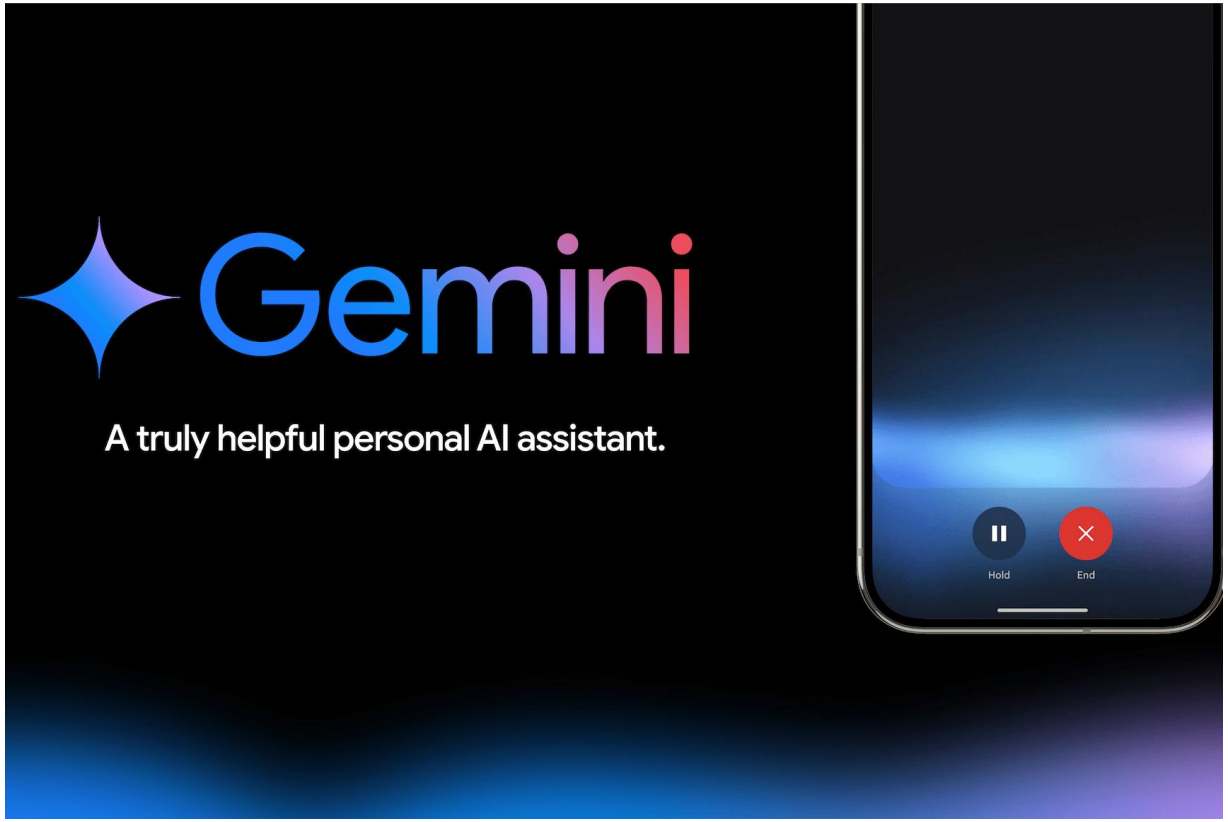
Gemini 2.0 Live

2025.12.30

SNU IDEA Lab.
정유립

01 Gemini 2.0 Live 소개



기술적 요소	기존 LLM 파이프라인	Gemini 2.0 Live (Native)
입력 처리	STT → Text	Audio Token 직접 처리
출력 처리	Text → TTS	Audio Token 직접 생성
지연 시간	수 초 (변환 과정 소요)	Sub-second (수백 밀리초)
비언어적 요소	대부분 소실 (텍스트 위주)	보존 및 표현 (감정, 톤)
비디오 이해	정지 이미지 분석 위주	실시간 스트림/시간적 맥락 이해

01 Gemini 2.0 Live 소개

모델을 실시간 대화 인터페이스로 구현하기 위해 추가된 아키텍처 및 전송 기술

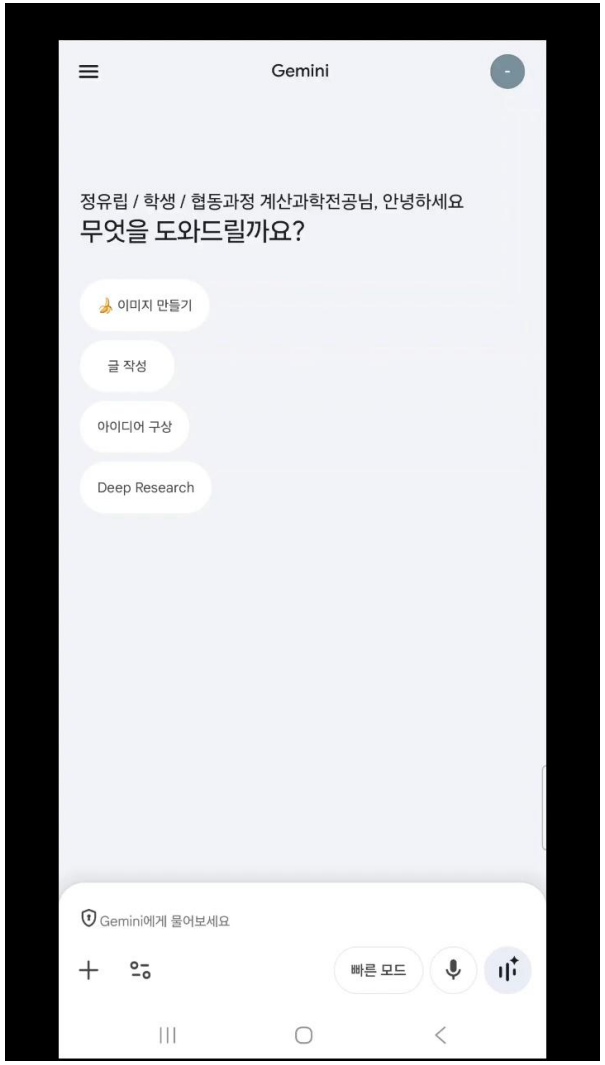
- **Streaming Speech-to-Speech (S2S)**:
  - 일반 모드: [오디오 파일 업로드 → 텍스트 답변] (단방향)
  - Live 모드: [실시간 오디오 스트림 → 실시간 오디오 생성] (양방향)

- **WebSocket 기반 양방향 통신**:
  - 일반적인 REST API(요청-응답)가 아니라, WebSocket을 통해 연결을 지속하며 데이터를 주고받는 방식

- **VAD (Voice Activity Detection) & Turn-taking**:
  - 사용자가 말을 끊거나 끼어들 때(Interruption) 즉시 반응하여 생성을 멈추는 로직

- **Low Latency Optimization (초저지연 최적화)**:
  - 모델의 응답 속도를 대화가 가능한 수준(수백 밀리초)으로 맞추기 위한 전용 서빙(Serving) 최적화 기술. 일반 텍스트 생성보다 훨씬 엄격한 타임 리미트가 적용.

02 Use Case

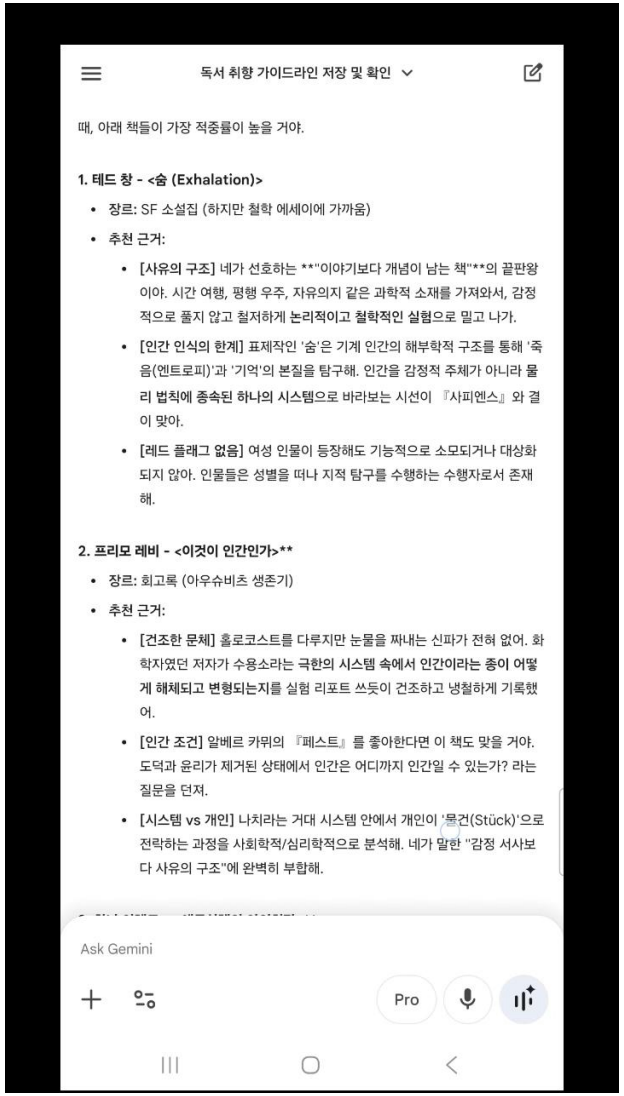


1. 사물 인식

2. 실시간 검색 (기존 LLM Chat 기능)

3. 장면 기억

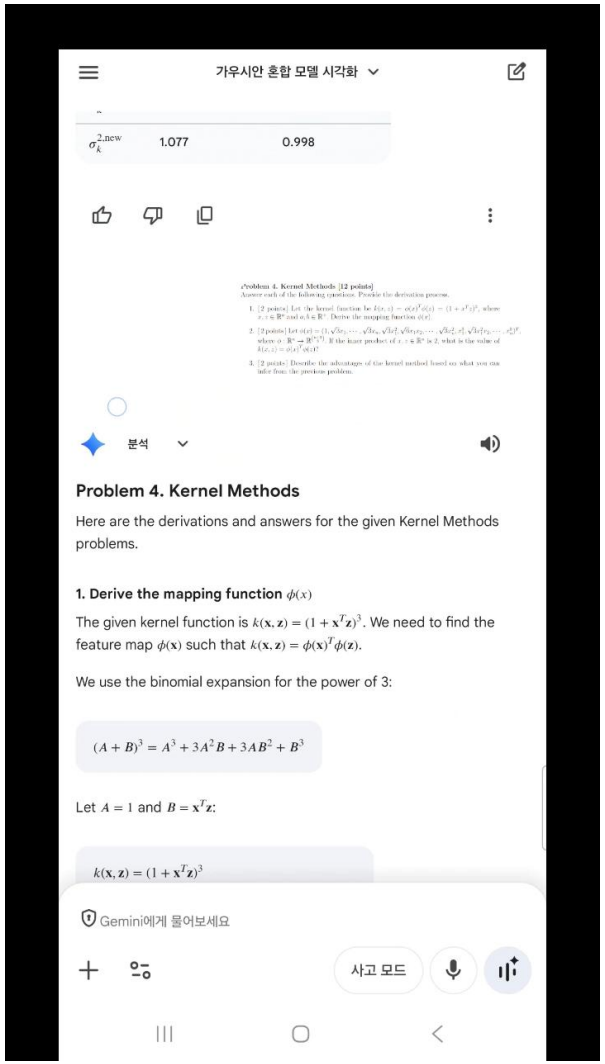
02 Use Case



4. 세션 대화 메모리 연계

5. Google Map 연계

02 Use Case



6. 끼어들기 인식

03 Summary

장점

·실시간성
·끼어들기로 실제 대화처럼 내용 수정 등 가능
·공간적(Spatial) & 시간적(Temporal) 비디오 맥락 이해

한계점

·텍스트 인식과 사물 구분 능력 한계
·비언어적 표현 인식에 한계
·Agentic 역할은 아직 미비함 (기존 Chat UI에서 사용하는 기능만 사용 가능)
·일시적 메모리ㄴ (Global한 메모리 없음)