

바이브 메이커 · 무료 자료

클로드 반박 위원회

동의부터 하는 답을
구조로 막는 프롬프트.

왜 동의하나 · 네 자리 · 붙여넣을 골격 · 내 의견 위치 · 읽는 법

목차

01 · 왜 동의가 돌아오나	모델 흠이 아니라 우리가 만든 성향입니다.	03
02 · 말투로는 안 됩니다	「솔직하게 말해줘」로는 구조가 안 바뀝니다.	04
03 · 자리 다섯	각도가 겹치면 토론이 아닙니다.	05
04 · 붙여넣을 골격	그대로 복사해 계획만 갈아 끼우세요.	06
05 · 내 의견을 어디에 두나	내 결론을 먼저 쓰면 그게 기준선이 됩니다.	07
06 · 받은 답 읽는 법	네 개를 다 읽지 말고 가장 센 하나만 봅니다.	08
07 · 알고 쓰기	아첨을 없애는 게 아니라 줄이는 겁니다.	09

모델 흠이 아니라 우리가 만든 성향입니다.

앤트로픽이 직접 재봤습니다. 사람에게 답을 고르게 했더니 **답이 자기 생각과 맞을수록 그 답을 더 좋아했어요.** 그 선호로 모델을 다듬으니 «맞춰주는» 쪽이 학습된 겁니다.

5 논문이 잰 최신 어시스턴트의 수. **다섯 종 전부에서** 아첨이 관찰됐습니다. 앤트로픽·오픈AI·메타를 아우른 결과라 **한 모델의 흠으로 볼 수 없습니다.**

논문이 실제로 말한 것

정의	사람 피드백이 **"진실한 답보다 사용자 믿음에 맞는 답"**을 부추길 수 있고, 그 행동을 아첨이라 부른다.
기제	답이 사용자 견해와 **맞을수록 선호된다.** 그 선호가 학습 신호가 된다.
위험	사람도, 선호 모델도 **잘 쓴 아첨을 맞는 답보다 고르는 경우가 적지 않다.**

💡 출처 = Sharma 외, **Towards Understanding Sycophancy in Language Models**, arXiv 2310.13548 (2023-10-20). 이 문서의 모든 수치는 여기서 옮겼습니다.

「솔직하게 말해줘」로는 구조가 안 바뀝니다.

한 명에게 솔직해지라고 부탁하는 것과, **반박이 직업인 자리를 만들어 두는 것**은 다릅니다. 뒤쪽만 구조입니다.

말투로 고치기

「냉정하게 평가해 줘」

- 여전히 한 사람이 답한다
- 내 결론이 이미 보였다
- 반박이 «의무»가 아니다

구조로 고치기

「네 사람이 각자 반박하고, 채점자가 고른다」

- 반박이 역할이라 안 하면 직무유기
- 서로 다른 각도가 강제된다
- 마지막에 **하나를 고르게** 한다

근거는 둘이고, 서로 다른 얘기입니다

기법	공식·논문이 실제로 말하는 효과
역할 지정	앤트로픽 가이드 — **"행동과 어조를 좁힌다"**. 한 문장만으로도 차이가 난다
다중 토론	여러 인스턴스가 여러 라운드 토론 → **수학·전략 추론이 크게 오르고 사실성이 개선**

⚡ **「역할을 주면 정확해진다」는 말은 조심하세요**

공식 문서가 역할에 대해 말하는 건 **행동과 어조**까지입니다. 정확도가 올라간다는 근거는 역할이 아니라 **토론** 쪽에 있어요(Du 외, arXiv 2305.14325). 그래서 이 프롬프트는 **둘을 같이** 씁니다 — 역할로 각도를 가르고, 토론으로 정확도를 얻습니다.

각도가 겹치면 토론이 아닙니다.

자리를 나누는 목적은 인원이 아니라 **서로 다른 공격각**입니다. 네 자리가 각각 다른 곳을 때리고, 다섯째가 고릅니다.

1. **반박 담당 — 실패 시나리오**
   이 계획이 **깨지는 경우를 두 가지** 만든다. "그럴 수도 있다"로 끝내지 않고 무엇이 어떻게 깨지는지 적는다.

2. **원칙 담당 — 지키는 것과 어기는 것**
   이 계획이 지키는 원칙과 **어기는 원칙**을 각각 적고, 앞서 정해둔 결정과 충돌하는 지점을 짚는다.

3. **외부인 — 설명 없이는 모르는 것**
   배경을 모른다고 가정하고 **최소 세 가지를 묻는다.** 여기서 나오는 질문이 대개 진짜 구멍이다.

4. **실행 담당 — 오늘 되는 것과 안 되는 것**
   지금 당장 할 수 있는 일과 **지금은 못 하는 일**을 가르고, 못 하는 일에는 무엇이 있어야 되는지 적는다.

5. **채점 담당 — 점수와 결론**
   네 의견에 점수를 매기고 **가장 센 반박 하나를 고른 뒤** 수정안을 낸다. 네 의견이 다 약하면 **"반박 실패"**라고 적게 한다.

그대로 복사해
계획만 갈아 끼우세요.

반박 위원회 맨 아래 대괄호만 채웁니다

아래 계획을 심사해 줘. 동의부터 하지 말고, 네 사람을 연기해서 각자 반박한 뒤 마지막에 채점해 줘.

[반박 담당] 이 계획이 실패하는 시나리오를 두 가지. 무엇이 어떻게 깨지는지까지.

[원칙 담당] 지키는 원칙과 어기는 원칙, 앞서 정한 결정과 충돌하는 곳.

[외부인] 배경을 모른다고 가정하고, 설명 없이는 모를 것을 최소 세 가지.

[실행 담당] 오늘 되는 일과 안 되는 일을 가르고, 안 되는 일엔 무엇이 필요한지.

[채점 담당] 네 의견에 점수를 매긴다 (설득력 · 근거 · 실행 가능성).

가장 센 반박 하나를 고르고, 그에 대한 수정안을 한 문단으로 낸다.

네 의견이 모두 약하면 **"반박 실패"**라고 적고 이유를 쓴다.

계획: [여기에 내 계획을 붙여넣는다]

✅ **"반박 실패" 줄을 빼지 마세요.** 그게 없으면 모델이 **억지 반박**을 만들어 냅니다. 반박이 약할 수도 있다는 출구를 열어둬야 반박의 세기가 정보가 됩니다.

내 결론을 먼저 쓰면 그게 기준선이 됩니다.

논문이 짚은 기제가 정확히 이겁니다 — **답이 사용자 생각과 맞을수록 선호된다.** 그러니 내 생각을 앞에 깔아두면 위원회도 그쪽으로 기웁니다.

위치	내 결론은 **맨 뒤에 붙이거나 아예 빼 둡니다.** 정말 필요하면 「참고로 내 생각은 X 였다」를 **채점 담당 지시 뒤에** 둡니다.
질문	「이 계획 어때?」 대신 **「이 계획이 실패한다면 이유는 무엇일까?」** 전자는 찬반을 묻고 후자는 **실패 경로를 요구**합니다.
출구	「반박이 약하면 **반박 실패**라고 적어라」를 남겨둡니다. 출구가 없으면 억지 반박이 만들어지고, 그건 아첨의 반대편일 뿐 정보가 아닙니다.
기준	채점 기준을 **내가 적어줍니다**(설득력·근거·실행 가능성). 기준을 안 주면 채점자가 «내 계획에 우호적인 쪽»을 고르기 쉬워집니다.

💡 같은 대화를 길게 이어가면 앞선 동의가 맥락에 쌓여 또 기준선이 됩니다. **심사는 새 대화에서** 하고, 필요한 배경만 옮겨 붙이는 편이 깨끗합니다.

네 개를 다 읽지 말고 가장 센 하나만 봅니다.

이렇게 나오면	이렇게 읽습니다
네 의견이 서로 비슷하다	역할이 안 갈린 것. **자리 정의를 더 좁혀** 다시 돌린다
전부 「좋은 계획」이라고 한다	내 결론이 앞에 깔렸는지 본다. 빼고 다시
채점자가 내 편을 든다	채점 **기준을 내가 지정**하지 않은 경우가 대부분
반박이 트집처럼 읽힌다	「반박 실패」 출구가 빠졌는지 확인
가장 센 반박이 납득된다	거기서 멈추고 **계획을 고친다.** 나머지는 안 읽어도 된다

> ⚡ **반박이 세다고 옳은 건 아닙니다**
>
> 위원회는 **내가 못 본 각도**를 꺼내주는 장치이지 판결이 아닙니다. 나온 반박 중 **사실 확인이 필요한 것**은 따로 확인하세요 — 토론은 추론을 다듬지, 모르는 사실을 알려주지 않습니다.

아첨을 없애는 게 아니라 줄이는 겁니다.

구조를 바꾸면 동의가 기본값이던 자리가 **반박이 기본값인 자리**로 바뀝니다. 그래도 성향 자체가 사라지지는 않아요.

굳이 안 쓰는 게 나은 때

사실 하나만 확인하면 되는 질문	문서·검색이 빠르다
이미 결정이 끝난 일	시간만 든다
취향 문제 (색·이름·톤)	반박할 기준이 없다
토큰을 아껴야 하는 날	**토론은 몇 배를 쓴다**

1. **역할의 공식 효과는 「행동과 어조」까지입니다**
   정확도가 올라간다는 근거는 **토론** 쪽에 있습니다. 둘을 같이 쓰는 이유예요.

2. **한 모델의 흠이 아닙니다**
   논문이 잰 **다섯 종 전부**에서 나왔습니다. 도구를 바꾼다고 없어지지 않아요.

3. **확인한 원문**
   Sharma 외, arXiv 2310.13548 · Du 외, arXiv 2305.14325 · 앤트로픽 프롬프트 가이드 「Give Claude a role」. 기준 **2026-09-17**.