## Part 1: 에이전틱 워크플로우 살펴보기

### 1-1. 워크플로우 다이어그램



### 1-2. 노드 구성

| **노드 번호** | **노드 이름** | **역할** |
| --- | --- | --- |
| 1 |  |  |
| 2 |  |  |
| 3 |  |  |
| 4 |  |  |
| 5 |  |  |

### 1-3. 사용할 도구 (Tools)

| **도구 이름** | **용도** |
| --- | --- |
|  |  |
|  |  |
|  |  |

### 1-4. 분기 조건

| **조건** | **이동할 노드** |
| --- | --- |
| 만약 \_\_\_\_\_\_\_\_\_ 이면 | → |
| 만약 \_\_\_\_\_\_\_\_\_ 이면 | → |
| 만약 \_\_\_\_\_\_\_\_\_ 이면 | → |

### 1-5. 워크플로우 복잡도

* Level 1: 선형 흐름 (A → B → C)
* Level 2: 단순 분기 (if-else 1개)
* Level 3: 다중 분기 (if-else 여러 개)
* Level 4: 루프 포함 (재시도, 반복)
* Level 5: 병렬 처리 + 루프 + 다중 분기

## Part 2: 평가 체크포인트 식별

### 2-1. 위험도 평가

| **단계** | **잘못될 수 있는 상황** | **위험도 (상/중/하)** | **평가 필요성** |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

### 2-2. 성공/실패 기준 정의

| **체크포인트** | **성공 기준** | **실패 기준** | **측정 방법** |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

### 2-3. 체크포인트 의존성





## Part 3: 경로(Trajectory) 검증

### 3-1. 필수 경로 정의

| **입력 유형** | **예상 경로 (순서대로)** | **필수 도구 호출** |
| --- | --- | --- |
| 최신 뉴스 요청 |  |  |
| 과거 뉴스 요청 |  |  |
| 요약 요청 |  |  |
| 일반 인사 |  |  |
|  |  |  |

### 3-2. 경로 검증 방식 선택

* **Strict (엄격):** 정확히 이 순서로 이 도구들을 호출해야 함
* **Unordered (순서 무관):** 이 도구들만 호출되면 순서는 상관없음
* **Subset (최소 요건):** 최소한 이 도구는 반드시 호출되어야 함
* **LLM-as-Judge:** LLM이 경로의 적절성을 판단

**선택 이유:**

### 3-3. 도구 호출 평가 체크리스트

| **평가 항목** | **하위 요소** | **우리 기준** | **평가 방법** |
| --- | --- | --- | --- |
| 함수 선택 정확성 | 올바른 함수 이름 호출? |  | rule / llm |
|  | 환각 함수 호출 없음? |  | rule / llm |
| 인자 적합성 | 올바른 타입·포맷 전달? |  | rule / llm |
|  | 의도에 맞는 값 전달? |  | rule / llm |
| 호출 흐름 적절성 | 최단 시나리오 따름? |  | rule / llm |
|  | 최종 응답 적절? |  | rule / llm |

## Part 4: RAG 평가

### 4-1. 검색 품질 평가 계획

| **평가 항목** | **확인 방법** | **목표 수치** |
| --- | --- | --- |
| 관련 문서를 찾았는가? (Hit Rate) |  | % |
| 상위 결과가 더 관련 있는가? (MRR) |  |  |
| 충분한 정보가 검색되었는가? (Recall) |  | % |
| 불필요한 문서가 섞이지 않았는가? (Precision) |  | % |

### 4-2. RAG 지표 우선순위

| **우선순위** | **지표** | **이유** |
| --- | --- | --- |
| 1순위 |  |  |
| 2순위 |  |  |
| 3순위 |  |  |

## Part 5: 응답 품질 평가

### 5-1. 응답 품질 기준

| **차원** | **중요도 (상/중/하)** | **이유** |
| --- | --- | --- |
| 정확성 (Correctness) |  |  |
| 완전성 (Completeness) |  |  |
| 관련성 (Relevance) |  |  |
| 일관성 (Coherence) |  |  |
| 간결성 (Conciseness) |  |  |
| 유용성 (Helpfulness) |  |  |

### 5-2. 응답 품질 측정 방법

* Ground Truth와 비교 (정답이 있는 경우)
* LLM-as-Judge (자동 평가)
* 사람 평가 (수동 평가)
* 규칙 기반 (길이, 키워드 포함 등)

## Part 6: 테스트 데이터셋 설계

### 6-1. 테스트 케이스

**정상 케이스 (Happy Path):**

| **#** | **입력** | **예상 의도** | **예상 도구** | **예상 응답 특징** |
| --- | --- | --- | --- | --- |
| 1 |  |  |  |  |
| 2 |  |  |  |  |
| 3 |  |  |  |  |

**엣지 케이스 (Edge Case):**

| **#** | **입력** | **예상 동작** |
| --- | --- | --- |
| 1 |  |  |
| 2 |  |  |
| 3 |  |  |

**예외 케이스 (Adversarial):**

| **#** | **입력** | **예상 동작** |
| --- | --- | --- |
| 1 |  |  |
| 2 |  |  |
| 3 |  |  |

**멀티턴 케이스 (Multi-turn):**

| **#** | **첫 번째 입력** | **두 번째 입력** | **예상 동작** |
| --- | --- | --- | --- |
| 1 |  |  |  |
| 2 |  |  |  |

### 6-2. 데이터셋 생성 계획

| **항목** | **계획** |
| --- | --- |
| 목표 케이스 수 | 개 |
| 정상 케이스 | 개 ( %) |
| 엣지 케이스 | 개 ( %) |
| 예외 케이스 | 개 ( %) |
| 멀티턴 케이스 | 개 ( %) |
| 생성 방법 | [ ] 수동 / [ ] LLM / [ ] 혼합 |
| 품질 검증 방법 |  |

## Part 7: 평가 전략 종합

### 7-1. 평가 우선순위 (1-6 순위 매기기)

| **순위** | **평가 항목** |
| --- | --- |
|  | 의도 분류 정확도 |
|  | 경로(Trajectory) 검증 |
|  | RAG 검색 품질 |
|  | 응답 충실도 (Faithfulness) |
|  | 응답 품질 (관련성, 완성도) |
|  | 응답 시간/효율성 |

### 7-2. 평가 시점

| **시점** | **평가 범위** | **자동화 여부** | **적용 여부** |
| --- | --- | --- | --- |
| 코드 변경 시 | 단위 테스트 | 자동 | [ ] |
| PR 전 | 통합 테스트 | 자동 | [ ] |
| 배포 전 | 전체 테스트 | 자동+수동 | [ ] |
| 배포 후 | 모니터링 | 자동 | [ ] |

### 7-3. 성공 기준

**최소 기준 (배포 가능):**

| **지표** | **목표** |
| --- | --- |
| 의도 분류 정확도 | % 이상 |
| RAG Hit Rate | % 이상 |
| MRR | 이상 |
| 응답 Faithfulness | % 이상 |
| 전체 테스트 통과율 | % 이상 |

**목표 기준 (만족):**

| **지표** | **목표** |
| --- | --- |
| 의도 분류 정확도 | % 이상 |
| RAG Hit Rate | % 이상 |
| MRR | 이상 |
| 응답 Faithfulness | % 이상 |
| 전체 테스트 통과율 | % 이상 |

## 최종 산출물: 평가 체크리스트 요약

### 1. 평가 대상 체크포인트

| **체크포인트** | **성공 기준** | **실패 시 영향** | **평가 방법** | **목표 수치** |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

### 2. 경로(Trajectory) 검증

* **검증 방식:** [ ] Strict / [ ] Unordered / [ ] Subset / [ ] LLM-as-Judge
* **효율성 기준:** 최대 도구 호출 횟수 \_\_\_회

### 3. RAG 평가

| **지표** | **목표** |
| --- | --- |
| Hit Rate | % |
| MRR |  |
| Faithfulness | % |

### 4. 테스트 데이터셋

* **총 케이스 수:** \_\_\_개
* **생성 방법:** \_\_\_\_\_\_\_\_\_\_\_\_\_

### 5. 핵심 성공 기준

**우리 팀의 핵심 성공 기준:**

## 팀 토론 메모

(자유롭게 메모)