Banya AI 6

Metacognition from Self-Output Distribution: Routing Certain, Vague, and Unknown States

자기 출력 분포에서 창발하는 메타인지: 확실, 모호, 모름의 라우팅

한혁진 (bokkamsun@gmail.com) · ORCID 0009-0007-4132-1707

반야 연구 시리즈 제6편

초록

메타인지(metacognition, 자기가 아는지 모르는지를 아는 능력)를 별도 회로 없이 모델이 자기 출력 분포(output distribution, 다음 낱말별 확률)를 읽어 뽑아낼 수 있다. 본고는 확실(아는 것), 모호(애매한 것), 모름(전혀 모르는 것)의 세 상태가 출력 분포의 통계량만으로 갈리는지를 측정한다. 통계량은 최대 확률(pmax, 가장 그럴듯한 낱말의 확률), 정규화 엔트로피(normalized entropy, 분포가 얼마나 확산했는지의 척도. 0이면 한 곳에 몰림, 1이면 고르게 확산), 봉우리 수(확률이 문턱을 넘는 낱말의 개수)이다. 측정한 결과, 확실은 최대 확률 1.00에 엔트로피 0.00으로 한 곳에 몰려 나머지와 명확히 갈린다. 모호(최대 확률 0.32, 엔트로피 0.35)와 모름(0.35, 0.37)은 확실과는 크게 갈리지만 서로는 통계량이 겹친다. 모름 문항 집합으로 설계한 낱말들(비행기, 컴퓨터, 공룡)이 실은 학습 말뭉치에 다수 등장해 문항 설계가 전혀 모름이 아니었던 것이 한 원인이다. 별도 학습이나 라벨 없이 출력 분포에서 바로 읽히는 것은 확실과 비확실의 갈림이고, 모델은 자기 분포를 모니터링하는 것만으로 자기가 확실한지 아닌지를 안다. 이 신호는 어떤 답을 그대로 내보낼지, 어떤 답을 더 다듬을지, 어떤 답을 모른다고 표시할지의 라우팅(routing, 경로 배정)에 쓸 수 있다. 나아가 이 신호는 트리거(trigger, 발동 조건)로도 작동한다. 실코퍼스(3절의 학습 말뭉치와 같은 동봉 말뭉치)에서 뽑은 가장 모호한 히든(최대 확률 0.11에서 0.13)을 직전 문맥을 전방에 배치해 문맥 동반 재해석하면 1회 만에 확실 문턱 0.55를 넘어 확실 상태로 전환되며, 문맥 없는 단독 재해석 대조는 같은 횟수에서 거의 오르지 않는다.

측정

이 글의 정량 수치는 집필 환경에서 측정한 실측값이다(초등 단계 얼린 모델, 환경은 부록 B). 이 글은 측정된 분포 통계량만을 다루며, 정서나 동기 같은 다른 제어 신호나 형이상학적 해석은 주장하지 않는다(5절).

목차

1. 서론 5
2. 방법 5
3. 결과 5
4. 모호함이 발동시키는 재해석 6
5. 한계와 다음 측정 과제 8
6. 결론 8
7. 후속 논문 예고 8

참고문헌 9
부록 A. 통계량 정의 10
부록 B. 측정 환경 10
부록 C. 재현 안내 10

용어 범례

용어를 아래 표에 정의한다. 같은 개념에는 같은 이름만 사용하며, 일곱 편의 본문과 그림이 모두 이 표를 따른다. 코드 기준은 동봉 코드에서 그 개념을 구현한 식별자이고, 일반 용어는 학계에서 통용되는 가장 가까운 이름이다. 다를 때만 칸을 채웠다.

용어 범례 1. 엔진과 신용

용어	뜻	코드 기준	일반 용어
정확한 전치	각 연산의 수반을 손으로 유도해 닫힌형으로 구현한 역방향 계산. 이 엔진과 신용 사슬의 고유명	`banya_core.py`	수동 역전파
수반	순전파의 짝 연산. 기울기를 정확히 거꾸로 되돌려 흘린다	`bwd` 커널들	adjoint
신용	어떤 지점의 은닉에 대한 손실 기울기. 파라미터가 결과에 진 책임	델타(δ)	그레이디언트
신용 사슬	캐시를 역순으로 되짚으며 수반을 곱해 신용을 입력까지 되돌리는 골격	`backward`	오차역전파
수반 정합	수반 식이 유한차분과 같은 값을 내는지 확인하는 검사	각 편 프로브	gradient check
순환행렬 채널 혼합	채널 혼합을 순환행렬로 두고 rfft 성분곱으로 계산하는 연산		순환 합성곱
상대 거리 혼합	거리에만 의존하는 공유 계수로 자리들을 섞는 혼합	`m_mat_w_lag`	Toeplitz 혼합
시간혼합	다른 자리의 정보가 현재 자리로 섞여 드는 경로의 총칭		토큰 혼합
혼합 헤드	혼합을 병렬로 나누는 머리 수		어텐션 헤드
필터 사다리	넓은 필터 몇 개부터 좁은 필터 여러 개까지 스케일이 깊이를 따라 변하는 콘볼루션 채널 변환	`m_mat_w_filter`	콘볼루션 필터뱅크
창	모델이 한 번에 보는 토큰 자리들. 측정 단위로는 고정 길이 텍스트 조각	`T`	컨텍스트 윈도
프로브	논문 수치를 목표값과 함께 재생하는 검증 스크립트	`probe` 폴더	재현 스크립트
얼린 모델	학습을 멈춰 고정한 발달 단계 모델	`model` 폴더 npz	frozen checkpoint
홀드아웃	학습에 쓰지 않은 평가 텍스트		held-out set

용어 범례 2. 바이토큰과 게이트

용어	뜻	코드 기준	일반 용어
바이토큰	토큰 하나를 데이터면과 연산면의 직교 쌍으로 두는 구조	`USE_BITOKEN`	
데이터면	토큰의 내용을 담는 임베딩 면. 더하기 자리로 들어간다	`m_mat_w_data_axis`	토큰 임베딩
연산면	토큰의 연산 지시를 담는 임베딩 면. 곱하기 자리로만 들어간다	`m_mat_w_operate_axis`	
더하기 자리, 곱하기 자리	데이터면이 실리는 잔차 덧셈 위치와 연산면이 작용하는 게이트 곱셈 위치		잔차 연결, 게이팅
게이트	각 채널의 통과량을 0에서 1로 정하는 밸브. 직전 토큰의 연산면이 주입된다	`m_mat_w_gate`	게이팅
한 칸 이동	직전 토큰의 연산면이 다음 자리의 게이트로 주입되는 한 자리 밀림		
연산이 곧 가중치	연산면 임베딩이 라벨 없이 가중치 역할로 창발하는 현상		
일관 연산자	어떤 데이터에도 일관된 방향으로 작용하는 연산자		
방향 일관성	같은 연산이 여러 데이터를 꺾는 방향들의 평균 코사인. 연산자 창발의 지표	`p_direction_consistency`	
회전 연산자 게이트	연산면을 오일러 회전각으로 읽는 게이트. 본문 축약은 회전 게이트	`paper7_rotation.py`	
기본 연산	게이트가 한 홉 지시로 곧바로 표현할 수 있는 연산		프리미티브
깊이 배제 판별	깊이를 1블록으로 제한해 게이트만 변수로 남긴 의도된 편파 측정	제7편 프로브	
자기 되먹임	자기 출력을 입력으로 되먹이며 끝 상태를 채점하는 채점 방식		자기회귀 롤아웃

용어 범례 3. 되새김과 발달

용어	뜻	코드 기준	일반 용어
되새김	기울기 없이 임베딩을 이웃끼리 당겨 뭉치게 하는 자기조직화 학습	제2편 프로브	자기조직화
동종군	씨앗에서 코사인 최근접 이웃을 모은 뜻 가까운 개념 집합		클러스터
무게중심 당김, 감쇠	평균 쪽으로 당기고 원본 쪽으로 되돌리는 되새김의 두 힘. 균형에서 평형이 창발한다	`ALPHA`, 감쇠율	
월드먼저	언어 이전에 감각, 공간, 논리 축을 먼저 발달시키는 커리큘럼 순서	`banya_world_data`	커리큘럼 학습
발달 단계	스텝 수로 정의하고 사람 성장 이름을 붙인 체크포인트 단계		
축	임베딩 안에 방향으로 형성되는 성질의 눈금. 감각, 순서, 범주		표현 축

용어 범례 4. 어휘와 메타인지

용어	뜻	코드 기준	일반 용어
음절 원자	어휘 바닥층의 음절 단위 토큰	`AtomTokenizer`	문자 단위 토큰
묶음	음절 원자를 묶어 사전에 올린 개념 토큰. 어휘 단위로만 사용한다		서브워드
캐시 사전 토큰화	원자 아래층에 묶음 사전 위층을 얹은 2층 어휘 체계	제5편 프로브	서브워드 토큰화와 대비
확실, 모호, 모름	자기 출력 분포로 갈리는 세 앎 상태	pmax, 엔트로피	불확실성 추정
라우팅	상태에 따라 답의 처리 경로를 배정하는 것		routing
재해석	모호 히든을 문맥과 함께 재순전파해 확실로 끌어올리는 절차	제6편 프로브	
유사도 정리	데이터면 유사도로 개념이 정돈되는 축		
추론 순회	연산면을 따라 추론을 이어 가는 축		
자기 발화	유사도 정리와 추론 순회를 엮어 스스로 말을 잇는 목표 상태		
반야프레임	시리즈의 배경을 이루는 공리 15개 체계. 본문 주장의 근거는 아님		

기호 범례

이 편의 수식과 본문이 함께 사용하는 기호를 정의한다. 한 식 안에서만 쓰는 국소 기호는 각 식 아래의 기호 줄에 적었다.

기호	한글 명칭	정의 및 의미	자료형
p_i	낱말 확률	다음 낱말 분포에서 낱말 i가 나올 확률	스칼라
$p_{\max}$	최대 확률	분포에서 가장 큰 확률, 곧 가장 그럴듯한 낱말의 확률. 측정값은 확실 1.00, 직접 연상 0.58, 모호 0.32, 모름 0.35	스칼라
$\tilde{H}$	정규화 엔트로피	전체 엔트로피를 최대값 $\log V$로 나눠 0에서 1로 맞춘 확산 척도. 0이면 한 곳에 몰림, 1이면 고르게 확산	스칼라
봉우리 수	봉우리 수	확률이 문턱 θ를 넘는 낱말의 개수. 봉우리가 여럿이면 후보가 많은 모호 상태다	정수
V	어휘 크기	모델 어휘에 든 낱말의 총 개수. 엔트로피 정규화의 분모 $\log V$에 사용된다	정수
θ	봉우리 문턱	봉우리 수를 셀 때 낱말 확률이 넘어야 하는 문턱. 설정값 0.05	스칼라
i	낱말 색인	어휘의 낱말을 가리키는 번호. 합과 최댓값 연산의 구간 변수	정수
k	깊이	재해석에서 모호 히든 앞에 두는 문맥 토큰 개수. 1에서 16까지 전부 시도해 최대 확률이 가장 높아지는 깊이의 히든을 채택한다	정수
$\max_i$	최댓값 연산	모든 낱말 i 가운데 가장 큰 값을 고르는 연산. 최대 확률의 정의에 사용된다	연산자
$\sum_i$	합 연산	모든 낱말 i에 걸쳐 더하는 연산. 엔트로피 계산에 사용된다	연산자
log	로그	로그 함수. 엔트로피의 각 항 $\log p_i$와 정규화 분모 $\log V$에 사용된다	함수

1. 서론

언어모델이 자기가 아는지 모르는지를 스스로 아는 능력은 안전한 생성과 직결된다. 아는 것은 그대로 내보내고, 모르는 것은 모른다고 표시하거나 더 다듬어야 한다. 본고는 이 메타인지를 별도의 학습된 판별기 없이, 모델이 이미 내는 출력 분포의 통계량만으로 뽑아낼 수 있는지를 측정한다. 자기 분포를 읽는 것이 자기 상태를 아는 저비용 방법이라는 가설이다. 학습에는 제1편 [2]의 엔진을 쓴다. 이론적 배경은 별도 공리 문헌 [1]에 있으나, 본고는 그와 독립적으로 측정에서만 결론을 낸다.

2. 방법

세 상태의 문항 집합을 준비한다. 확실(모델이 잘 아는 것), 모호(후보가 여럿인 것), 모름(전혀 모르는 것)이다. 세 문항 집합은 모두 사용자와 반야의 대화 형식으로 묻는다. 여기에 중간 대조군으로 직접 연상 문항 집합을 둔다. 대화 틀 없이 학습된 연상을 문장 이어쓰기로 입력해 다음 낱말 분포를 읽는 문항 집합이며, 문항은 "봄 여름 가을", "눈으로" 같은 부분 문장 네 개다. 같은 지식이라도 대화 틀에 담긴 확실보다 최대 확률이 낮게 나오는지를 보는 대조다. 각 문항의 다음 낱말 분포에서 세 통계량을 측정한다.

식 2-1. 분포 통계량

$$p_{\max} = \max_i p_i, \qquad \tilde{H} = \frac{-\sum_i p_i \log p_i}{\log V}, \qquad \text{봉우리 수} = \left|\{\, i : p_i > \theta \,\}\right|$$

수식

기호: p_i는 낱말 i의 확률, $p_{\max}$는 그중 가장 큰 값(최대 확률), $\tilde{H}$는 정규화 엔트로피(전체 엔트로피를 최대값 $\log V$로 나눠 0에서 1로 맞춘 것), V는 어휘 크기, θ는 봉우리 문턱(0.05), 봉우리 수는 확률이 그 문턱을 넘는 낱말의 개수이다.
작동 방식: 최대 확률이 크고 엔트로피가 작으면 분포가 한 곳에 몰린 것(확실). 최대 확률이 작고 엔트로피가 크면 넓게 확산한 것(모름). 봉우리가 여럿이면 후보가 많은 것(모호)이다.
실제 동작: 세 통계량은 학습이나 라벨 없이 이미 나온 분포에서 바로 계산된다. 별도의 판별기를 학습할 필요가 없다.

3. 결과

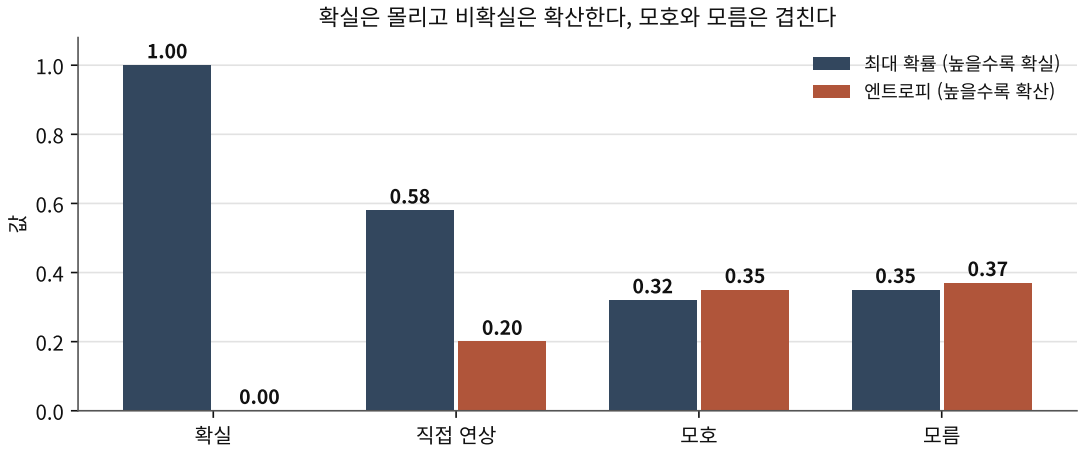
측정

세 상태 문항 집합의 분포 통계량을 측정한 결과, 확실은 나머지와 명확히 갈리고 모호와 모름은 서로 겹친다.

상태	최대 확률 $p_{\max}$	정규화 엔트로피 $\tilde{H}$
확실 (아는 것)	1.00	0.00
직접 연상 (관련 개념)	0.58	0.20
모호 (후보 여럿)	0.32	0.35
모름 (전혀 모름)	0.35	0.37

확실은 최대 확률 1.00과 엔트로피 0.00으로 한 곳에 몰려 명확히 갈린다. 엔트로피는 확실에서 모름으로 갈수록 0.00, 0.20, 0.35, 0.37로 증가하지만 모호와 모름 사이의 차이는 작고, 최대 확률은 모호 0.32와 모름 0.35로 역전되어 두 상태가 갈리지 않는다. 상태 판정 문턱은 최대 확률 0.55 이상이면 확실, 0.20 이하이면 모름으로 정한다. 원인을 확인해 보면 모름 문항 집합의 낱말들이 학습 말뭉치에 이미 다수 등장한다(비행기 14,579회, 컴퓨터 1,023회, 공룡 121회). 즉 전혀 모름으로 설계한 문항이 이 체크포인트에는 이미 본 것이었고, 실제로 모름 4문항 중 3개가 모호로 라우팅된다. 별도 학습 없이 분포만으로 갈리는 것은 확실과 비확실의 경계이며, 이 경계가 4절 재해석 트리거의 발동 조건이다. 봉우리 수는 정의만 두고 이 표에서는 측정하지 않았으며, 그 측정은 5절의 다음 과제로 미룬다.

그림 3-1. 확실은 몰리고 비확실은 확산한다



확실은 최대 확률(남색 막대) 1.00과 엔트로피(주홍 막대) 0.00으로 한 곳에 몰려 명확히 갈린다. 엔트로피는 확실에서 모름으로 갈수록 증가하지만 모호(0.35)와 모름(0.37)의 차이는 작고, 최대 확률은 모호 0.32와 모름 0.35로 역전되어 두 상태가 겹친다. 별도 학습이나 라벨 없이 갈리는 것은 확실과 비확실의 경계다.

재현

3절의 분포 통계 수치는 부록 C의 준비를 마친 뒤 다음 한 줄로 재생된다.

```
cd probe && python3 paper6_metacog_probe.py
```

출력에서 최대 확률 확실 1.00, 직접 연상 0.58, 모호 0.32, 모름 0.35 와 엔트로피 0.00, 0.20, 0.35, 0.37 이 본문 수치와 대응한다.

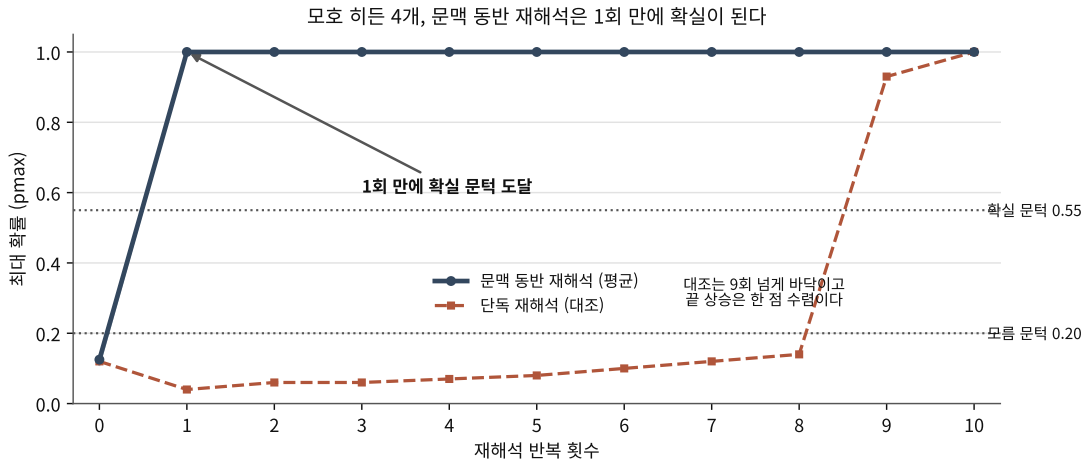
4. 모호함이 발동시키는 재해석

3절의 상태 갈림은 읽기만 하는 신호가 아니라 행동의 트리거(trigger, 발동 조건)가 된다. 확실이면 그대로 내보내고, 확실이 아니면 재해석을 발동시켜 최대 확률(pmax)을 올린 뒤 다시 판정하는 회로이다. 이 절은 그 회로의 핵심 고리인 모호에서 확실로의 전환을 측정한다.

측정 방법은 다음과 같다. 실코퍼스(동봉 자작 초등 대화 말뭉치로, 3절의 학습 말뭉치와 같다)를 순전파해 다음 낱말 예측의 최대 확률이 가장 낮은 자리 네 곳을 고르고, 그 자리의 히든(hidden, 층을 전부 통과한 내부 표현 벡터)을 모호 히든으로 삼는다. 재해석 한 회는 그 자리 직전의 문맥 토큰 16개를 전방에 배치하고, 앞 k개 토큰 뒤에 모호 히든을 배치한 시도를 깊이 k가 1에서 16까지 전부 한 배치로 순전파한 뒤, 최대 확률이 가장 높아지는 깊이의 히든을 채택하는 것이다. 이를 10회 반복하며 매회 최대 확률을 기록한다. 대조군은 문맥 없이 모호 히든 단독으로 재해석을 반복한 것이다. 상태 판정 문턱은 3절과 같다(확실 0.55, 모름 0.20).

> **측정**
>
> 실코퍼스에서 뽑은 모호 히든 네 개의 시작 최대 확률은 0.11에서 0.13으로 전부 모름 대역(모름 문턱 0.20 이하의 구간)이다. 문맥 동반 재해석은 네 개 전부 1회 만에 확실 문턱 0.55를 넘었다. 재해석 3회까지의 최대 확률 상승은 문맥 동반 쪽이 +0.88, 단독 대조가 -0.06이다. 대조도 9회 이상 반복하면 문턱에 도달하나, 그 상승은 입력과 무관하게 한 점으로 수렴하는 고정점 성격이어서 문맥 동반 재해석과 다르다.

그림 4-1. 재해석 반복에 따른 최대 확률 상승



가는 선은 모호 히든 네 개의 개별 궤적, 굵은 선은 평균이다. 문맥 동반 재해석은 1회 만에 확실 문턱 0.55를 넘어 1.0 부근에서 유지된다. 단독 재해석 대조(점선)는 9회 넘게 바닥에 머무르며, 끝의 급상승은 한 점 수렴이다.

이 결과로 라우팅의 세 경로 중 다듬기 경로가 실제로 작동함이 확인된다. 모호나 모름으로 판정된 히든이라도 자기 문맥 위에서 재해석하면 한 회 만에 확실 대역(확실 문턱 0.55 이상의 구간)으로 넘어온다. 본고가 측정한 것은 최대 확률(pmax)의 상승까지이며, 재해석이 낸 답의 정답 여부는 측정하지 않았다. 또한 이 회로는 프로브(probe, 측정 코드)가 바깥에서 반복을 돌린 것이고 문턱 판정으로 자동 발동하는 배선이 모델 안에 들어 있는 것은 아니다.

> **재현**
>
> 이 절의 재해석 트리거는 부록 C의 준비를 마친 뒤 다음 한 줄로 재생된다.
>
> ```
> cd probe && python3 paper6_vague_trigger_probe.py
> ```

출력에서 모호 히든 4개 전부 1회 만에 확실 문턱 0.55 도달, 재해석 3회까지 상승 문맥 동반 +0.88 대 단독 -0.06 이 본문 수치와 대응한다.

5. 한계와 다음 측정 과제

기준선은 1인 개발, 자작 정확한 전치 역전파 엔진, 소비자 그래픽 처리장치 한 대라는 레퍼런스 클래스이다. 아래는 결함이 아니라 다음 측정 과제이다.

- 보정과 하류 라우팅 : 이 상태 갈림이 실제 과제에서 잘못된 답을 걸러 내는 데 얼마나 도움이 되는지의 병행 측정(보정, calibration 지표 포함)이 다음 과제이다.
- 정서와 동기 신호 : 자기 좌표와 기하에서 정서나 학습 동기 같은 다른 제어 신호를 뽑는 부분은, 본고가 측정한 얼린 모델에는 그 회로가 아직 없어 측정되지 않았다. 이는 그 회로를 갖춘 뒤 단계의 몫이며 본고는 주장하지 않는다. 특히 정서를 생리 지표에 대응시키는 해석은 측정 근거가 없으므로 다루지 않는다.
- 모호와 모름의 세분 : 이 문항 집합에서는 두 상태의 최대 확률과 엔트로피가 겹쳐 갈리지 않았고, 모름 문항의 낱말들이 학습 말뭉치에 이미 등장한 설계 결함이 확인됐다. 학습에 없는 낱말로 문항 집합을 다시 만들고 봉우리 수까지 함께 사용하는 세분 라우팅의 재측정이 다음 과제이다.
- 표준 불확실성 대비 : 온도 스케일링이나 앙상블 같은 표준 불확실성 추정과의 직접 비교가 다음 과제이다.

6. 결론

본고는 모델이 자기 출력 분포의 통계량(최대 확률, 엔트로피, 봉우리 수)만으로 확실과 비확실을 별도 학습 없이 갈라냄을 측정으로 보였다. 확실은 최대 확률 1.00과 엔트로피 0.00으로 몰리고 비확실은 확산한다. 모호와 모름의 세분은 이 문항 집합에서 통계량이 겹쳐 갈리지 않았는데, 모름으로 설계한 낱말들이 실은 학습에 다수 등장한 것이 한 원인이라 문항 집합 확장과 함께 다음 과제로 남긴다. 이 신호는 답을 그대로 낼지, 다듬을지, 모른다고 표시할지의 라우팅에 쓸 수 있다. 나아가 이 신호를 트리거로 쓴 재해석이 실제로 작동함을 보였다. 실코퍼스에서 뽑은 모호 히든을 직전 문맥을 전방에 배치해 문맥 동반 재해석하면 1회 만에 확실 문턱을 넘는다. 정서와 동기 같은 다른 제어 신호는 해당 회로를 갖춘 뒤 단계의 측정 과제로 둔다. 닫힌 기계는 스스로 시작하지 못한다. 본 시리즈가 모름 신호에 이토록 공을 들이는 이유는 그 한 문장에 있다.

7. 후속 논문 예고

후속 논문 예고

- 마지막 편(미완) — 유사도 정리와 추론 순회의 결합 : 데이터면 유사도 정리(제2편)와 연산면 추론 순회를 엮어, 입력 없이 스스로 도는 자기 발화를 목표로 한다. 본편의 모름 신호가 그 발화의 연료가 될 수 있다. 이 자기 발화는 반야프레임의 열다섯째 공리를 기계로 구현해 보려는 시도이기도 하다. 최종 조립이 끝나지 않아 아직 다루지 않으며, 측정이 확립되면

별도로 다룬다.

참고문헌

[1] 한혁진 (2026). Banya Framework: An Axiom-Based Science Mining Engine for Deriving Fundamental Physical Constants (반야프레임 공리 15개). Zenodo. https://doi.org/10.5281/zenodo.20301304 . (본 시리즈의 배경 자연관, 선행 공리 문헌)

[2] 한혁진 (2026). 자동미분 없이 수동으로 유도한 정확한 전치로 학습하는 언어모델 엔진. 반야 연구 시리즈 제1편. Zenodo. https://doi.org/10.5281/zenodo.21385447

부록 A. 통계량 정의

표 A-1. 분포 통계량

통계량	뜻	확실할 때
최대 확률 $p_{\max}$	가장 그럴듯한 낱말의 확률	높음
정규화 엔트로피 $\tilde{H}$	분포가 확산한 정도 (0에서 1)	낮음
봉우리 수	확률 0.05 초과 낱말의 개수	적음

부록 B. 측정 환경

표 B-1. 측정에 사용한 컴퓨터 사양

구분	사양
그래픽 처리장치(GPU)	NVIDIA GeForce RTX 3090 Ti (24 GB)
운영체제	Ubuntu 24.04.4 LTS
소프트웨어	Python 3.12.3, numpy 2.5.1, scipy 1.18.0, cupy 14.1.1, CUDA 13.3.1
측정 대상	초등 단계 얼린 모델의 확실, 모호, 모름 문항 집합 출력 분포

3절과 4절의 실측 수치는 위 환경에서 측정했다.

부록 C. 재현 안내

본고의 실측 수치는 공개 재현 패키지로 재생된다. 아래 다섯 단계를 순서대로 하면 된다.

1. 요구 환경. NVIDIA 그래픽 처리장치(GPU)와 CUDA, Python 3.12 가 필요하다. 파이썬 패키지는 압축을 푼 폴더에서 다음으로 설치한다. cupy 는 설치된 CUDA 버전에 맞는 배포판을 사용한다(예: CUDA 13 은 cupy-cuda13x).

```
pip install -r requirements.txt
```

2. 코드 받기. 압축 묶음(zip)을 내려받아 푼다. 저장소로 받으려면 github.com/ubmscoin/banya/banya_ai_paper_code 폴더다.

```
wget https://ubmscoin.github.io/banya/banya_ai_paper_code/
banya_ai_paper_code.zip
unzip banya_ai_paper_code.zip && cd banya_ai_paper_code
```

3. 모델 받기. 얼린 모델 3개(합계 471MB)는 제노도 doi.org/10.5281/zenodo.21383724 에 있다. 다음 한 줄이 내려받기와 무결성 검증(sha256, 파일 지문 대조)을 함께 한다. 손으로 받으려면 위 기록에서 세 파일을 받아 model 폴더에 넣고 `sha256sum -c checksums.sha256` 으로 검증한다.

```
cd model && bash download.sh && cd ..
```

4. 말뭉치 만들기. 말뭉치는 원문 배포 없이 생성기가 만든다. 다음 한 줄이 banya_world_data 폴더에 말뭉치 23종을 전부 생성한다.

```
cd corpus_build && bash build_all.sh && cd ..
```

5. 실측 재생. probe 폴더의 프로브가 각 절의 수치를 재생한다. 본고의 수치는 다음 프로브가 담당한다. 출력에 목표값이 함께 인쇄되므로 본문과 바로 대조된다.

프로브	재생하는 수치	확인할 출력
paper6_metacog_probe.py	3절 분포 통계	최대 확률 1.00, 0.58, 0.32, 0.35 와 엔트로피 0.00, 0.20, 0.35, 0.37
paper6_vague_trigger_probe.py	4절 재해석 트리거	4개 전부 1회 도달, 3회 상승 +0.88 대 -0.06

패키지의 폴더 구성과 파일별 설명은 목차 페이지에 있다.

반야 연구 시리즈 제6편(메타인지, 분포 라우팅). 모든 수치는 이 글에서 부록 B 환경으로 직접 측정한 값이다. 배경 자연관은 선행 공리 문헌 [1]로 인용하되 이 글 주장의 근거가 아니다. 제1편을 인용한다. 정서와 동기는 해당 회로를 갖춘 뒤 단계의 측정 과제로 두었다. 유사도 정리와 추론 순회를 엮는 자기 발화는 최종 조립 뒤 마지막 편에서 다룬다.