Finetuned Language Models Are Zero-Shot Learners

FLAN Finetuned Language Net

Wei, Bosma, Zhao, Guu, Yu, Lester, Du, Dai, Le · Google Research · ICLR 2022

Instruction tuning — finetuning LLMs on tasks phrased as natural language instructions — substantially improves zero-shot performance on unseen tasks.

Problem: LLM의 Zero-Shot 성능 부족

Few-Shot vs Zero-Shot

Few-Shot

예시를 몇 개 제공한 후 질문

→ GPT-3가 강한 영역

Zero-Shot

예시 없이 instruction만으로 질문

→ GPT-3가 현저히 약한 영역

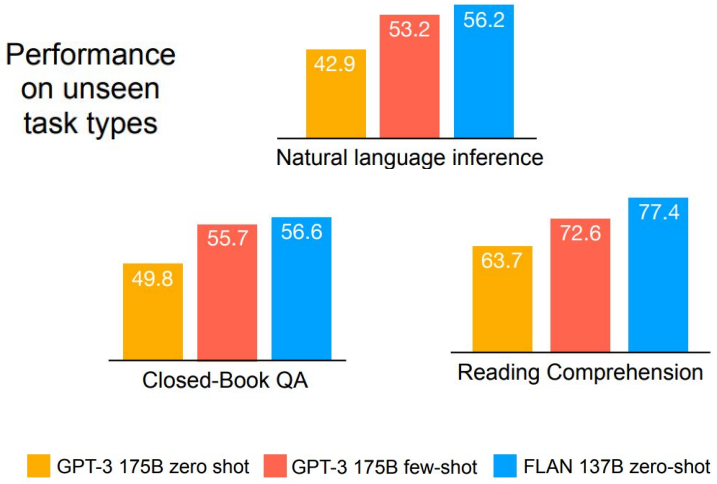
왜 약한가?

Pretraining 데이터 형식과

Task 입력 형식의 mismatch.

→ 모델이 '무엇을 해야 하는지' 파악 못함

GPT-3 (175B) 성능 격차



핵심 질문: 예시(exemplar) 없이, instruction만으로 LLM이 새로운 task를 잘 수행할 수 없을까?

Idea: Instruction Tuning

(A) Pretrain-Finetune

BERT, T5

Task마다 별도 finetuning
→ 새 task = 새 모델 필요
→ 확장성 낮음

>

+

(B) Prompting

GPT-3

Few-shot 예시로 성능 향상
→ Zero-shot에서는 약함
→ 예시 설계에 의존

>

=

(C) Instruction Tuning

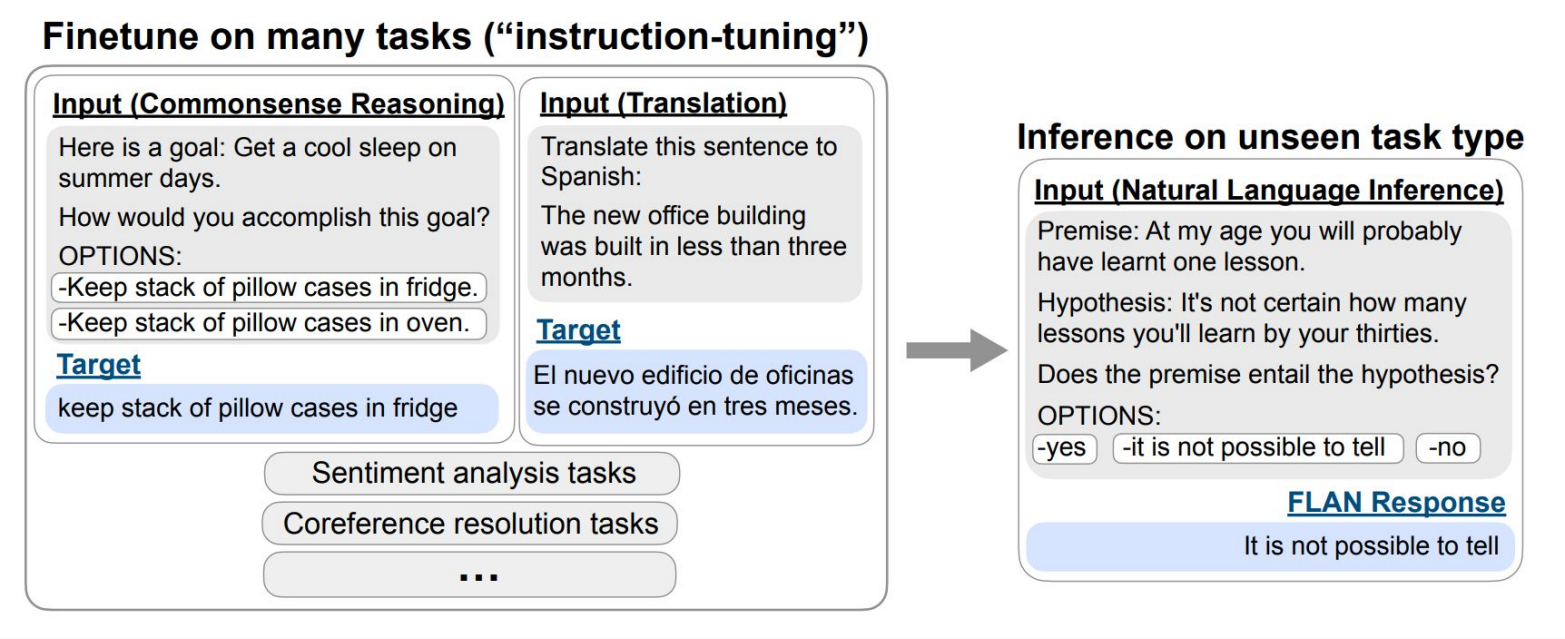
FLAN ★

다양한 task를 instruction 형태로 동시에 학습 → 새 task도 instruction만으로 수행 가능

NLP task를 자연어 instruction으로 표현하여 학습시키면,
모델은 instruction을 이해하는 능력 자체를 습득한다

→ 학습에서 본 적 없는 task(unseen task)도 instruction만으로 수행 가능

Method : Overview



Method ① 데이터 구성

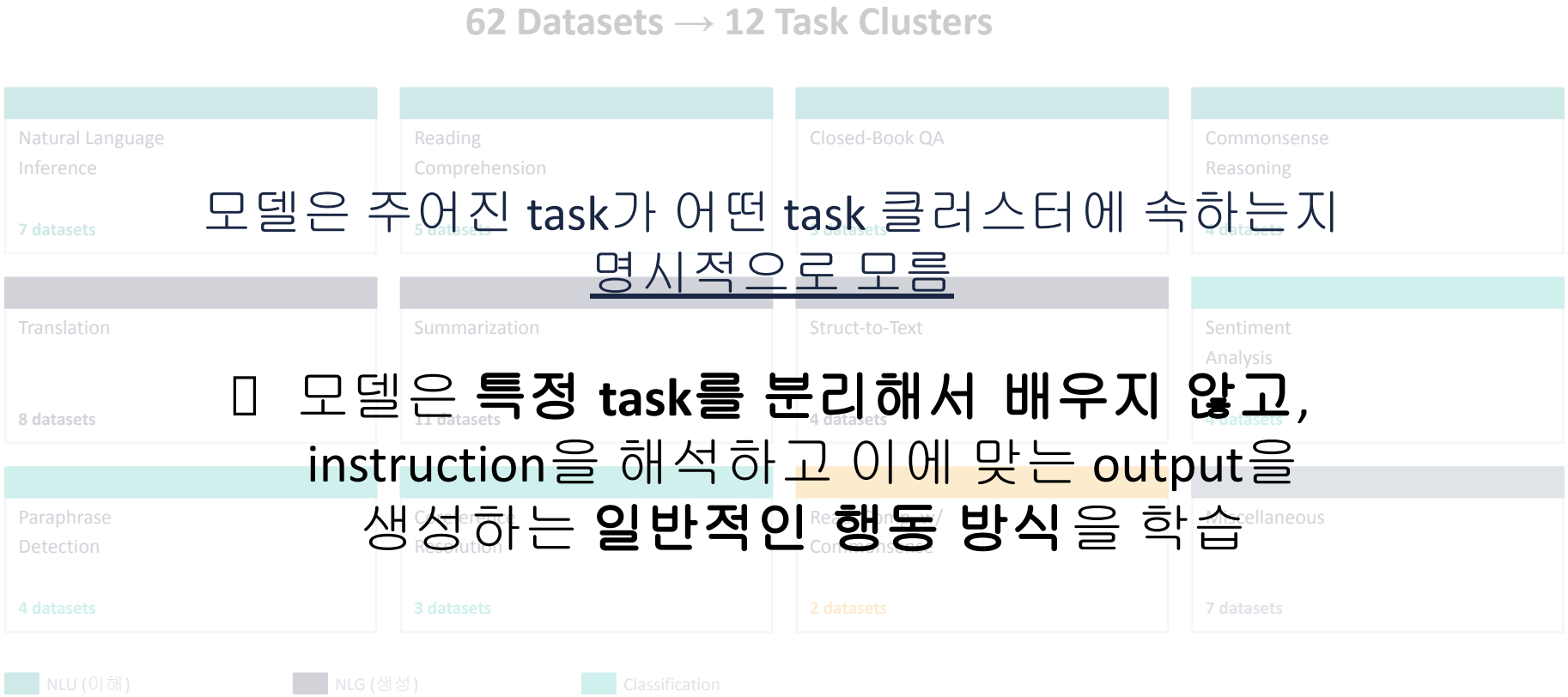
62 Datasets → 12 Task Clusters

- 각 Dataset당 최대 3만개 예시 사용,
- 총 90만개 가량의 Q&A pair 사용

Natural Language Inference **7 datasets**	Reading Comprehension **5 datasets**	Closed-Book QA **3 datasets**	Commonsense Reasoning **4 datasets**
Translation **8 datasets**	Summarization **11 datasets**	Struct-to-Text **4 datasets**	Sentiment Analysis **4 datasets**
Paraphrase Detection **4 datasets**	Coreference Resolution **3 datasets**	Read. Comp. w/ Commonsense **2 datasets**	Miscellaneous **7 datasets**

NLU (이해) NLG (생성) Classification

Method ① 데이터 구성



모델은 주어진 task가 어떤 task 클러스터에 속하는지 명시적으로 모름

 모델은 **특정 task를 분리해서 배우지 않고**, instruction을 해석하고 이에 맞는 output을 생성하는 **일반적인 행동 방식**을 학습

Method ② Instruction 변환 & Unseen Task 평가 설계

Instruction Template 다양화

Template 1	Based on the paragraph, can we conclude that <hypothesis>?
Template 2	Can we infer the following? <premise> / <hypothesis>
Template 3	Read the following and determine if the hypothesis can be inferred…
Template 4+	"Turned task around": premise를 생성하도록요청 등 (diversity 확보)

Dataset당 10개 template을 직접 메뉴얼하게 작성

* 다양성을 위해 일부 템플릿은 **task를 뒤집어서 (turn around)** 사용한다.
→ 양방향 mapping + 한 데이터로도 여러 task 경험 → **일반화 능력**

Good

Unseen Task 평가 설계

Cluster 단위 Hold-Out

- **같은 종류의 task 전체를 학습에서 제외**
  e.g. NLI 평가 시 → NLI 관련 dataset을 학습에서 통째로 제외
- **Dataset 단위 hold-out과의 차이**
  같은 task의 다른 dataset을 학습에 쓰면 task를 '이미 아는' 효과
  → 실제 generalization 측정 불가
- **12개 task cluster 평가 (12개 모델 각각 학습)**

목표: 본 적 없는 task 유형에서 얼마나 일반화되는지 평가

학습 설정: LaMDA-PT 137B · 30k gradient steps · Dataset당 최대 30k examples (불균형 방지) · 약 60시간 (TPUv3 128코어)

Results: Zero-Shot FLAN vs GPT-3

20/25

datasets에서
zero-shot GPT-3 초과

10/25

datasets에서
few-shot GPT-3도 초과

137B FLAN

vs 175B GPT-3

instruction tuning을 하면
더 작은 모델도 더 큰 모델을 이길 수 있다

주요 Task 성능 비교 (Zero-Shot)

Task	LaMDA-PT (base)	GPT-3 Zero-Shot	GPT-3 Few-Shot	FLAN Zero-Shot ★
NLI (avg)	47.6	42.9	53.2	56.2
Reading Comp (avg)	61.6	63.7	72.6	77.4
Closed-Book QA (avg)	48.2	49.8	55.7	56.6
Translation (EN→FR)	11.2	25.2	32.6	33.9

주의: Commonsense, Coreference 등 문장 완성 형태의 task에서는 효과가 제한적 → instruction이 불필요한 형식이기 때문

Analysis

FLAN이 잘 되는 이유를 세 가지 요인으로 분리하여 검증한다

1

Task 다양성

얼마나 다양한 종류의 task를 학습에 포함했는가?

2

모델 스케일

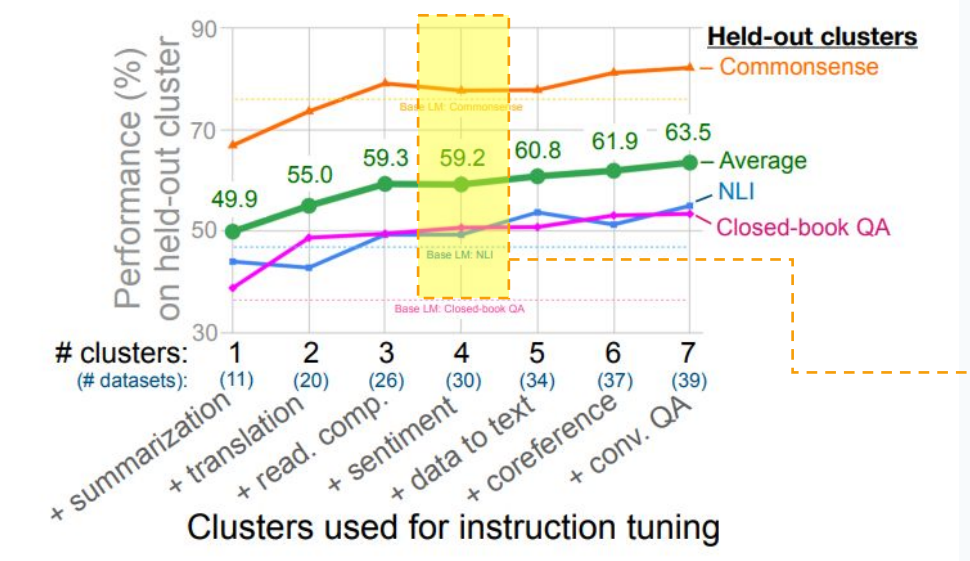
모델 크기가 instruction tuning의 효과에 미치는 영향

3

Instruction 형식

Multi-task finetuning 자체가 아닌 instruction이 핵심인가?

1 Task 다양성 : 많을수록 좋다



핵심 발견

▲ **Task cluster 수가 늘수록**
unseen task 성능 향상

→ 7개 cluster에서도 성능 포화 (saturation) 미발생
→ 더 다양하면 더 좋아질 가능성

! Sentiment analysis cluster는 추가시 기여 미미
→ 단순 수량 증가가 아닌 '다양성'이 핵심

62 Datasets → 12 Task Clusters

Natural Language Inference

7 datasets

Reading Comprehension

5 datasets

Closed-Book QA

3 datasets

Commonsense Reasoning

4 datasets

Translation

8 datasets

Summarization

11 datasets

Struct-to-Text

4 datasets

Sentiment Analysis

특정 리뷰나 평가가 담긴 문장에서 긍정적/부정적 감정을 판별 하는 task

4 datasets

Paraphrase Detection

두 개의 문장이 표면적으로는 다르더라도, 의미적으로 유사한지를 판별하는 task

4 datasets

Coreference Resolution

문서나 대화 속에서 서로 다른 표현(명사구, 대명사 등)이 실세계의 같은 개체(Entity)를 가리키고 있을 때, 이를 찾아내어 하나로 연결해 주는 task

3 datasets

Read. Comp. w/ Commonsense

2 datasets

Miscellaneous

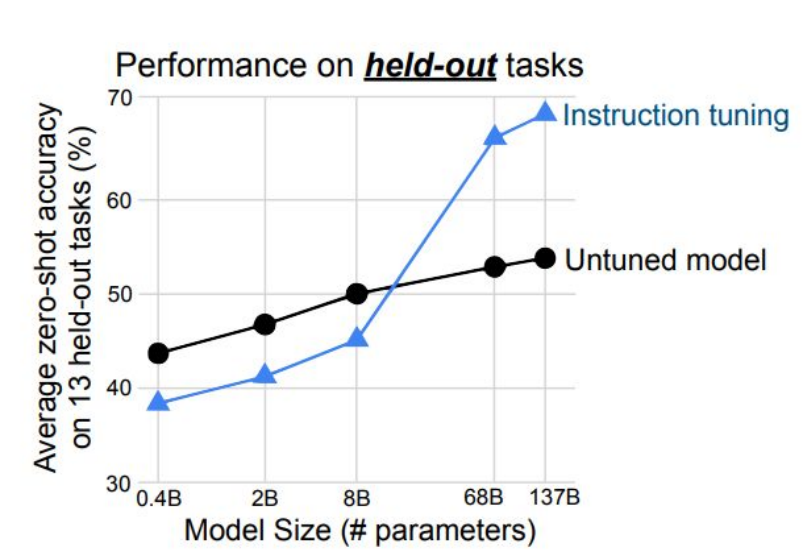
7 datasets

NLU (이해)

NLG (생성)

Classification

2 모델 스케일: 충분히 커야 효과가 있다



핵심 발견

소형 모델 (≤8B)

Instruction tuning 후 오히려 unseen task 성능 하락

→ 여러가지 task를 외우는 데 전체 capacity 소진

→ train task외의 task들에 일반화하지 못 함

대형 모델 (≥68B)

Instruction tuning 후 성능 대폭 향상

→ Task를 외우고도 남는 capacity로 instruction 이해 능력 습득

→ Instruction tuning은 대형 모델이 전제 조건

3 Instruction의 역할: Multi-task만으로는 부족하다

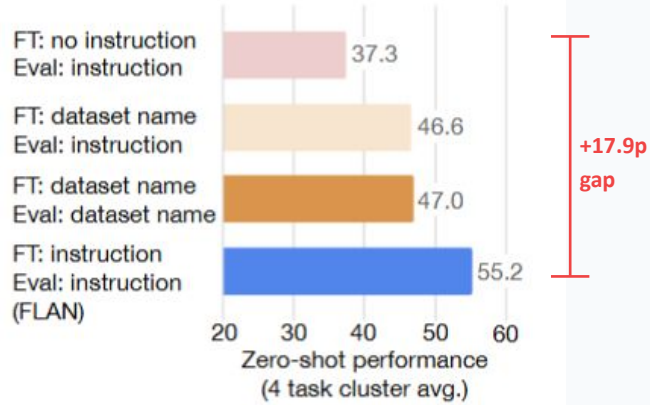


Figure 8: Ablation study result using models with instructions removed from finetuning (FT).

4 task cluster 평균 (NLI / Reading Comp / Closed-Book QA / Translation)

핵심 발견

Multi-task finetuning ≠ Instruction tuning

Task명 + dataset name을 붙여서 학습
→ instruction 없는 것과 성능 차이 미미
(37.3 → 47.0, vs FLAN 55.2)

Instruction 형식 자체가 핵심

자연어 instruction으로 학습해야
제로샷 일반화 능력이 생김

성능 향상이 multi-task 학습의
부수 효과가 아님을 직접 증명!!

Analysis: 세 요인은 각자 독립적이고 모두 필요하다

1 Task 다양성

다양한 종류의 task를 학습에 포함할수록 unseen task 성능 향상

- Cluster 수 ↑ → 성능 ↑ (saturation 없음)
- Dataset 수가 template 수보다 중요
- 다양성 > 양(quantity)

2 모델 스케일

대형 모델에서만 instruction tuning이 효과적

- ≤8B: instruction tuning이 오히려 역효과
- ≥68B: 성능 대폭 향상
- 전제 조건: 충분한 모델 capacity

3 Instruction 형식

Multi-task finetuning 자체가 아닌 instruction 형식이 핵심

- No instruction: 37.3
- Dataset name only: 47.0
- Instruction (FLAN): 55.2
- +17.9p gap

추가 발견

Few-shot exemplar + Instruction Tuning = 상호보완적 ·
Few-shot FLAN이 Zero-shot FLAN보다 모든 task cluster에서 성능 향상 (특히 복잡한 출력 형식 요구 task에서 효과적)

Takeaway

"모델은 specific task가 아니라 instruction을 따르는 방법을 배운다"

→ 충분히 큰 모델 + 다양한 task + instruction 형식이 모두 갖춰질 때 zero-shot generalization이 가능하다

의의

- 이 논문 이후 instruction tuning은 현대 LLM의 핵심 학습 방식으로 정착
- InstructGPT → ChatGPT → Claude 등 대화형 AI의 직접적 선조
- labeled data가, 즉, supervised learning이 specialist model이 아닌 generalist model을 강화하는 데 쓰일 수 있음을 실증

한계

- 소형 모델(≤8B)에는 역효과 — 범용 적용 불가
- 영어 중심, 1024 토큰 제한 — 긴 문서 처리 불가
- Commonsense reasoning 등 LM 형식 task에서는 효과 제한적
- Task cluster 분류에 주관성 개입 가능

Scaling Instruction-Finetuned Language Models

Flan / FLAN-T5

After FLAN (2022)

Chung, Hou, Longpre, Zoph, Tay, Fedus, Wei et al. · Google · NeurIPS 2022

Task 수 × 모델 크기 × Chain-of-Thought 데이터 — 세 축을 동시에 scaling하여 instruction tuning의 한계를 돌파하고 범용 방법론으로 확립한다.

Motivation: FLAN이 보여준 것, 그리고 남은 질문

FLAN (2022)

FLAN이 보여준 것

- 62개 dataset, 12개 cluster로 instruction tuning
- Zero-shot FLAN 137B > Zero-shot GPT-3 175B (25개 중 20개)
- Instruction 형식, task 다양성, 모델 크기가 핵심

→

남은 세 가지 질문

1. Task를 62개 → 1,836개로 늘리면?
2. 모델을 137B → 540B로 키우면?
3. CoT 추론 데이터도 포함하면?

핵심 문제: CoT 없이 Instruction Tuning만 하면 추론 능력이 망가진다

- FLAN 방식으로 학습(non-CoT만)하면 unseen NLP task에선 강하지만, CoT가 필요한 추론 task에서는 오히려 성능 하락
- 이유: instruction tuning이 '직접 답하는' 패턴을 강화 → 단계별 추론 능력이 억제됨
- 목표: NLP task와 추론 task 모두에서 잘 동작하는 단일 체크포인트를 만들 수 있는가?

Idea: 세 가지 축의 동시 Scaling

1 Task 수

BEFORE
62개

AFTER
1,836개

Muffin + T0-SF + NIV2 + CoT
4개 mixture 결합
473 datasets, 146 categories

2 모델 크기

BEFORE
137B
(LaMDA-PT)

AFTER
540B
(PaLM)

Flan-PaLM (8B/62B/540B)
Flan-T5 (80M~11B)
다양한 아키텍처에 적용

3 CoT 데이터 추가

BEFORE
없음

AFTER
9개 추론
dataset 포함

산술 추론, 다단계 추론,
NLI + CoT annotation
인간이 직접 작성한 추론 과정

이 세 가지를 동시에 적용한 finetuning procedure를 "Flan finetuning"이라 부르며,
Flan-PaLM / Flan-T5 등 여러 모델에 동일하게 적용

Finetuning tasks 데이터 구조

Muffin

69 datasets | 80 tasks

- NLI, Closed-book QA
- Code 생성·수정
- 원래 FLAN 62개 포함

T0-SF

55 datasets | 193 tasks

- Commonsense reasoning
- Adversarial QA
- Struct-to-text
- T0 중 Muffin 미포함분

NIV2

372 datasets | 1,554 tasks

- 전체 task 중 가장 큰 비중
- Named entity recognition
- Toxic language detection

CoT (Reasoning)

9 datasets | 9 tasks

- 산술 추론, 다단계 추론
- 자연어 추론 + 설명
- 인간이 작성한 CoT 주석

합계:

473 datasets	146 task categories	1,836 total tasks	공개 오픈소스 ✓

- **총 1,000만개 가량의 Q&A pair 사용**

Finetuning tasks 데이터 구조

Muffin

69 datasets | 80 tasks

- NLI, Closed-book QA
- Code 생성·수정
- 원래 FLAN 62개 포함

T0-SF

55 datasets | 193 tasks

- Commonsense reasoning
- Adversarial QA
- Struct-to-text
- T0 중 Muffin 미포함분

NIV2

372 datasets | 1,554 tasks

- 전체 task 중 가장 큰 비중
- Named entity recognition
- Toxic language detection

CoT (Reasoning)

9 datasets | 9 tasks

- 산술 추론, 다단계 추론
- 자연어 추론 + 설명
- 인간이 작성한 CoT 주석

합계:

473 datasets | **146 task categories** | **1,836 total tasks** | **공개 오픈소스 ✓**

Finetuning tasks 데이터 구조

Muffin	T0-SF	NIV2	CoT (Reasoning)
69 datasets	55 datasets	372 datasets	9 datasets
80 tasks	193 tasks	1,554 tasks	9 tasks

Held-out tasks

MMLU

Abstract algebra
College medicine
Professional law
Sociology
Philosophy
...

57 tasks

BBH

Boolean expressions
Tracking shuffled objects
Dyck languages
Navigate
Word sorting
...

27 tasks

TyDiQA

Information seeking QA

8 languages

MGSM

Grade school math problems

10 languages

합계:	473 datasets	146 task categories	1,836 total tasks	공개 오픈소스 ✓

Finetuning tasks 데이터 구조



이전 FLAN에서는 task cluster 단위로 hold-out

vs

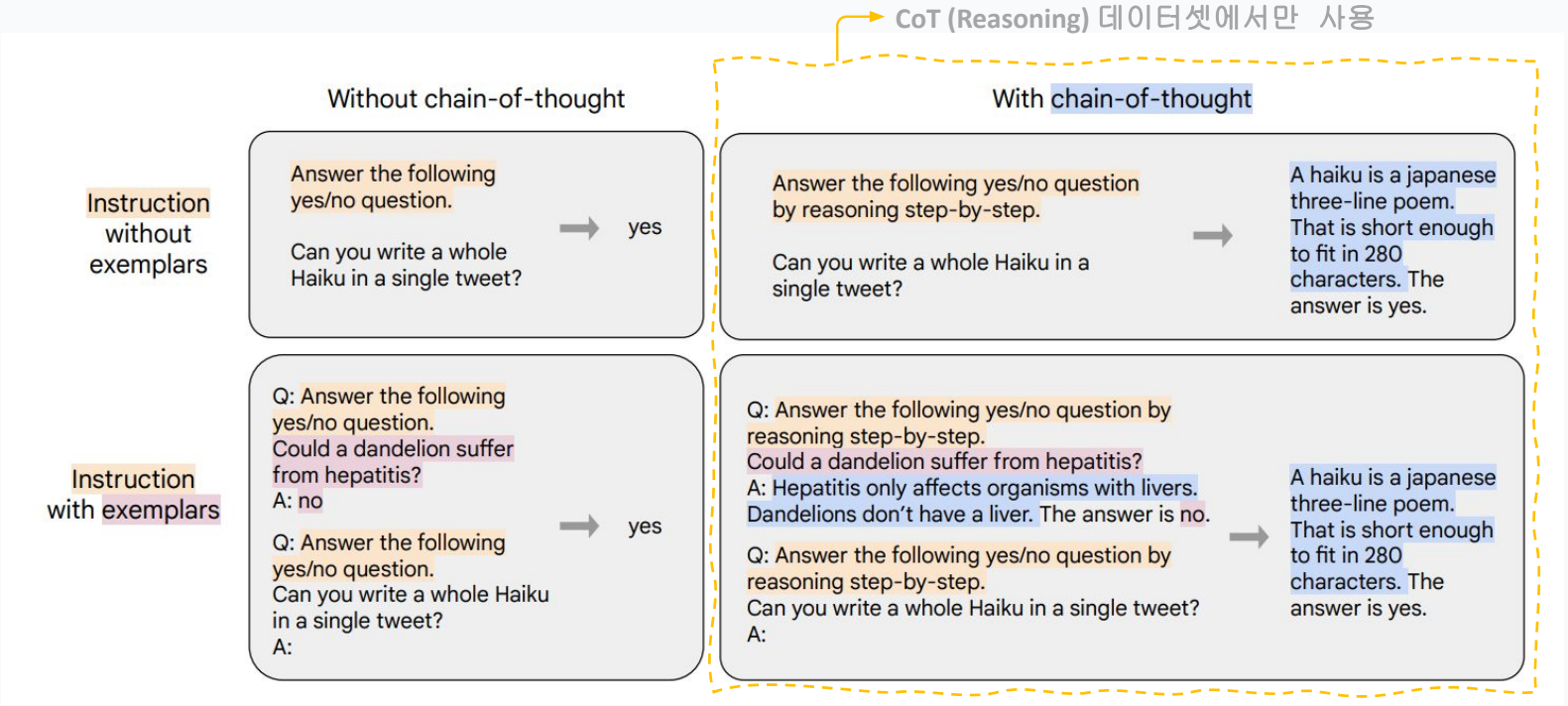
FLAN T5에서는 dataset 단위 hold-out

□ 완전한 Unseen task를 풀게 하지는 않음

단지 새로운 dataset를 해결한다는 것에 의의를 둠

Method ① 학습 데이터 포맷: 4가지 조합

핵심 설계



Method ① 학습 데이터 포맷: 4가지 조합

핵심 설계



핵심: 이 4가지 형식을 동시에 학습 → 어떤 입력 형식에도 대응 가능한 단일 체크포인트 (CoT mixture는 9개 dataset만 사용)

Method ② 평가 Benchmark & 모델 패밀리

평가 Benchmark (4종)

FLAN 기존 평가셋은 이미 학습 데이터에 포함
→ 더 어려운 **benchmark** 사용
(= 현재 언어 모델이 전문가 수준의 인간 평가자보다 훨씬 낮은 성능을 보이는 벤치마크)

MMLU	57개 분야 문제 (수학, 역사, 법, 의학 등) 5-shot, direct/CoT 평가
BBH	27개 어려운 추론 task 3-shot, direct/CoT 평가
TyDiQA	8개 언어 정보탐색 QA, 다국어 능력 측정 1-shot, exact match
MGSM	10개 언어 수학 문제, 다국어 추론 능력 8-shot CoT 평가

모델 패밀리 (Flan-T5 공개)

Flan-T5-Small	80M	encoder-decoder
Flan-T5-Base	250M	encoder-decoder
Flan-T5-Large	780M	encoder-decoder
Flan-T5-XL	3B	encoder-decoder
Flan-T5-XXL	11B	encoder-decoder
Flan-PaLM	8~540B	Decoder-only

GPT-3 초과

최고 성능

★ 공개

Finetuning compute = pretraining의 0.2%에 불과 · Flan-T5 checkpoint 공개 오픈소스

Results: Flan Finetuning의 성과

+9.4%

Flan-PaLM 540B
평균 성능 향상

75.2%

MMLU (당시 SOTA)
CoT + Self-consistency

Flan-T5 11B > PaLM 62B

BBH-direct에서
소형 모델이 대형 초과

주요 모델 성능 비교 (Normalized Avg)

Model	Params	No Finetuning	Flan Finetuned	향상폭
T5-XXL → Flan-T5-XXL	11B	-2.9	23.7	+26.6
PaLM → Flan-PaLM	8B	6.4	21.9	+15.5
PaLM → Flan-PaLM	62B	28.4	38.8	+10.4
PaLM → Flan-PaLM	540B	49.1	58.4	+9.3

Flan-T5 XL (3B)의 MMLU 52.4% > GPT-3 175B의 43.9% · compute는 pretraining의 0.2%만 사용

Analysis

세 가지 확장 각각이 얼마나 기여했는가, 그리고 서로 어떻게 상호작용하는가

1 **Task 수 Scaling**
더 많은 task를 포함할수록 unseen task 성능이 어떻게 변하는가?

+ Dataset 수 vs Template 수의 상대적 기여

2 **모델 크기 Scaling**
8B / 62B / 540B — 크기에 따라 finetuning 효과가 달라지는가?

FLAN 논문과의 비교: 소형 모델에서의 차이

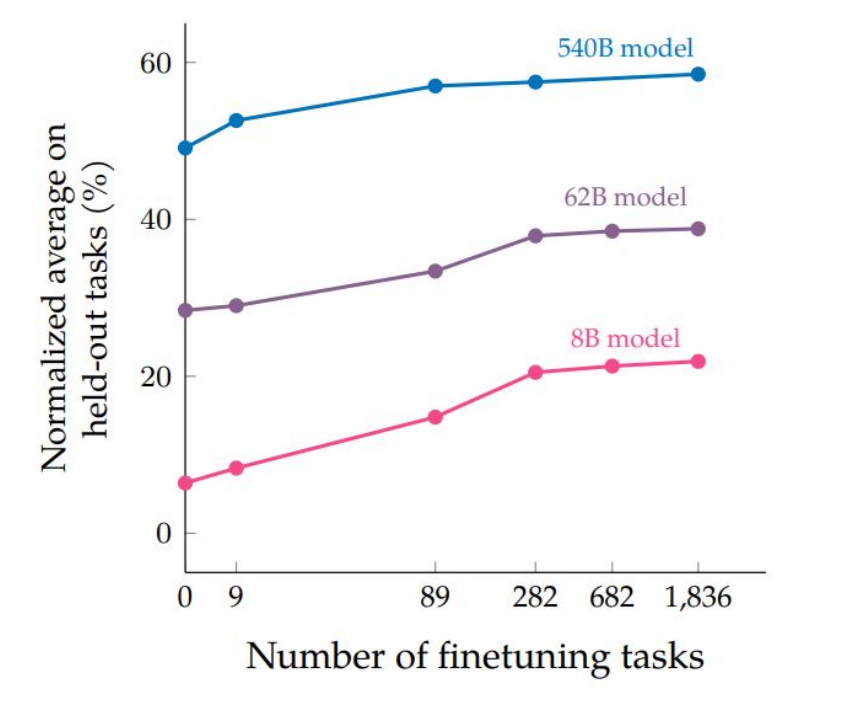
3 **CoT 데이터의 역할**
Non-CoT만 학습하면 추론이 망가진다 — CoT 9개 데이터로 어떻게 해결하는가?

Zero-shot CoT: 'let's think step-by-step'이 가능해진다

4 **종합 & 범용성**
세 요인의 상호작용 + 다양한 아키텍처에서의 범용성 확인

Flan-T5 공개 오픈소스의 의의

① Task 수 Scaling: 많을수록 좋지만 diminishing returns



핵심 발견

Task 수 ↑ → 성능 ↑

모든 모델 크기에서 finetuning이 성능 향상 (절대적)

282개 이후 Diminishing Returns

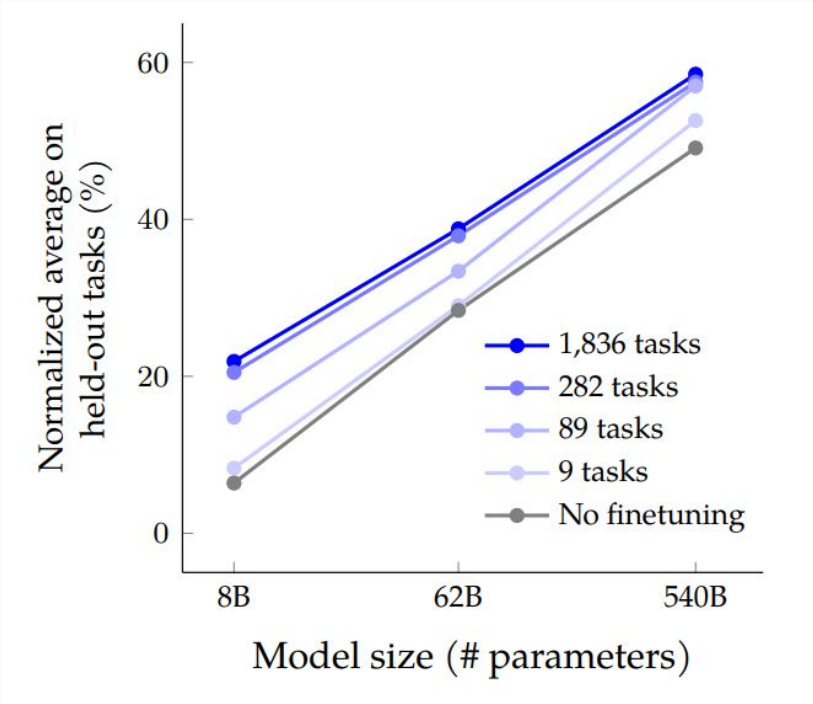
282개까지는 급격히 향상, 이후는 증분 효과

→ 여전히 positive

해석

1. **추가 task가 기존과 비슷해 다양성 기여 적음**
2. 대부분 이득은 '아는 것을 잘 표현하는 법'을 배우는 것
   — pretraining 지식 활용

② 모델 크기: 모든 크기에서 효과적 (FLAN과의 차이)



FLAN 논문과의 차이

FLAN (2022) — 소형 모델 역효과

8B 이하에서 instruction tunin으로 오히려 성능 하락

(모델이 특정 task 패턴을 암기하는데 capacity를 다 씀)

Flan-T5 — 8B에서도 향상 ✓

더 많은 task 덕분에 암기가 불가능.

결국 instruction 자체를 이해하도록 강제되어,

소형 모델도 충분한 instruction 이해 능력 습득

→ 모든 크기에서 finetuning 효과적

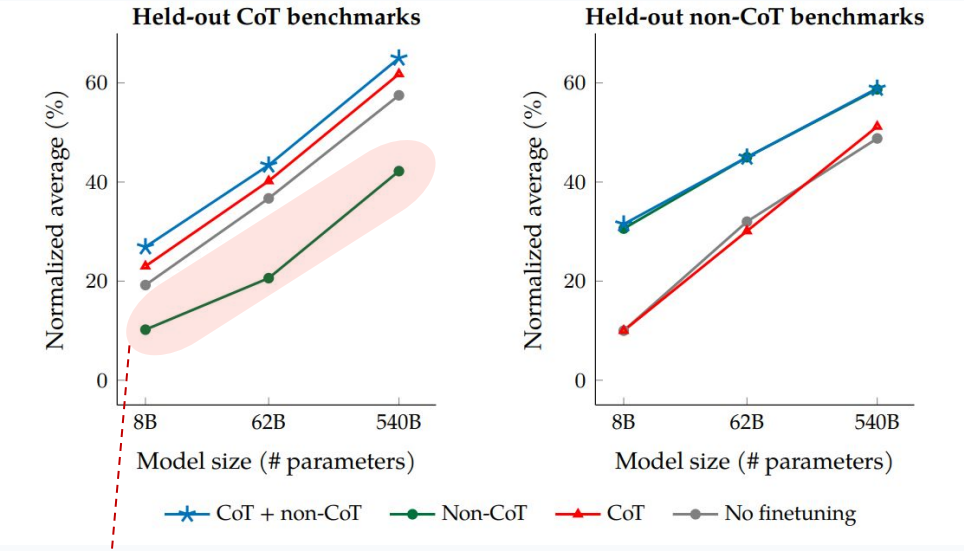
향상폭 해석

절대값: 8B > 540B (15.5% vs 9.4%)

Error rate 감소: 540B > 8B (18.4% vs 16.6%)

→ **큰 모델이 더 효율적으로 개선**

③ CoT 데이터 : 9개 데이터셋이 추론 능력을 살린다



⚠️ **Non-CoT만 학습하면 CoT 성능이 "No finetuning"보다도 낮아진다**

→ CoT 9개 데이터셋만 추가해도 이 문제가 해결되고 non-CoT 성능도 유지됨

③ CoT 데이터 : Zero-Shot CoT — PaLM vs Flan-PaLM

	Disambiguation QA	Snarks (Sarcasm)
Input	Q: 다음 문장에서 대명사의 선행사를 설명하거나, 모호하다고 판단하라. Sentence: ~, Options: (A), (B), (C) Ambiguous	Which is sarcastic? Wow thats crazy, I've never had a takeout meal that used (A)hardwood/(B) plastic containers before!
PaLM 540B	The reporter and the chef will discuss their favorite dishes….→ ❌ 질문에 답하지 않음	What is sarcasm? The use of irony… → ❌ 무한 생성 (멈추지 않음)
Flan-PaLM 540B	Does not indicate whose dishes → ✓ The answer: (C) Ambiguous	플라스틱 컵은 흔함 → ✓ The answer: (B)

Zero-shot CoT 능력 획득

"let's think step-by-step"

Flan-PaLM은 이 한마디로 zero-shot CoT 가능

PaLM (finetuning 없음)

동일한 trigger를 줘도 CoT 생성 실패

→ 답 틀리거나 무한 반복

왜 중요한가?

Few-shot exemplar 없이도 CoT 유도 가능

= 실사용성 대폭 향상

Prompt engineering 없이도 복잡한 추론 task 수행

④ 종합: 세 요인은 독립적이며 범용 방법론으로 작동한다

1 Task 다양성

더 많은 task = 더 나은 generalization

- 282개까지 급격히 향상, 이후 diminishing returns
- 여전히 positive — 계속 늘리는 게 유익
- 다양성 > 단순 양(quantity)

2 모델 크기

모든 크기에서 효과적 (FLAN과 차이)

- FLAN: 8B 이하 역효과
- Flan-T5: 80M에서도 향상
- 더 큰 모델이 error rate 감소 효율 높음

3 CoT 데이터

9개 데이터셋으로 추론 능력 보존+강화

- Non-CoT only → CoT 성능 대폭 하락
- CoT 추가 시 양쪽 모두 유지/향상
- Zero-shot CoT 능력 획득

범용성 확인 — Flan finetuning은 아키텍처를 가리지 않는다

T5 (encoder-decoder)	PaLM (decoder-only)	U-PaLM (prefix LM)	cont-PaLM
80M ~ 11B	8B ~ 540B	540B	62B

Flan-T5 (80M~11B) 공개 오픈소스 → 연구 생태계 기여 · 소형이지만 강력 · Hugging Face에서 현재도 널리 사용

Takeaway

"Instruction tuning은 모델 크기, 아키텍처, task 수에 무관하게 범용적으로 작동하는 방법론이다"

→ 충분한 task 다양성 + CoT 포함 + 대형 모델 선호 — 세 조건이 갖춰지면 zero-shot, few-shot, CoT 모두에서 강력한 단일 체크포인트

FLAN (2022- 2월) → Flan-T5 / Flan-PaLM (2022- 10월) → InstructGPT → ChatGPT / Claude

이 논문은 instruction tuning을 62개 task/137B → 1,836개 task/540B로 확장하며 현대 LLM의 핵심 학습 방법론으로 확립시킨 연구

의의

- Flan-T5 오픈소스 공개

  → 현재 Hugging Face 최다 사용 instruction-tuned 모델 중 하나
- 모든 모델 패밀리에 적용 가능한 범용 방법론 확립
- CoT finetuning이 추론 능력 보존에 필수임을 최초로 체계적으로 증명
- Compute 0.2%로 평균 9~15% 성능 향상 — 비용 효율성 입증

한계

- 영어 중심 (다국어는 TyDiQA/MGSM에서 부분 검증에 그침)
- 독성·편향 문제 여전히 존재 (Appendix C 참고)
- 코드 생성 등 전문화 모델 (Codex)에는 여전히 못 미치는 영역 존재
- CoT finetuning도 충분히 큰 모델에서만 효과적

흐름

각 연구가 어떻게 영향을 주었는가

0

GPT-3 (2020)

논문: *Language Models are **Few-Shot Learners***

- “instruction 없이도 prompt만 잘 주면 된다”는 패러다임 생성

1

FLAN (2021)

논문: ***Finetuned** Language Models Are **Zero-Shot Learners***

- ‘예시 없이도 문제 풀 수 있게 할 수 없어?’
  → instruction이 뭔지 모델에게 이해시키자!
- 90만개 Q&A 쌍으로 SFT
- **instruction tuning의 시작점**

2

Chain-of-Thought (CoT) (2022)

논문: *Chain-of-Thought Prompting Elicits Reasoning in LLMs*

- reasoning 능력 폭발적으로 향상
- **“instruction tuning + reasoning” 결합 시작**

3

FLAN-T5 (2022)

논문: *Scaling Instruction-Finetuned Language Models*

- CoT로 해결해야 하는 reasoning task 들을 기존 FLAN이 잘 못 푸네?
  → CoT 데이터 포함, 모델 크기와 학습 데이터 크게 늘림
  ⇒ reasoning task 성능 크게 상승
- instruction tuning을 제대로 scale하면 LLM 성능 크게 올라간다