



STEELMAN

논쟁에서 이기고 싶으면, 먼저 상대를 이해하세요

"같은 팩트, 다른 결론. 당신은 왜 그 결론에 도달했나요?"

Product Requirements Document v2.0
Gemini 3 Seoul Hackathon 2026 | Track: Gemini for Good
February 28, 2026

1. 우리가 해결하는 진짜 문제

1.1 기존 접근법이 실패하는 이유

지금까지 나온 미디어 리터러시 툴들은 모두 같은 가정에서 출발해요.

“당신이 편향됐다는 걸 알면, 스스로 고치려 할 것이다”

이 가정이 틀렸어요. 심리학 연구에서 반복적으로 증명된 사실이 있어요:

- 편향을 지적받으면 사람들은 반성하는 게 아니라 방어적이 돼요 (심리적 반발, Reactance)
- 이미 비판적 사고에 관심 있는 사람만 미디어 리터러시 툴을 써요
- 정작 편향이 심한 사람일수록 이런 툴을 안 써요

결론: ‘당신의 편향을 고쳐드립니다’는 프레임 자체가 제품을 죽이는 구조예요.

1.2 SteelMan의 다른 출발점

SteelMan은 다른 인간의 본능에서 출발해요.

사람들은 논쟁에서 이기고 싶어합니다.
그 욕구가 있을 때 — 마찰이 없습니다.
상대방 논리를 이해하는 것이 이기기 위한 수단이 됩니다.
그 과정에서 의도치 않게 비판적 사고가 작동합니다.

기존 접근	SteelMan 접근
진입 동기	'공부하러 왔어요'
프레임	'당신이 편향됐어요'
저항	높음 — 방어적 반응
사용 빈도	가끔 (의식적 노력)
결과	외부에서 강요된 관점 변화

1.3 해커톤 문제 정의와의 연결

Gemini for Good 트랙: '모두를 위한 더 나은 사회적, 경제적, 문화적 결과'

임팩트 영역	현재 문제	SteelMan의 기여
사회적	정치적 극단화, 사회 분열	이해하면서 반대하는 문화 형성
문화적	에코챔버, 확증 편향	논쟁 문화의 질적 향상
경제적	허위정보로 인한 의사결정 왜곡	팩트 기반 판단력 강화

2. 제품 컨셉

2.1 SteelMan이란

SteelMan은 '비판적 사고 분석 플랫폼'이에요. 챗봇이 아니에요.

철학 용어 '*Steel-manning*'에서 온 이름이에요. 상대방의 논리를 가장 약한 버전이 아닌, 가장 강한 버전으로 만들어 이해하는 것. 진짜 논쟁은 그때부터 시작돼요.

사용자가 콘텐츠를 입력하면 —
AI 에이전트 3개가 동시에 자율 작동하며 —
3개의 분석 패널을 실시간으로 채워나갑니다.
사용자는 대화가 아닌 '분석 대시보드'를 보고 있습니다.

2.2 핵심 차별점: 챗봇이 아닌 플랫폼

챗봇	SteelMan 플랫폼
메인 화면	대화창
결과물	텍스트 답변
에이전트 동작	보이지 않음
재방문 이유	없음
공유	불가
느낌	질문하는 곳

2.3 3개 분석 레이어 — 순서가 설득의 구조

이 3가지가 순서대로 작동하는 게 핵심이에요. 순서를 바꾸면 제품이 망해요.

Layer 1: Primary Source 검증

'이 주장의 근거가 진짜인가?'

- 에이전트가 인용된 소스의 원본을 자율 탐색
- 인용이 왜곡됐는지, 맥락이 잘렸는지 확인
- '이 기사는 A 연구를 인용했는데, 원본엔 이렇게 써있어요'

→ 감정이 아닌 팩트 레벨에서 시작. 방어심이 낮은 상태.

Layer 2: 다른 관점 탐색

'같은 사실을 왜 다르게 해석하는가?'

- 동일 이슈의 반대 관점 소스 자유훈 수집
- 단순히 '반대 의견'이 아니라 — 같은 팩트를 다른 프레임으로 보는 것
- '같은 데이터인데 왜 결론이 다른가'를 시각화

→ Layer 1을 경험한 후라 '아, 팩트는 같구나'가 전제된 상태에서 봄.

Layer 3: 편향성 분석

'나는 왜 이 관점에 끌렸는가?'

- 1, 2를 다 경험한 후에 등장
- Hans Rosling 10 Instincts 기반 편향 패턴 진단
- '당신이 처음에 동의한 기사는 이런 편향 패턴을 썼어요'

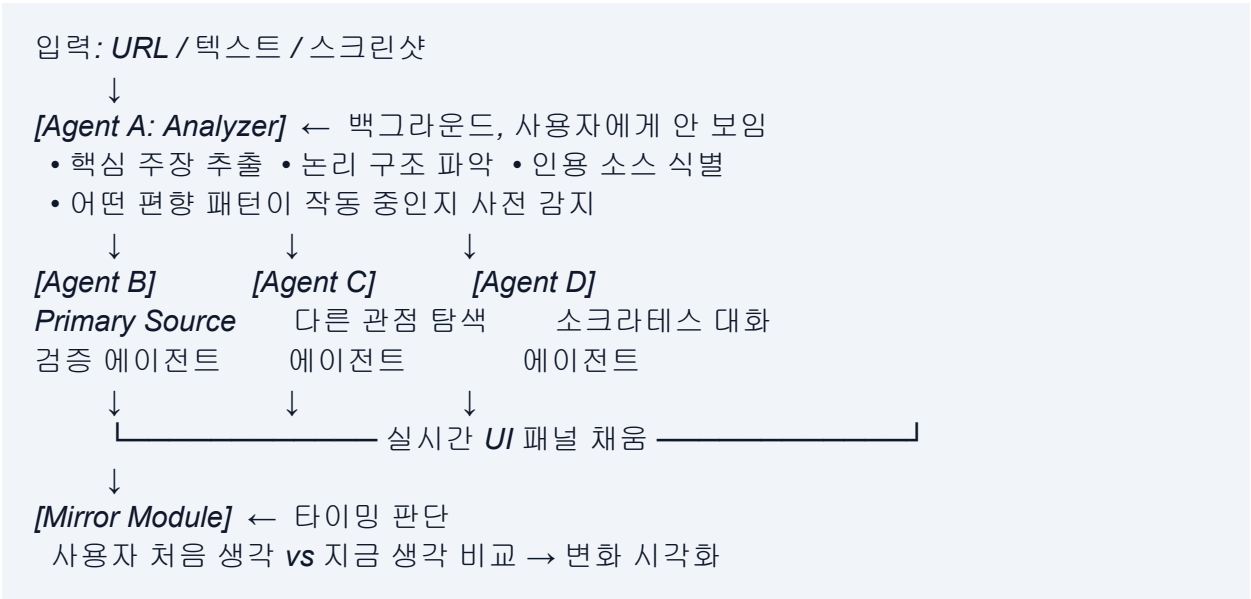
→ 1, 2를 경험한 후라 방어적이지 않음. 스스로 '아, 이래서구나'가 됨.

3번을 처음에 보여주면 → 방어적 반응. 1, 2를 먼저 경험하면 → 자연스러운 자기 인식.
순서 자체가 설득의 구조입니다.

3. 에이전트 워크플로우

3.1 전체 아키텍처

사용자가 콘텐츠를 입력하는 순간, 3개의 에이전트가 독립적으로 + 협력해서 작동해요.



3.2 각 에이전트 상세

Agent A: Analyzer (오케스트레이터)

모든 에이전트의 출발점. 사용자는 볼 수 없어요.

작업	방법	Gemini 3 기능
콘텐츠 파싱	URL → 본문 추출, 스크린샷 → 텍스트+이미지 동시 분석	Multimodal Vision
주장 구조화	핵심 주장 3개, 근거, 인용 소스 추출	Deep Think Reasoning
편향 사전 감지	어떤 본능/편향 패턴이 작동 중인지 분류	Structured Output
에이전트 지시	B, C, D에게 무엇을 탐색할지 지시	Agentic Orchestration

Agent B: Source Verifier

Primary Source 패널을 채우는 에이전트.

작업	방법	출력
원본 소스 탐색	Google Search Grounding으로 인용 원본 자율 탐색	원본 링크 + 실제 내용
왜곡 감지	원본 vs 기사 인용 비교 분석	일치/왜곡/맥락 누락 판정
신뢰도 스코어	소스 발행처, 날짜, 인용 횟수 기반	신뢰도 점수 0-100

Agent C: Perspective Explorer

다른 관점 패널을 채우는 에이전트.

작업	방법	출력
반대 관점 수집	동일 이슈 다른 프레임 소스 3-5개 자율 탐색	관점 카드 리스트
프레임 분석	같은 팩트를 왜 다르게 해석하는지 추출	프레임 차이 시각화
스펙트럼 매핑	좌-우, 공포-희망, 단순-복잡 축으로 배치	관점 스펙트럼 맵

Agent D: Socrates (대화 에이전트)

분석 결과를 바탕으로 사용자와 대화하는 에이전트. B, C와 실시간 협력.

중요: **Agent D**는 분석을 바로 주지 않아요. 먼저 물어봐요.

*Agent D*의 질문 순서:

Q1 (*Layer 1* 전): '이 주장에서 가장 말이 안 된다고 생각하는 부분이 어디예요?'

Q2 (*Layer 1* 후): '원본 데이터 보니까 어떤 생각이 들어요?'

Q3 (*Layer 2* 후): '반대 관점 중 가장 말이 되는 게 뭐예요?'

Q4 (*Layer 3* 후): '지금도 처음이랑 같은 생각이예요?'

3.3 에이전트 간 실시간 협력

Agent B, C가 팩트를 찾으면 바로 대화에 삽입하는 게 아니에요.

- Agent B/C → Agent D에게 먼저 전달
- Agent D가 '지금 이 팩트 넣을 타이밍인가?' 판단
- 대화 흐름에 자연스럽게 삽입: '참고로, 방금 원본 데이터를 찾아봤는데요...'

*Thought Signatures*로 멀티턴 대화 전체에서 추론 일관성 유지.

3.4 사용자 포기 대응 — 3가지 유형

포기 유형	감지 신호	에이전트 대응
A: 귀찮음	답변이 짧아짐, 속도 느려짐	스냅샷 모드 전환 — 핵심 1개만 30초 요약
B: 방어적	'나 설득하려는 거죠?' 반문	역설적 개입 — '맞아요, 가장 강한 버전 알아야 이기니까요'
C: 과부하	'모르겠어요' '너무 복잡해요'	감정 리셋 — '가장 화나는 부분이 어디예요?'

3.5 마음이 안 바뀔 때 — 이것도 성공

실패: 사용자가 상대 의견에 동의하게 만드는 것
성공: 사용자가 상대 논리를 '이해'하게 만드는 것

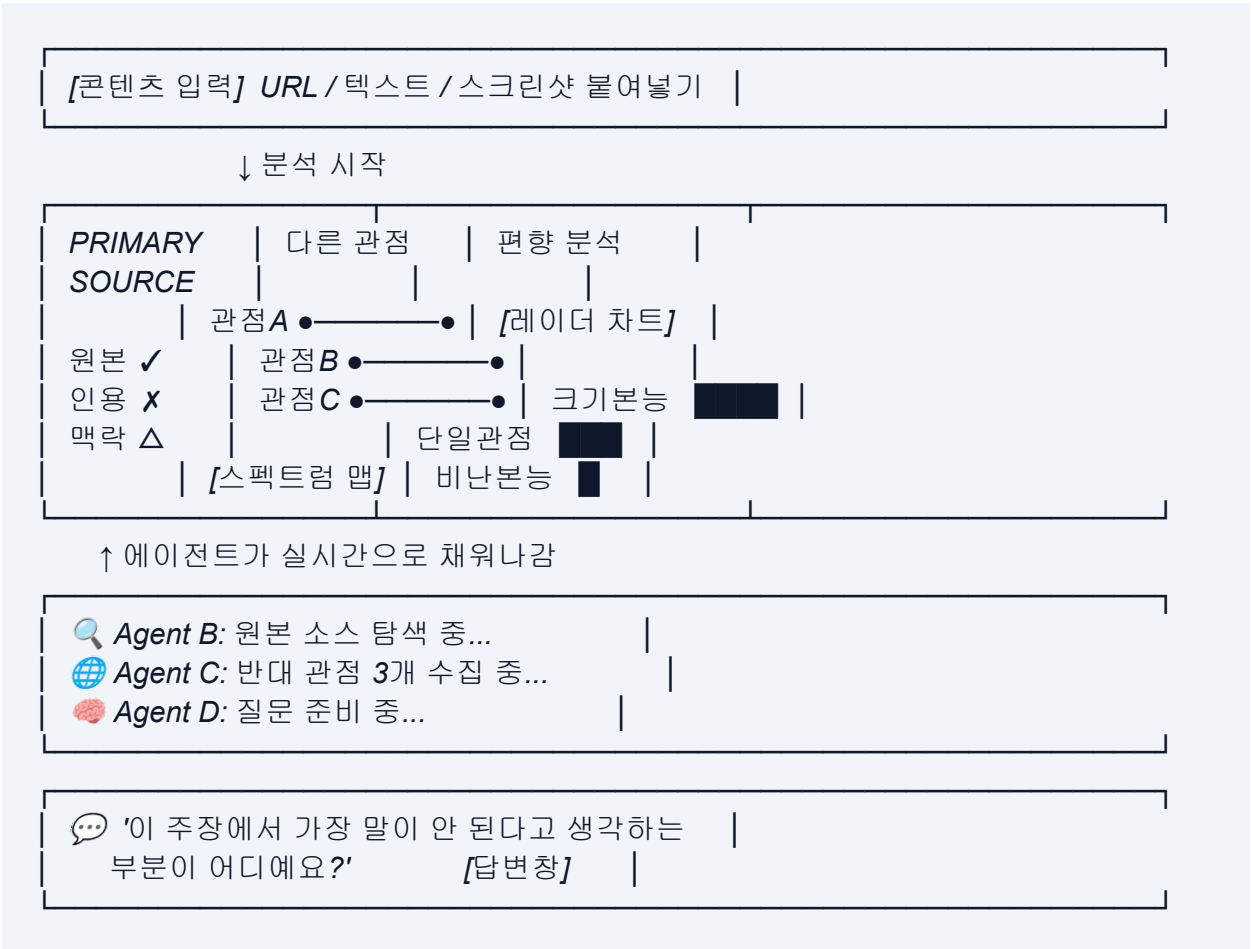
동의 ≠ 이해. 이해하면서 반대할 수 있습니다. 그게 진짜 비판적 사고입니다.

마음이 바뀐 경우	마음이 안 바뀐 경우
아웃풋	사고 확장 지수 시각화
메시지	'오늘 관점이 여기서 여기까지 확장됐어요'
제공물	관점 변화 타임라인

4. UI — 플랫폼 설계

4.1 메인 화면 구조

입력 → 3패널 대시보드가 실시간으로 채워지는 구조예요.



에이전트가 뭘 하는지 실시간으로 보여주는 것 — 이게 챗봇과 시각적으로 가장 다른 부분이에요.

4.2 3개 패널 상세

패널 1: Primary Source

표시 항목	내용
원본 링크	인용된 소스의 실제 원본 URL

일치 여부	✓ 일치 / ✗ 왜곡 / △ 맥락 누락 — 시각적 뱃지
원본 vs 인용 비교	두 텍스트 나란히, 다른 부분 하이라이트
신뢰도 점수	0-100 스코어 + 근거

패널 2: 다른 관점

표시 항목	내용
관점 카드	각 관점 소스 카드 — 제목, 핵심 주장, 프레임
스펙트럼 맵	좌-우 / 공포-희망 / 단순-복잡 축으로 배치
공통점 추출	'모든 관점이 동의하는 팩트' 하이라이트
차이점 추출	'어디서 결론이 갈리는가' 시각화

패널 3: 편향 분석

표시 항목	내용
편향 레이더 차트	Hans Rosling 10 Instincts 기반 6개 축
주요 편향 Top 3	이 콘텐츠에서 가장 강하게 작동한 편향
텍스트 예시	어떤 문장이 이 편향을 만드는지 하이라이트
대안 프레임	'같은 팩트를 편향 없이 쓰면 어떻게 될까'

4.3 최종 아웃풋 카드 — 공유 가능

세션이 끝나면 공유 가능한 분석 카드가 생성돼요. 논쟁 상대한테 바로 보낼 수 있어요.

SteelMan 분석 결과

📌 Primary Source: 인용 왜곡 감지 ✗

🌐 관점 다양성: 낮음 (단일 프레임)

🧠 주요 편향: 비난 본능, 단일관점

💡 상대 논리 SteelMan 버전:

[가장 강한 반박 논리 요약]

🎯 내 주장 강화 포인트:

[취약점 보완된 내 논리 키트]

[공유하기]

5. Gemini 3 활용 전략

5.1 기능 매핑

Gemini 3 기능	SteelMan에서 활용	왜 필수인가
Multimodal Vision	스크린샷/URL → 텍스트+이미지+레이아웃 동시 분석	헤드라인 이미지 선택도 편향 분석 대상
Deep Think	논리 구조 파악, 편향 패턴 진단, SteelMan 생성	단순 요약이 아닌 다층적 추론 필요
Search Grounding	Primary Source 자율 탐색, 반대 관점 실시간 수집	환각 없는 실제 소스 기반 분석
Agentic Workflow	3개 에이전트 병렬 자율 실행 + 협력	이게 챗봇과 구조적으로 다른 핵심
Thought Signatures	멀티턴 대화에서 추론 연속성 유지	Q1~Q4 대화가 하나의 맥락으로 연결
Long Context (1M)	이전 세션 분석 히스토리 전체 컨텍스트	개인화된 편향 패턴 추적 가능
Structured Output	편향 점수, 스펙트럼, 분석 카드 생성	UI 패널에 바로 렌더링 가능한 데이터
Dynamic Thinking	질문 생성은 Flash, 편향 분석은 Pro	비용/속도 최적화

5.2 에이전틱 설계 핵심 — Thought Signatures

Gemini 3의 Thought Signatures가 이 제품에서 특히 중요한 이유:

- Agent D의 Q1~Q4 대화가 진행되면서 '이 사람의 인식이 어디까지 왔는가'를 추론의 연속선 위에서 유지
- 에이전트가 중간에 판단해야 하는 순간들 — 언제 분석 공개할지, 어떤 질문 할지 — 이 모두 이전 추론 위에서 결정
- 없으면 매 턴마다 처음부터 다시 파악해야 함 → 대화의 일관성 붕괴

6. 데모 시나리오 (해커톤 발표용)

6.1 3분 데모 스크립트

심사위원이 볼 것: '이건 챗봇이 아니네. 에이전트들이 실제로 움직이고 있어.'

🕒 0:00-0:15 — 입력

논란이 되고 있는 기사 URL 하나 붙여넣기.

→ 즉시 3개 에이전트 상태 표시 시작. 패널이 채워지기 시작.

🕒 0:15-0:45 — Primary Source 패널

Agent B가 원본 소스를 찾아냄. '이 기사가 인용한 연구, 원본엔 이렇게 써있어요.' 왜곡 감지 **x** 뱃지 등장.

→ Agent D 첫 번째 질문: '이거 보니까 어떤 생각이 들어요?'

🕒 0:45-1:30 — 다른 관점 패널 (3~5개의 주제 고려)

Agent C가 반대 관점 소스 4개 수집. **스펙트럼 맵**에 배치됨. '같은 데이터인데 왜 결론이 다른가' 시각화.

→ Agent D: '반대 관점 중 가장 말이 되는 게 뭐예요?'

🕒 1:30-2:15 — **편향** 분석 패널

레이더 차트 등장. '비난 본능 82%, 단일관점 본능 71%. 이 기사는 이런 패턴을 사용했어요.'

→ Agent D: '지금도 처음이랑 같은 생각이에요?'

🕒 2:15-2:45 — 최종 아웃풋

분석 카드 생성. SteelMan 버전 + 내 논리 강화 키트. [공유하기] 버튼.

🕒 2:45-3:00 — 임팩트 한 마디

'이 사람은 URL 하나 넣었을 뿐인데 — 3개 에이전트가 자율적으로 팩트를 검증하고, 반대 관점을 탐색하고, 편향을 분석했습니다. 논쟁에서 이기려다 비판적 사고를 하게 됩니다.'

7. 해커톤 빌드 플랜

7.1 Tech Stack

레이어	선택
AI Core	Gemini 3 Pro API (Deep Think mode)
Agentic Search	Gemini API + Google Search Grounding
Multimodal	Gemini 3 Multimodal Vision
Frontend	React + Tailwind CSS
Backend	Python FastAPI
시각화	D3.js (레이더 차트, 스펙트럼 맵)
Credit	\$20 Gemini API Credit (해커톤 제공)

7.2 7시간 빌드 계획

시간	빌드 항목	담당
09:00-10:00	Gemini 3 API 연동 + Agent A (Analyzer) 구현	Backend
10:00-11:30	Agent B (Source Verifier) + Search Grounding	Backend
11:30-12:00	Agent C (Perspective Explorer) 기본 버전	Backend
12:00-13:00	점심 + 3패널 UI 기초 (React)	Frontend
13:00-14:30	Agent D (Socrates) + 대화 플로우	Backend+PM
14:30-16:00	UI 완성 — 에이전트 상태 표시, 시각화	Frontend
16:00-16:45	분석 카드 + 공유 기능	Full
16:45-17:00	데모 시나리오 리허설 + 마무리	All

7.3 MVP vs Nice-to-Have

MVP — 반드시 데모	Nice-to-Have
--------------	--------------

3패널 대시보드 UI	분석 히스토리 (My Bias Map)
Agent A: Analyzer	Chrome Extension
Agent B: Primary Source	소셜 공유 기능
Agent C: 다른 관점 (기본)	모바일 최적화
Agent D: 소크라테스 대화 3턴	다국어 지원
에이전트 실시간 상태 표시	사용자 계정 / 히스토리
최종 분석 카드	관점 스펙트럼 인터랙티브 맵

8. Why SteelMan Wins

“논쟁에서 이기고 싶은 인간의 본능을 이용해서, 의도치 않게 비판적 사고를 하게 만드는 AI 분석 플랫폼”

8.1 Judging Criteria 대응

심사 기준	SteelMan의 강점
Impact 25%	사회 분열·극단화 직접 공략. 이기려는 동기 이용 → 저항 없음. 글로벌 확장 가능.
Demo 50%	URL 하나 → 60초 안에 3패널 채워짐. 에이전트가 움직이는 게 눈에 보임. 실시간 시연.
Creativity 15%	미디어 리터러시를 '이기고 싶음'으로 재프레이밍. 플랫폼 형태의 독창적 접근.
Pitch 10%	'같은 팩트, 다른 결론. 당신은 왜 그 결론에 도달했나요?' — 강한 내러티브.

8.2 Anti-Project 체크

✓ AI 정신건강 어드바이저 아님 ✓ 기본 RAG 아님 ✓ 단순 챗봇 아님 ✓ 이미지 분석기 아님 ✓ 교육 챗봇 아님 ✓ 의료 조언 아님 — 완전 적합

8.3 핵심 인사이트 3줄 요약

- 프레임: 기존 도구는 '편향을 고쳐라'고 말해서 실패했어요. SteelMan은 '이기려면 상대를 이해해야 한다'고 말해요.
- 차별점: 챗봇이 아닌 플랫폼. 에이전트 3개가 실시간으로 움직이는 게 보여요. 심사위원이 '이건 뭔가 다르네'를 느껴요.
- 데모 파워: URL 하나 넣으면 60초 안에 Primary Source + 다른 관점 + 편향 분석이 완성돼요. 데모가 제품이에요.

9. Library — 분석이 쌓이면 달라지는 것

9.1 컨셉

일반 서비스의 '히스토리'는 내가 본 것을 다시 보여줘요. **SteelMan**의 라이브러리는 달라요.

일반 추천: 내가 관심 있는 것 → 비슷한 것 더 보여줌 → 에코챔버 강화
SteelMan 추천: 내 편향 패턴 → 그 편향이 다른 이슈에서 어떻게 작동하는지 → 편향 인식 확장

추천의 기준이 '내 관심사'가 아니라 '내 편향 패턴'입니다.
비슷한 기사를 보여주는 게 아니라, 같은 편향이 다른 맥락에서 작동하는 걸 보여줍니다.

9.2 핵심 기능

① 편향 패턴 클러스터링

내가 분석한 콘텐츠들이 쌓이면 패턴이 보여요.

예시:
'당신은 경제 이슈에서 비난 본능이 반복적으로 작동해요'
'환경 이슈에서는 다급함 본능이 강하게 나타납니다'
'외교 이슈에서는 단일관점 본능이 지배적이에요'

→ 이슈별로 내 편향이 어떻게 다르게 작동하는지 시각화

② **SteelMan** 특화 추천

'비슷한 기사'가 아닌 '같은 편향의 다른 발현'을 추천해요.

일반 추천 로직	SteelMan 추천 로직
기반 데이터	내가 클릭한 것, 읽은 시간
추천 기준	관심사와 유사한 콘텐츠
목적	더 많이 보게 함 (체류시간↑)

결과	에코챔버 강화
----	---------

추천 문구 예시:

"비난 본능이 경제 이슈에서 작동했는데 —
이 외교 이슈에서도 같은 패턴이 보여요. 한번 분석해볼래요?"

③ 관점 갭 분석

내가 소비한 콘텐츠의 관점 분포를 보여줘요.

예시:

"당신이 분석한 23개 콘텐츠 중"

진보 프레임: 16개 (70%) | 보수 프레임: 4개 (17%) | 중립: 3개 (13%)

→ '당신이 아직 탐색하지 않은 관점의 콘텐츠' 추천

→ 강요가 아님 — '이런 세계도 있어요' 초대

④ 분석 카드 컬렉션

내가 분석한 모든 콘텐츠의 결과 카드가 저장돼요. 검색, 필터, 태그 가능.

필터 기준	내용
이슈 카테고리	경제 / 정치 / 환경 / 사회 / 국제
주요 편향 유형	비난 본능 / 다급함 / 단일관점 / 크기 본능 등
Primary Source 결과	검증됨 / 왜곡 감지 / 맥락 누락
날짜	이번 주 / 이번 달 / 전체

9.3 해커톤 MVP vs 애드온 범위

기능	해커톤 MVP	애드온
분석 히스토리 저장	세션 내 임시 저장	영구 저장 + 계정
편향 패턴 클러스터링	레이더 차트만	폴 클러스터링 알고리즘
SteelMan 특화 추천	슬라이드 1장으로 소개	실제 추천 엔진 구현
관점 갭 분석	슬라이드 1장으로 소개	실제 분포 계산 + 시각화
분석 카드 컬렉션	세션 종료 후 카드 생성	저장 + 검색 + 필터

해커톤에서는 '이게 쌓이면 이렇게 됩니다'를 슬라이드 한 장 + 목업으로 보여주는 것으로 충분해요. 실제 구현은 애드온 범위.

10. 리스크 & 미티게이션

10.1 데모 리스크 — 가장 중요

해커톤에서 가장 큰 리스크는 기술적 완성도가 아니라 데모 안정성이에요.

리스크	문제	미티게이션
에이전트 응답 속도	3개 에이전트 동시 실행 시 30초+ 지연 가능	Agent B/C 결과 pre-fetch 후 실시간처럼 스트리밍
Search Grounding 불안정	실시간 검색 결과가 관련 없는 소스 반환	데모용 특정 기사 소스를 사전 캐싱
Deep Think 지연	편향 분석에 시간 소요	Flash 모델로 먼저 빠른 결과, Pro로 심층 분석 후 업데이트
API 크레딧 소진	\$20 크레딧이 데모 중 소진될 수 있음	Flash 모델 기본, 핵심 분석만 Pro 사용
UI 렌더링 오류	3패널 동시 업데이트 시 버그	각 패널 독립적으로 업데이트, 에러 시 graceful fallback

핵심 원칙: 화려한 에이전틱 구조보다 안정적으로 작동하는 데모가 훨씬 강해요.
심사위원이 '와 대단하다'보다 '이거 실제로 되네'를 느끼는 게 더 중요해요.

10.2 임팩트 근거 — Q&A 대비

'실제로 사회 분열을 줄이는 효과가 있나요?' 질문에 대한 준비된 답변.

전략: 임팩트 주장을 낮추고 더 구체적으로 만들어요. 과장된 주장보다 검증 가능한 주장이 신뢰를 얻어요.

과장된 주장 (피할 것)	현실적 주장 (사용할 것)
'사회 분열을 해결합니다'	'논쟁의 질을 높입니다'
'편향을 없앱니다'	'편향을 인식하게 합니다'
'극단화를 막습니다'	'상대 논리를 이해하고 반박하는 문화를 만듭니다'
'허위정보를 차단합니다'	'Primary Source를 직접 확인하게 합니다'

참고할 수 있는 근거:

- **Reuters Institute:** 전 세계 28%가 반대 관점 뉴스를 의도적으로 회피 — 이 문제가 실재함을 증명
- **Steel-manning** 기법은 법학, 철학, 토론 교육에서 이미 검증된 방법론
- '논쟁 전 상대 논리 이해' 접근법이 협상, 갈등 해결 연구에서 효과 입증

10.3 'Anti-Project 걸리는 거 아니에요?' 대비

해커톤 금지 항목과 SteelMan의 명확한 차이.

금지 항목	SteelMan과의 차이
AI 정신건강 어드바이저	심리적 지원 X — 논리 분석 도구
기본 RAG 애플리케이션	단순 검색 X — 4개 에이전트 협력 에이전틱 워크플로우
AI for Education 챗봇	교육 목적 X — 실제 논쟁에서 이기기 위한 실용 도구
이미지 분석기	이미지 분석 자체가 목적 X — 멀티모달은 수단
Streamlit 앱	단순 UI X — 3패널 실시간 분석 플랫폼

10.4 7시간 빌드 리스크 관리

시간이 부족할 때 어떤 것부터 포기할지 미리 정해두는 게 중요해요.

우선순위	기능	이유
🔴 절대 포기 불가	3패널 UI + Agent A + 데모 시나리오	이게 없으면 데모 자체가 안 됨
🟡 시간 있으면	Agent B (Source) + Agent C (관점)	있으면 훨씬 강하지만 목업으로 대체 가능
🟢 없어도 됨	Agent D 소크라테스 대화	텍스트 Q&A로 시뮬레이션 가능
⚪ 애드온	라이브러리, 히스토리, 공유	슬라이드로 비전 보여주기

최악의 시나리오 대비: **Agent B/C** 결과를 하드코딩한 '데모 모드'를 미리 만들어두세요. 실제 에이전트가 실패해도 데모는 돌아가야 해요.