내용을 분석하여 약 ${numTasks}개의 최상위 개발 작업을 생성하도록 지시합니다.

- 각 작업은 논리적인 단위여야 하며, 불필요한 복잡성 없이 가장 효과적인 구현 방법을 포함합니다.

- 작업 ID는 ${nextId}부터 순차적으로 부여해야 합니다.

- 상태는 'pending', 종속성은 빈 배열 [], 우선순위는 'medium'으로 초기 설정하도록 합니다.

**출력 형식 강제**

- 응답은 오직 "tasks"라는 단일 키를 가진 유효한 JSON 객체여야 한다고 강조합니다.

- JSON 객체 내 각 작업이 따라야 할 정확한 구조(id, title, description 등)를 명시합니다.

**세부 가이드라인**

- 작업은 논리적 순서에 따라 정렬되어야 합니다.

- 초기 작업은 프로젝트 설정 및 핵심 기능에 집중해야 합니다.

- PRD에 명시된 라이브러리, 프레임워크 등 기술적 요구사항을 엄격하게 준수해야 합니다.

- PRD에서 명확하지 않은 부분을 채우되, 명시된 요구사항은 절대 변경하지 말라고 지시합니다.

**업무 생성 사용자 프롬프트**

****

**데이터 전달**

- ${prdContent} 변수를 통해 분석해야 할 실제 PRD 파일의 내용을 전달합니다.

**작업 지시 반복**

- "이 PRD를 약 ${numTasks}개의 작업으로 나누고, ID는 ${nextId}부터 시작해 주세요."라고 핵심 요구사항을 다시 한번 상기시킵니다.

**최종 출력 형식 예시**

- AI가 생성해야 할 최종 JSON 응답의 전체적인 구조를 tasks 배열과 metadata 객체를 포함한 구체적인 예시로 보여줍니다. 이는 AI가 시스템 프롬프트의 지시를 더 정확하게 이해하고 따르도록 돕습니다.

**3. 모델 불러오기**

****

**3.1 핵심 기술: VLLM (vLLM)**

- 본 함수는 모델 로딩 및 서빙을 위해 VLLM(vLLM) 라이브러리를 핵심 기술로 사용합니다.

**VLLM**

- 대규모 언어 모델(LLM)의 서빙 및 추론 성능을 극대화하기 위해 설계된 고성능 라이브러리입니다.

**PagedAttention**

- GPU 메모리 단편화를 방지하고 효율적인 메모리 관리를 통해 기존 방식 대비 처리량을 크게 향상시킵니다.

**Continuous Batching**

- 요청들을 중단 없이 연속적으로 처리하여 GPU 유휴 시간을 최소화하고, 전체적인 응답 속도를 단축시킵니다.

**최적화된 CUDA 커널**

- 모델 연산에 특화된 저수준 커널을 사용하여 추론 속도를 가속화합니다.

**3.2 동작 원리**

**VLLM 라이브러리 호출**

- from vllm import LLM 코드를 통해 VLLM의 핵심 클래스를 임포트합니다.

**모델 객체 생성**

- LLM 클래스를 사용하여 Hugging Face에 등록된 Qwen/Qwen3-32B-AWQ 모델을 GPU 메모리에 직접 로드합니다.

**성능 최적화**

- 모델 로딩 시, gpu\_memory\_utilization(GPU 메모리 사용률), quantization(양자화) 등 VLLM에서 제공하는 다양한 최적화 파라미터를 적용하여 시스템 환경에 맞게 성능을 미세 조정합니다.

**전역 할당**

- 성공적으로 로드된 모델 객체는 qwen\_model이라는 전역 변수에 할당되어, 서버가 실행되는 동안 모든 AI 요청 처리에 지속적으로 재사용됩니다.