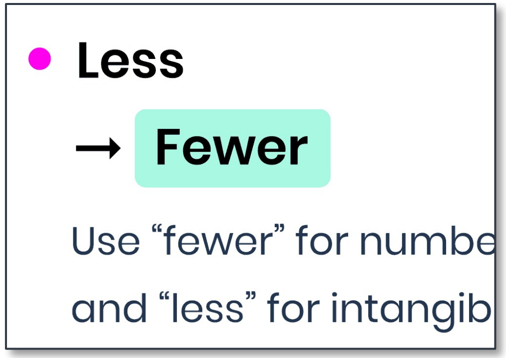
# The Quest Toward Generality in Natural Language Understanding

Daniel Khashabi

Allen Institute for AI

# AI-driven Language Interfaces

... are everywhere!

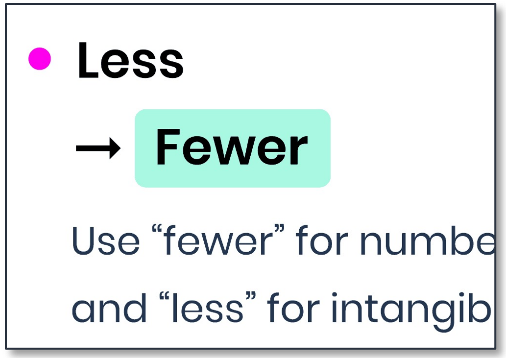




# AI-driven Language Interfaces

... are everywhere!

.... have narrow targets!

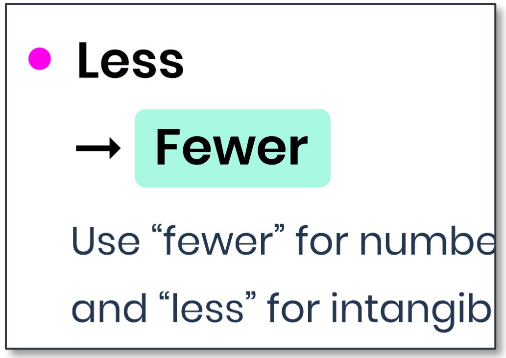




# AI-driven Language Interfaces

… are everywhere!

…. have narrow targets!





Why no single “general” system?

# AI's Inception w/ a Broad Vision

*"By 'general intelligent action' … a behavior appropriate to the ends of the system and adaptive to the demands of the environment can occur."*


[Newell and Simon, 1959 & 1976]

# AI’s Inception w/ a Broad Vision

> *“By ‘general intelligent action’ … a behavior appropriate to the ends of the system and adaptive to the demands of the environment can occur.”*

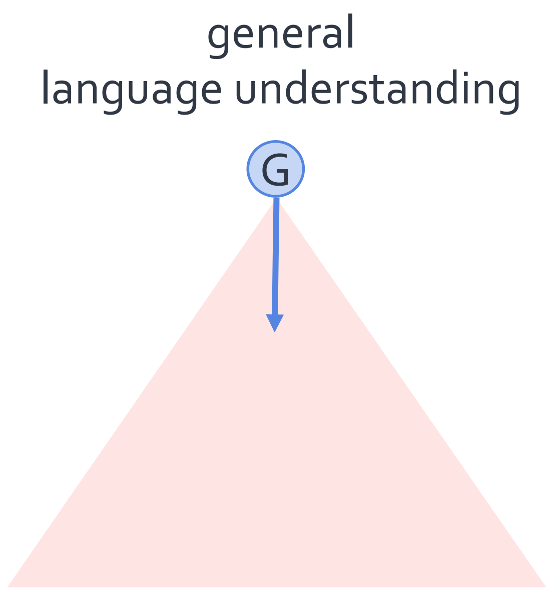
General Problem Solver

[Newell and Simon, 1959 & 1976]

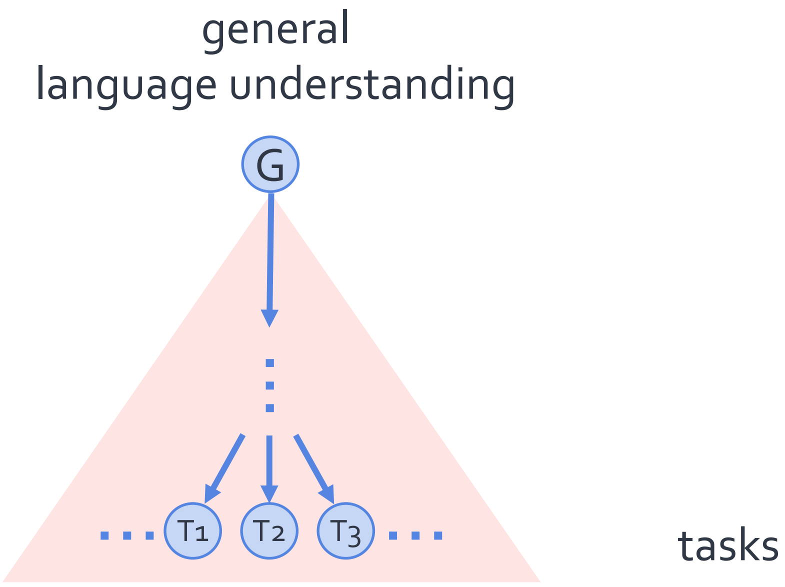
# The Great Separation

- "General language understanding" broken into many narrowed tasks



# The Great Separation

- "General language understanding" broken into many narrowed tasks



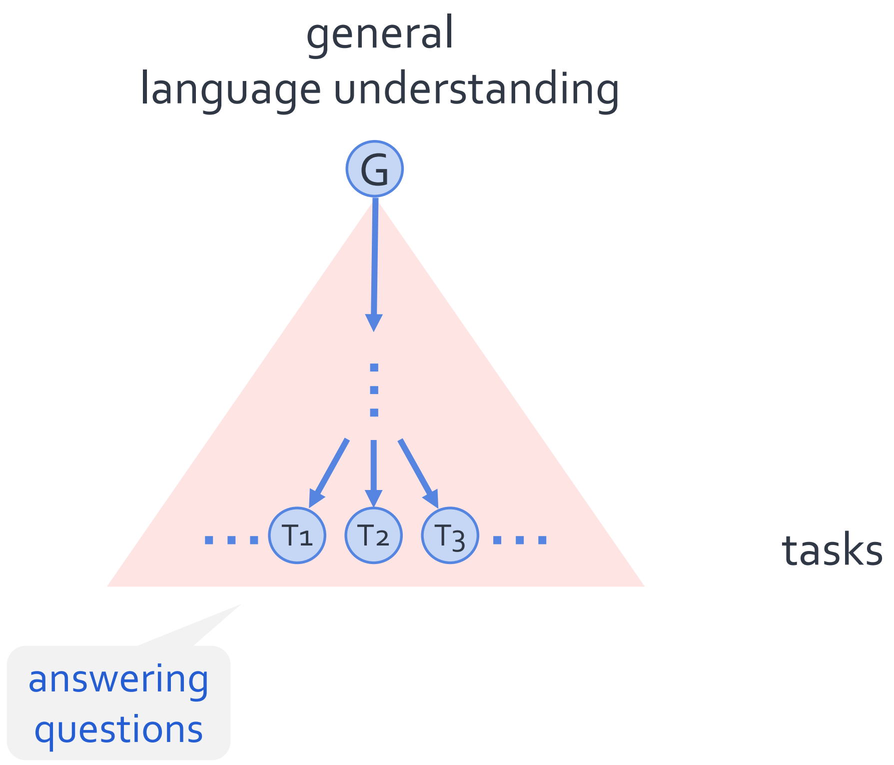

Generalization Hierarchy [Chollet, '19]

# The Great Separation

- "General language understanding" broken into many narrowed tasks

T1 Task: answering questions

$\mathbf{x}$ = "How long did ....." $\longrightarrow$ $\mathbf{y}$ = "2 years"




Generalization Hierarchy [Chollet, '19]

# The Great Separation

- "General language understanding" broken into many narrowed tasks

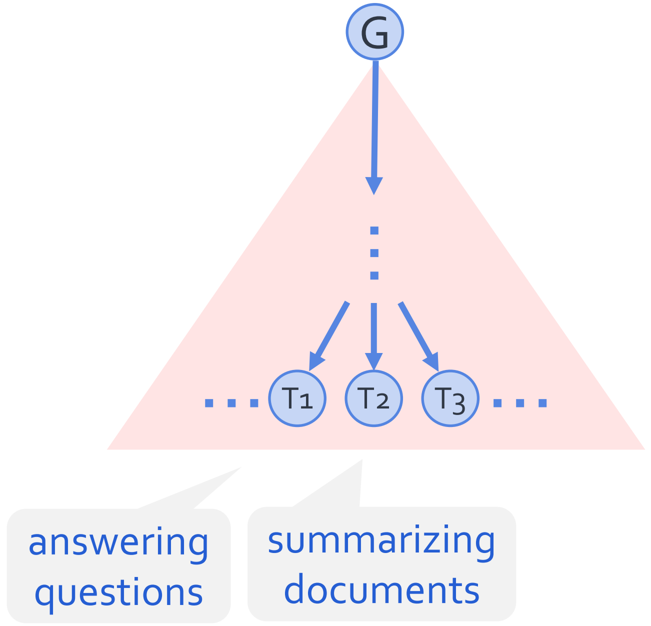
T1 Task: answering questions

$\mathbf{x}$ = "How long did ....."  $\longrightarrow$  $\mathbf{y}$ = "2 years"

T2 Task: summarizing documents

$\mathbf{x}$ =   $\longrightarrow$  $\mathbf{y}$ = " U.S. troops will …"

general
language understanding




tasks

Generalization Hierarchy [Chollet, '19]

# The Great Separation

- "General language understanding" broken into many narrowed tasks

**T1 Task: answering questions**

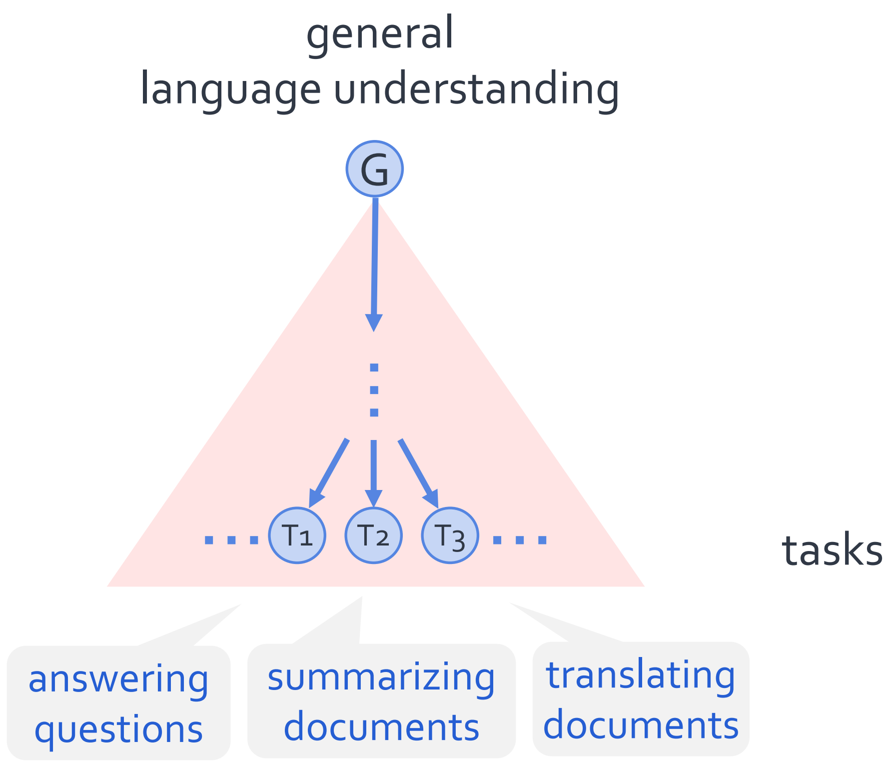
$\mathbf{x} =$ "How long did ... ." $\longrightarrow$ $\mathbf{y} =$ "2 years"

**T2 Task: summarizing documents**

$\mathbf{x} =$  $\longrightarrow$ $\mathbf{y} =$ " U.S. troops will ..."

**T3 Task: translating documents**

$\mathbf{x} =$ "Enjoyed ... ! " $\longrightarrow$ $\mathbf{y} =$ "¡Disfruté ...!"




Generalization Hierarchy [Chollet, '19]

# The Great Separation

- "General language understanding" broken into many narrowed tasks

T1 Task: answering questions

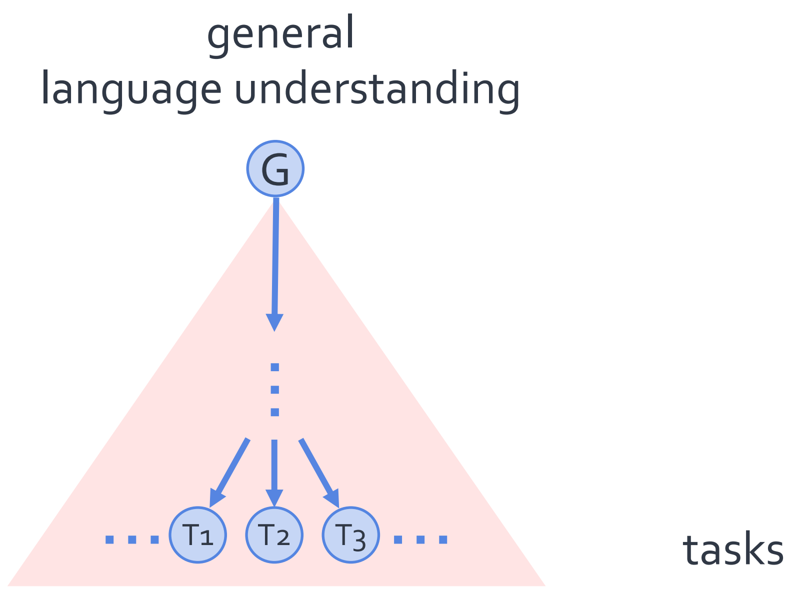
$\mathbf{x}$ = "How long did ....." $\longrightarrow$ $\mathbf{y}$ = "2 years"

T2 Task: summarizing documents

$\mathbf{x}$ =  $\longrightarrow$ $\mathbf{y}$ = " U.S. troops will …"

T3 Task: translating documents

$\mathbf{x}$ = "Enjoyed ... ! " $\longrightarrow$ $\mathbf{y}$ = "¡Disfruté …!"




Generalization Hierarchy [Chollet, '19]

# The Great Separation

- "General language understanding" broken into many narrowed tasks

T1 Task: answering questions

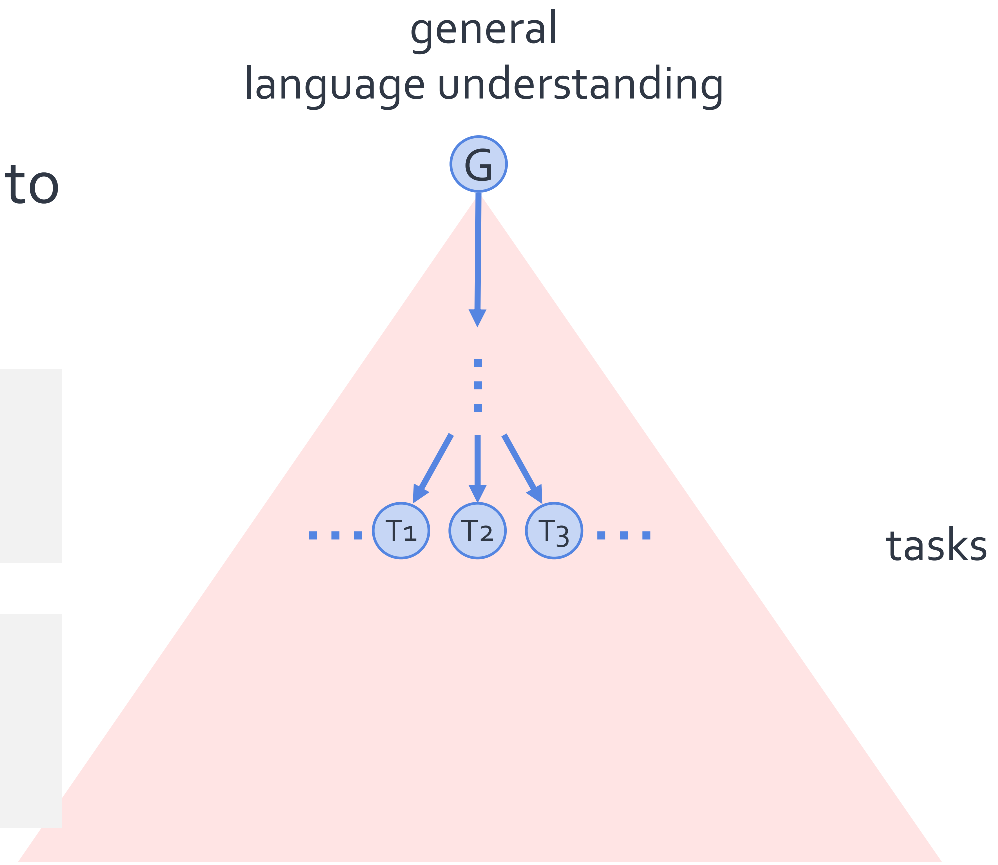
$\mathbf{x}$ = "How long did ... ." $\longrightarrow$ $\mathbf{y}$ = "2 years"

T2 Task: summarizing documents

$\mathbf{x}$ = $\longrightarrow$ $\mathbf{y}$ = " U.S. troops will ..."

T3 Task: translating documents

$\mathbf{x}$ = "Enjoyed ... ! " $\longrightarrow$ $\mathbf{y}$ = "¡Disfruté ...!"




Generalization Hierarchy [Chollet, '19]

# The Great Separation

- “General language understanding” broken into many narrowed tasks

(T1) Task: answering questions

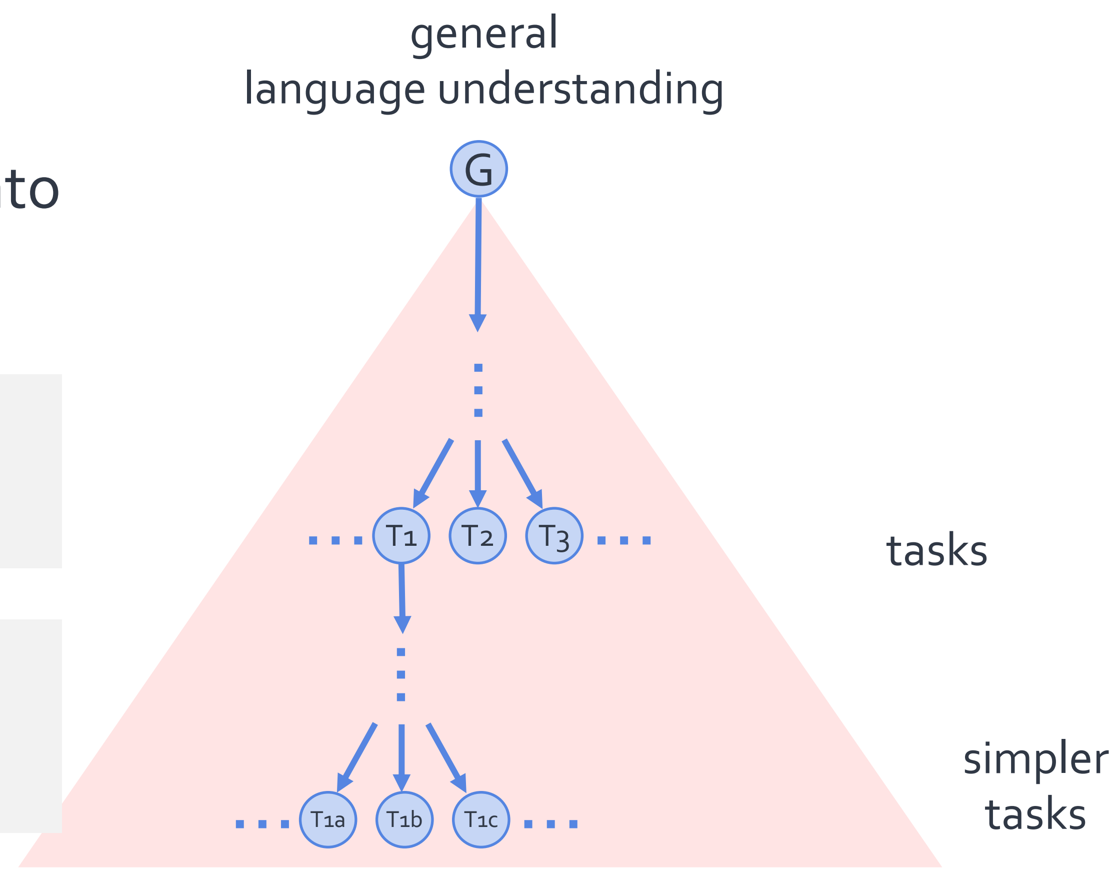
$\mathbf{x}$ = ”How long did ... .” $\longrightarrow$ $\mathbf{y}$ = ”2 years”

(T2) Task: summarizing documents

$\mathbf{x}$ =  $\longrightarrow$ $\mathbf{y}$ = ” U.S. troops will ...”

(T3) Task: translating documents

$\mathbf{x}$ = ”Enjoyed ... ! ” $\longrightarrow$ $\mathbf{y}$ = ”¡Disfruté ...!"



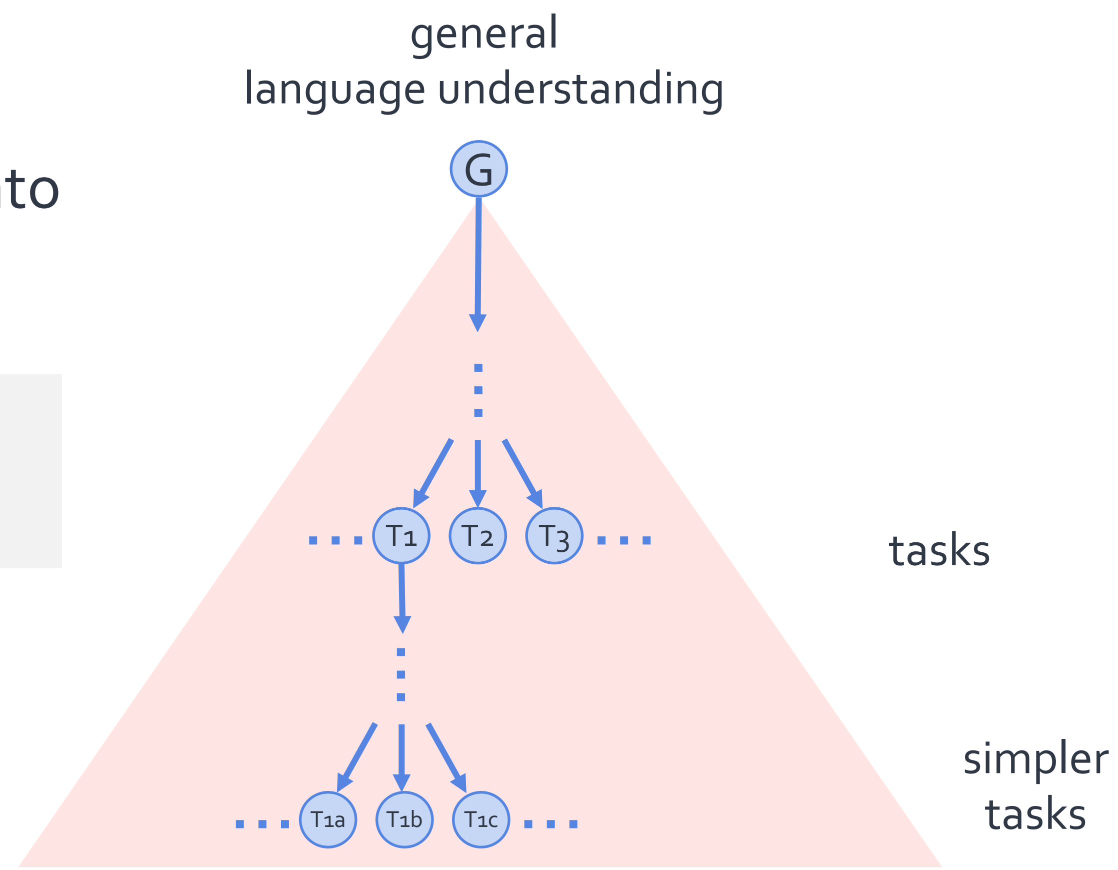

Generalization Hierarchy [Chollet, ‘19]

# The Great Separation

- "General language understanding" broken into many narrowed tasks

T1 Task: answering questions

$\mathbf{x}$ = "How long did …." $\longrightarrow$ $\mathbf{y}$ = "2 years"




Generalization Hierarchy [Chollet, '19]

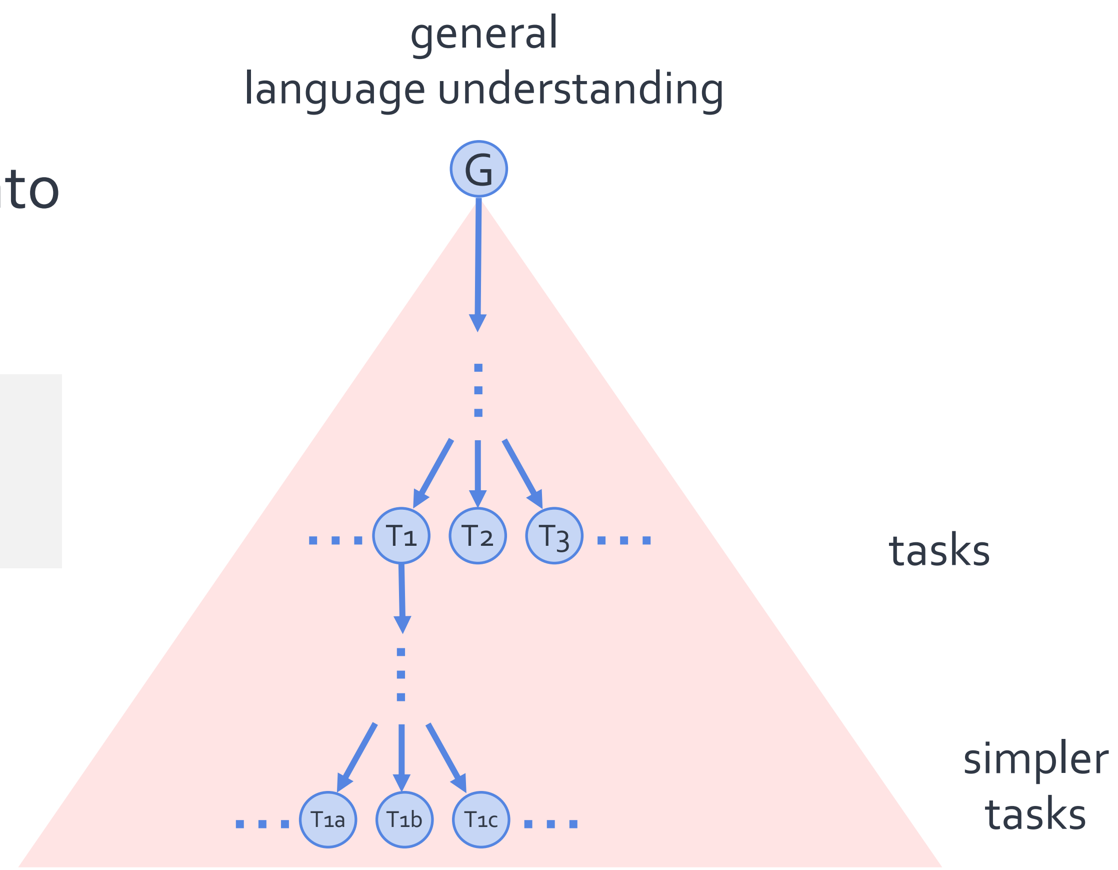
# The Great Separation

- "General language understanding" broken into many narrowed tasks

T1 Task: answering questions

$\mathbf{x}$ = "How long did ....." $\longrightarrow$ $\mathbf{y}$ = "2 years"

T1a answering simple factual questions




Generalization Hierarchy [Chollet, '19]

# The Great Separation

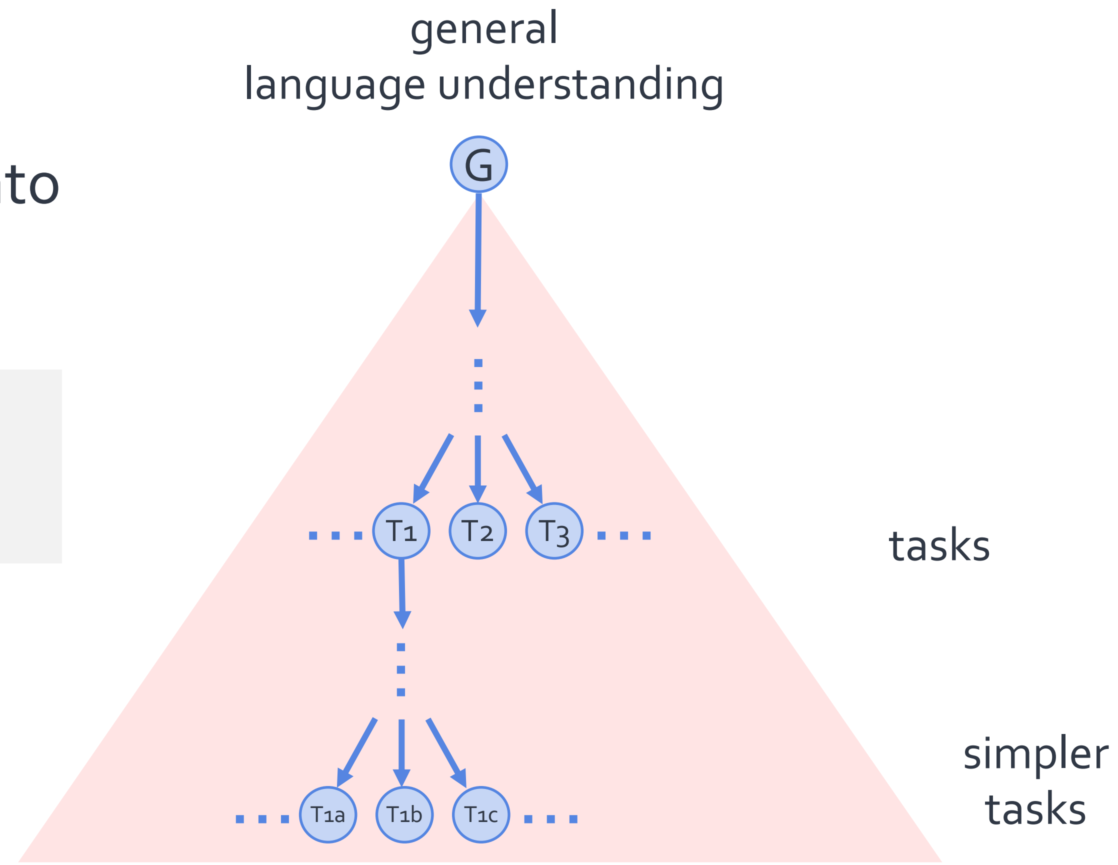
- "General language understanding" broken into many narrowed tasks

T1 Task: answering questions

$\mathbf{x} =$ "How long did ....." $\longrightarrow$ $\mathbf{y} =$ "2 years"

T1a answering simple factual questions

T1b answering algebra questions




Generalization Hierarchy [Chollet, '19]

# The Great Separation

- "General language understanding" broken into many narrowed tasks

T1 Task: answering questions

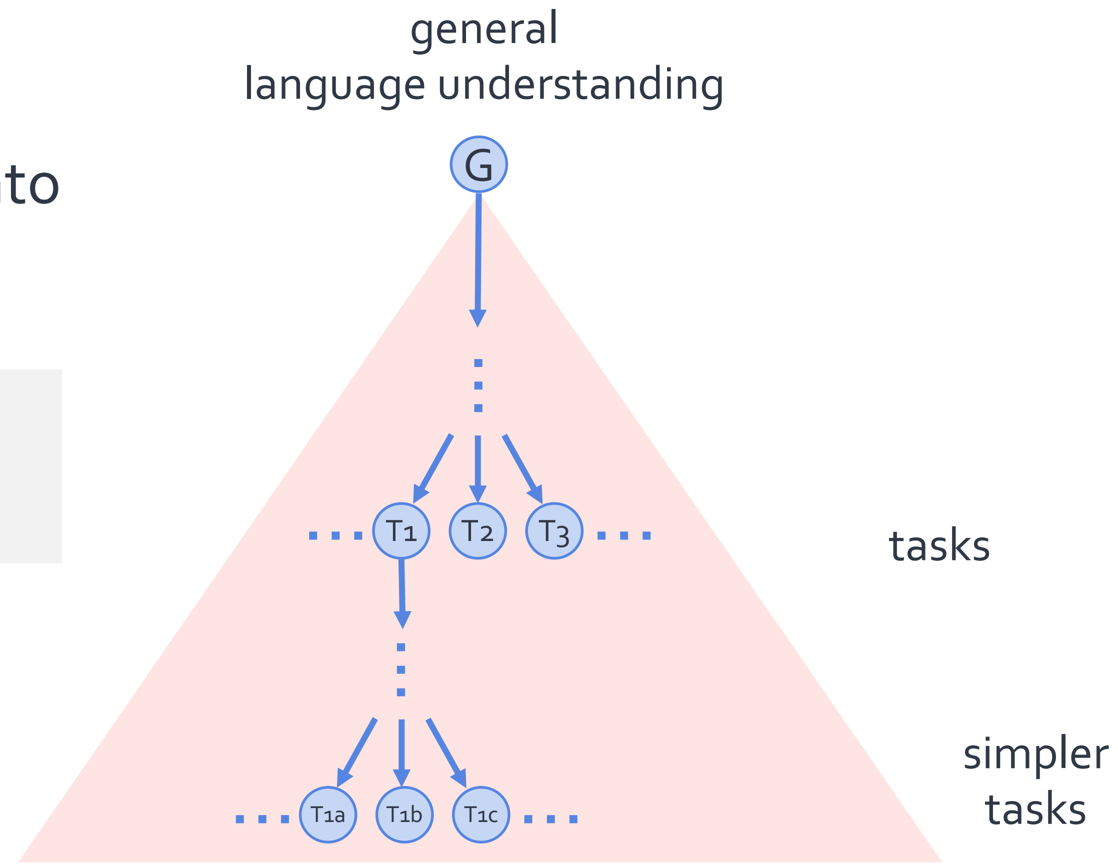
$\mathbf{x}$ = "How long did ...." $\longrightarrow$ $\mathbf{y}$ = "2 years"

T1a answering simple factual questions

T1b answering algebra questions

T1b answering elementary school questions

⋮

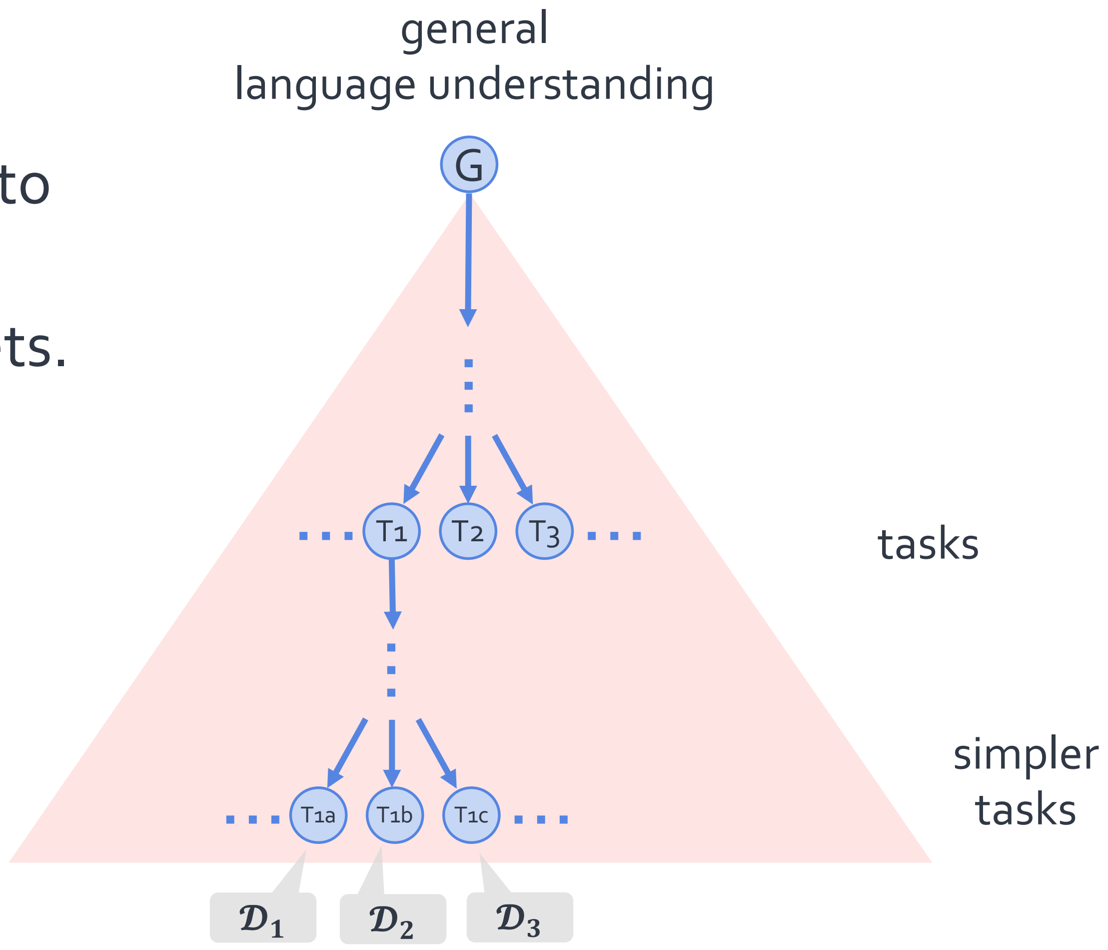



Generalization Hierarchy [Chollet, '19]

# The Dataset Heaven

- "General language understanding" broken into many narrowed tasks
- Subtasks instantiated as input-output datasets.

$$(\mathbf{x}, \mathbf{y}) \sim \mathrm{T} \quad \rightarrow \mathcal{D} = \{(\mathbf{x}, \mathbf{y})\}$$



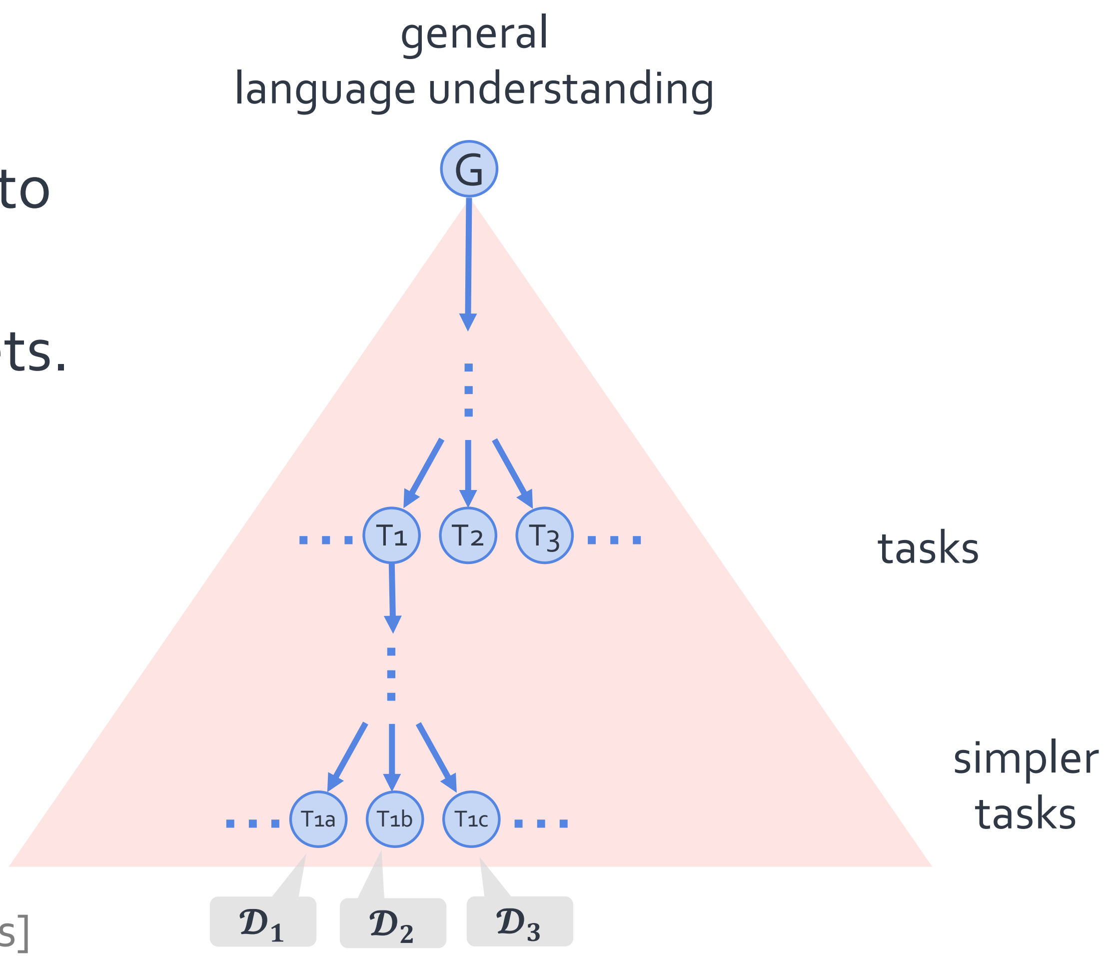

Generalization Hierarchy [Chollet, '19]

# Success at Dataset Level

- "General language understanding" broken into many narrowed tasks
- Subtasks instantiated as input-output datasets.

$$(\mathbf{x}, \mathbf{y}) \sim \text{T} \quad \rightarrow \mathcal{D} = \{(\mathbf{x}, \mathbf{y})\}$$

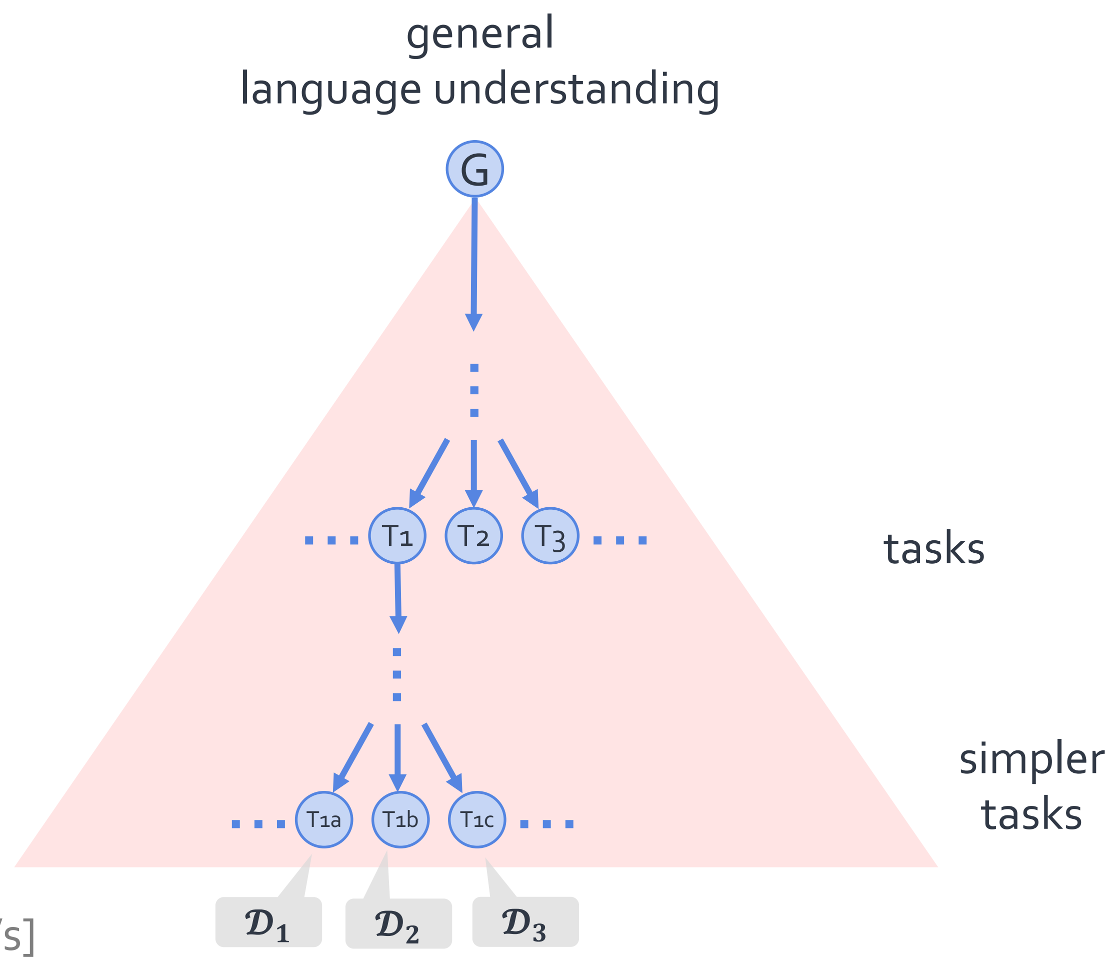
- Statistical models [Brown, Jelinek and others, late 80's]
  - Fitting a parameterized model to datasets



# Success at Dataset Level

**Neural Language Models**

[Bengio et al. '04, Peters et al. '18, Raffel et al. '20, Brown et al. '20, …]

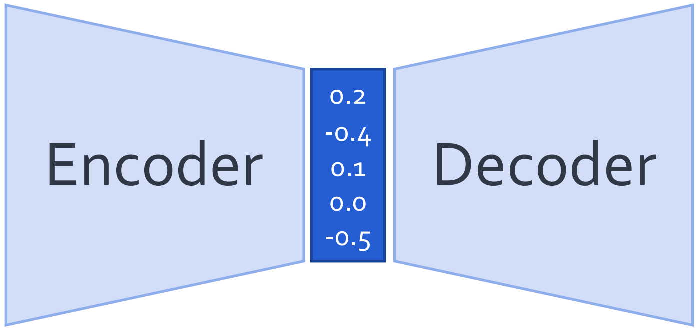


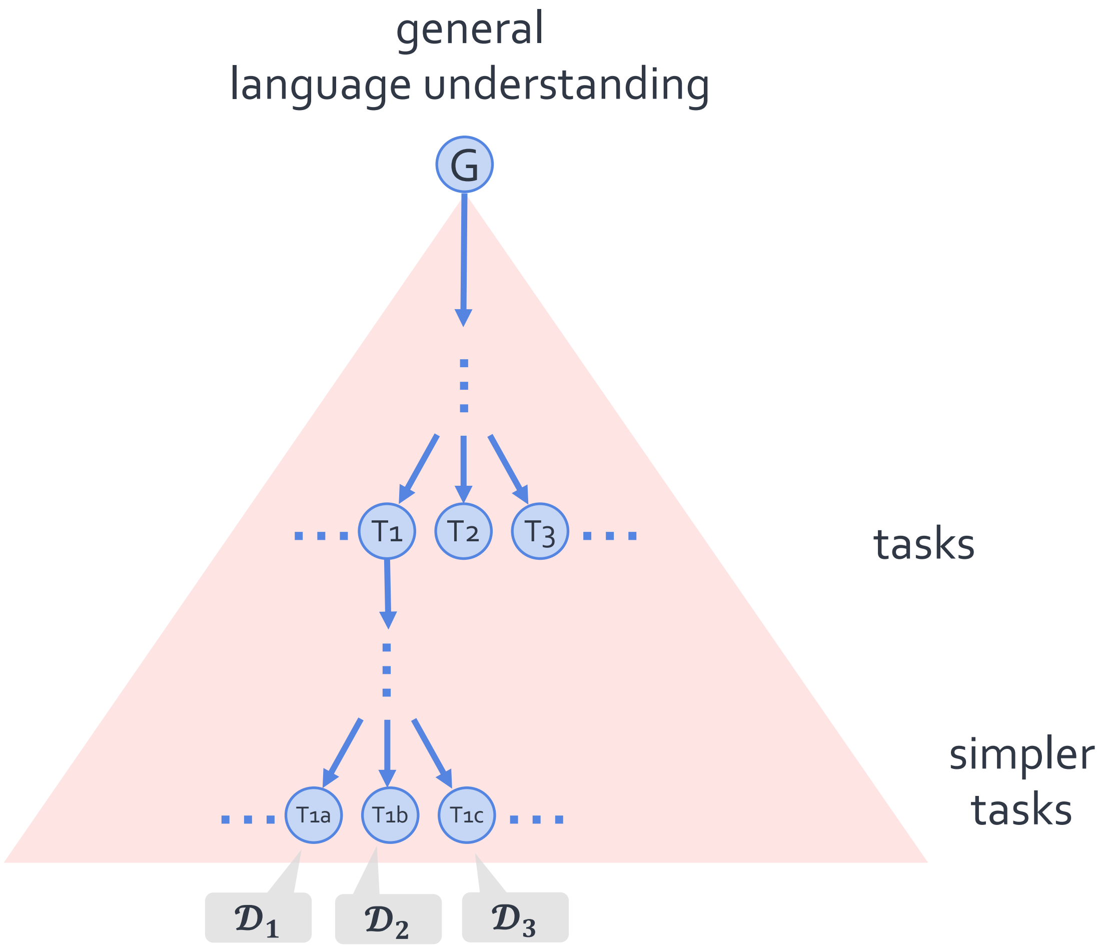

- Statistical models [Brown, Jelinek and others, late 80's]
  - Fitting a parameterized model to datasets

# Success at Dataset Level

## Neural Language Models

[Bengio et al. '04, Peters et al. '18, Raffel et al. '20, Brown et al. '20, ...]

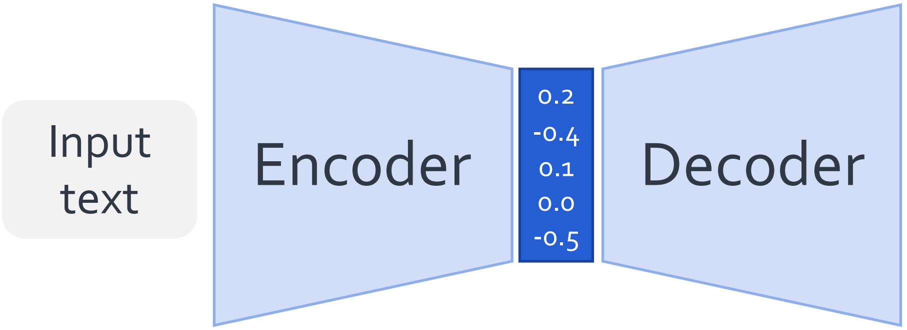


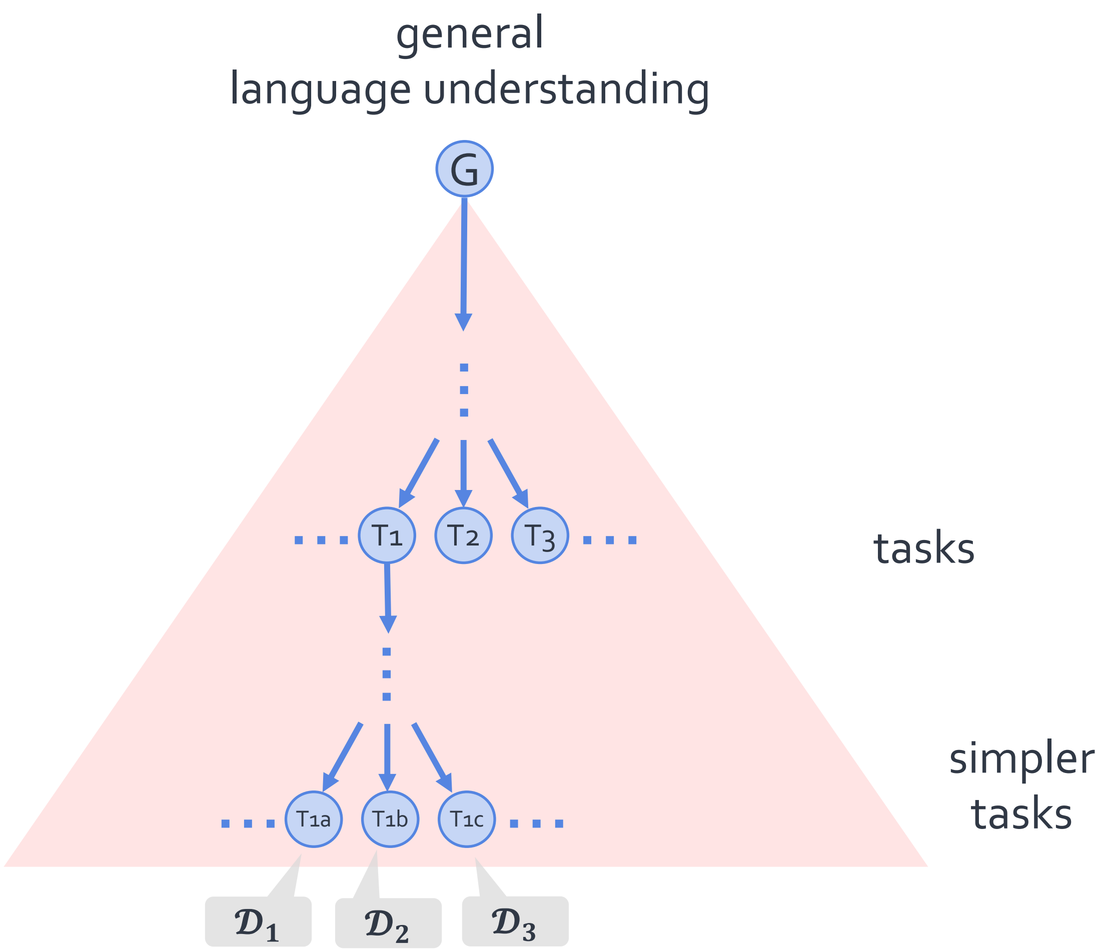



# Success at Dataset Level

## Neural Language Models

[Bengio et al. '04, Peters et al. '18, Raffel et al. '20, Brown et al. '20, ...]

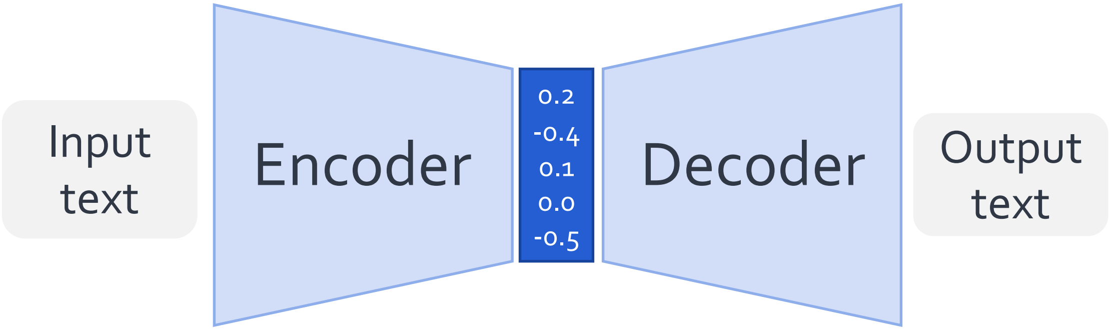


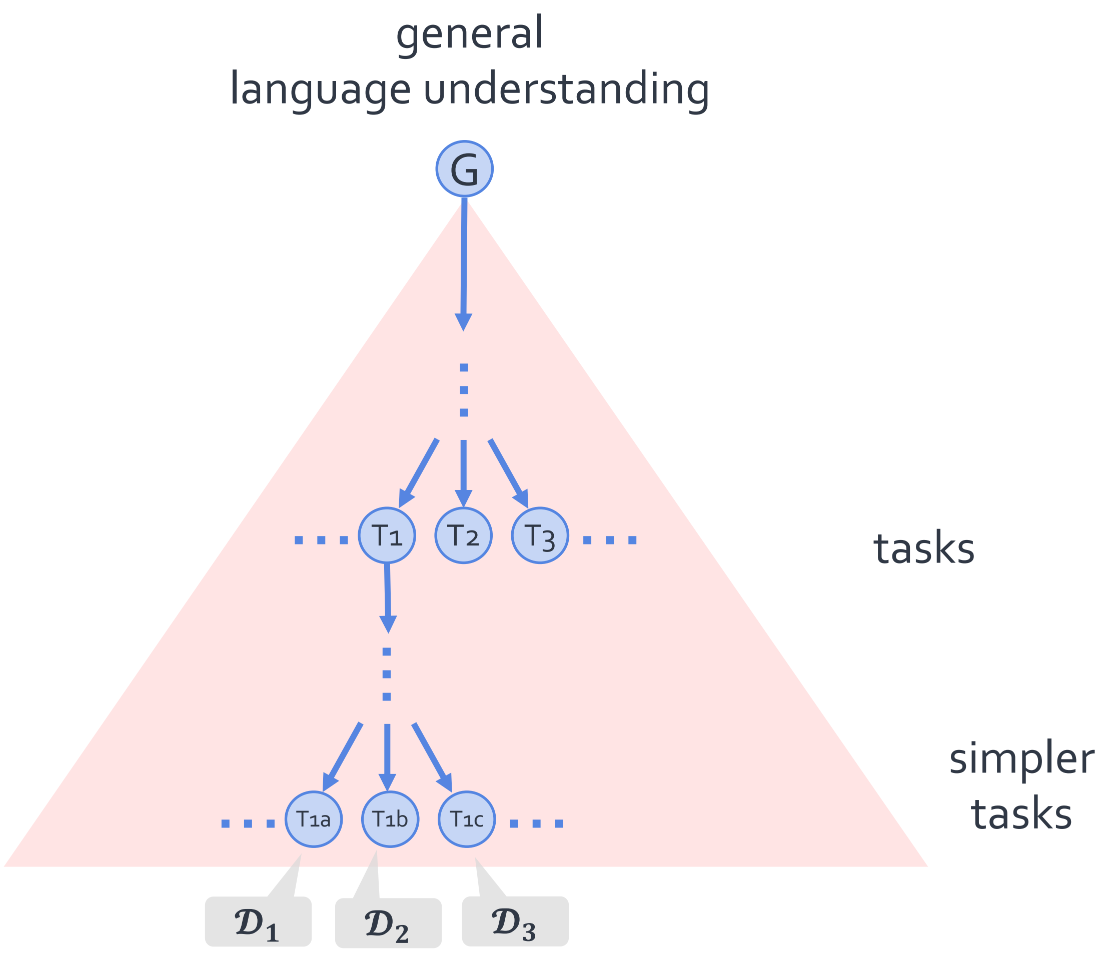



# Success at Dataset Level

**Neural Language Models**

[Bengio et al. '04, Peters et al. '18, Raffel et al. '20, Brown et al. '20, ...]

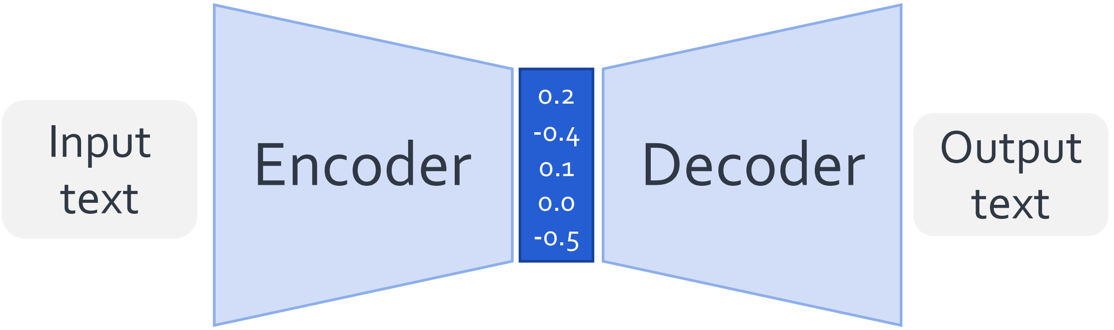


# Success at Dataset Level

## Neural Language Models

[Bengio et al. '04, Peters et al. '18,
Raffel et al. '20, Brown et al. '20, ...]

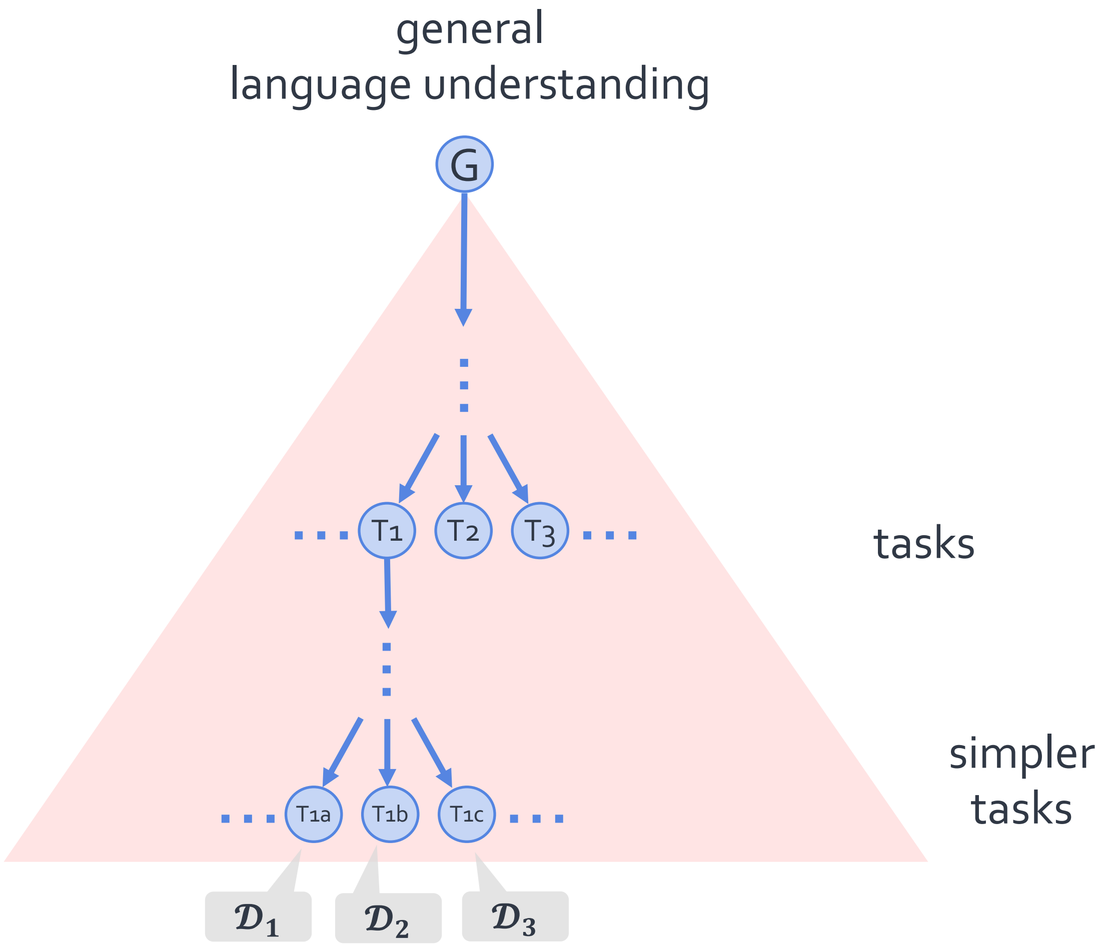



**T5**

Google AI

**GPT-3**

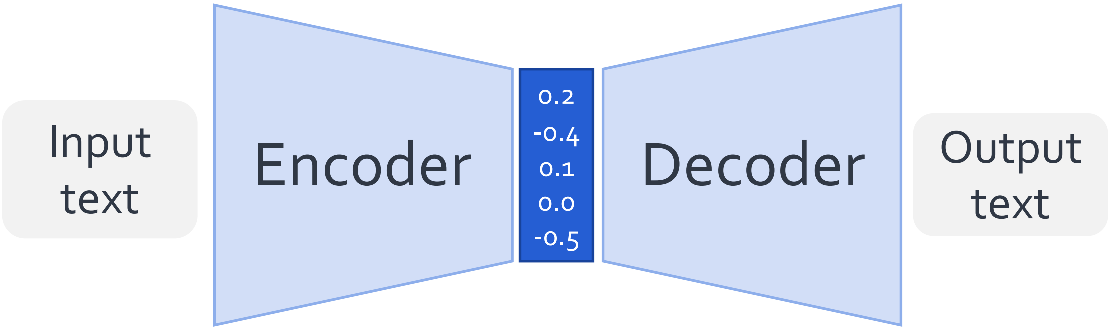
OpenAI



# Success at Dataset Level

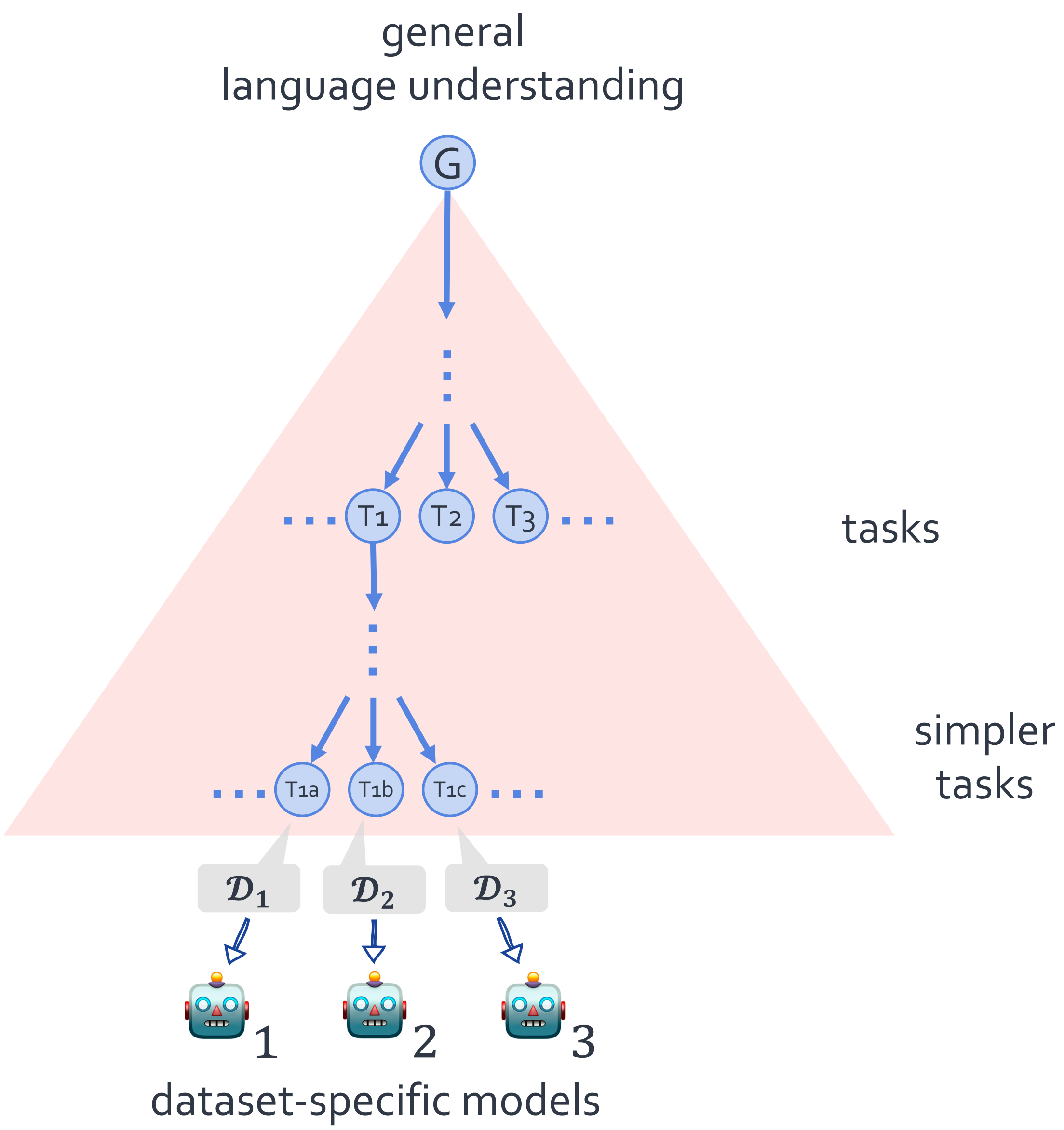
## Neural Language Models

[Bengio et al. '04, Peters et al. '18,
Raffel et al. '20, Brown et al. '20, ...]

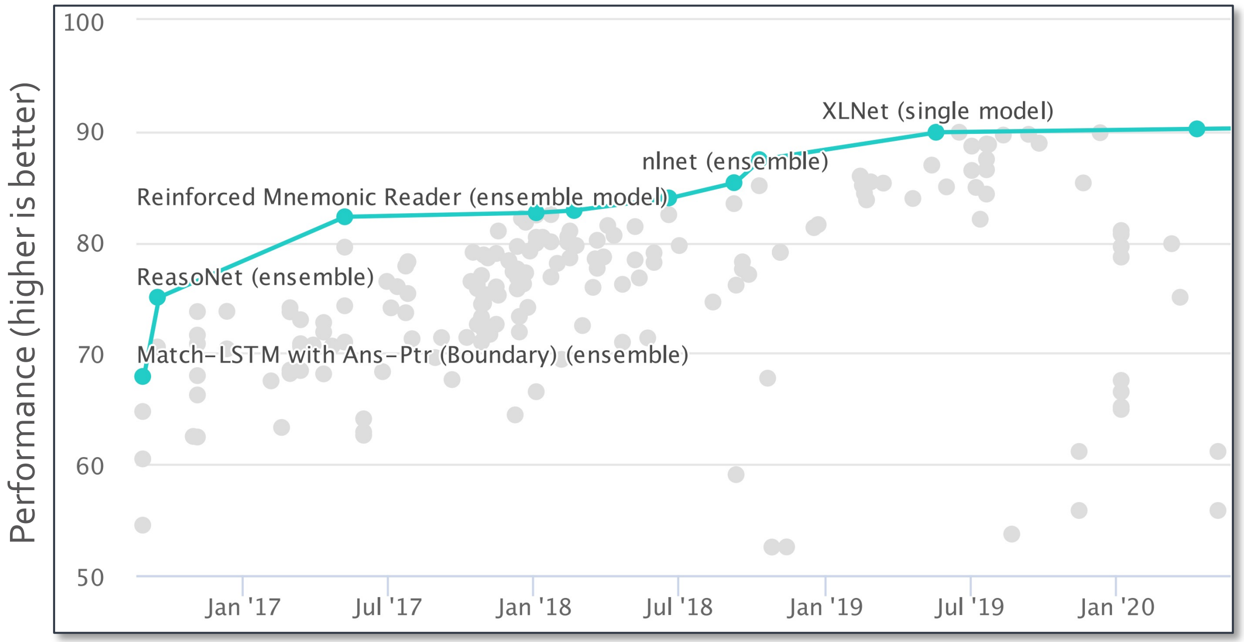



**T5**
Google AI

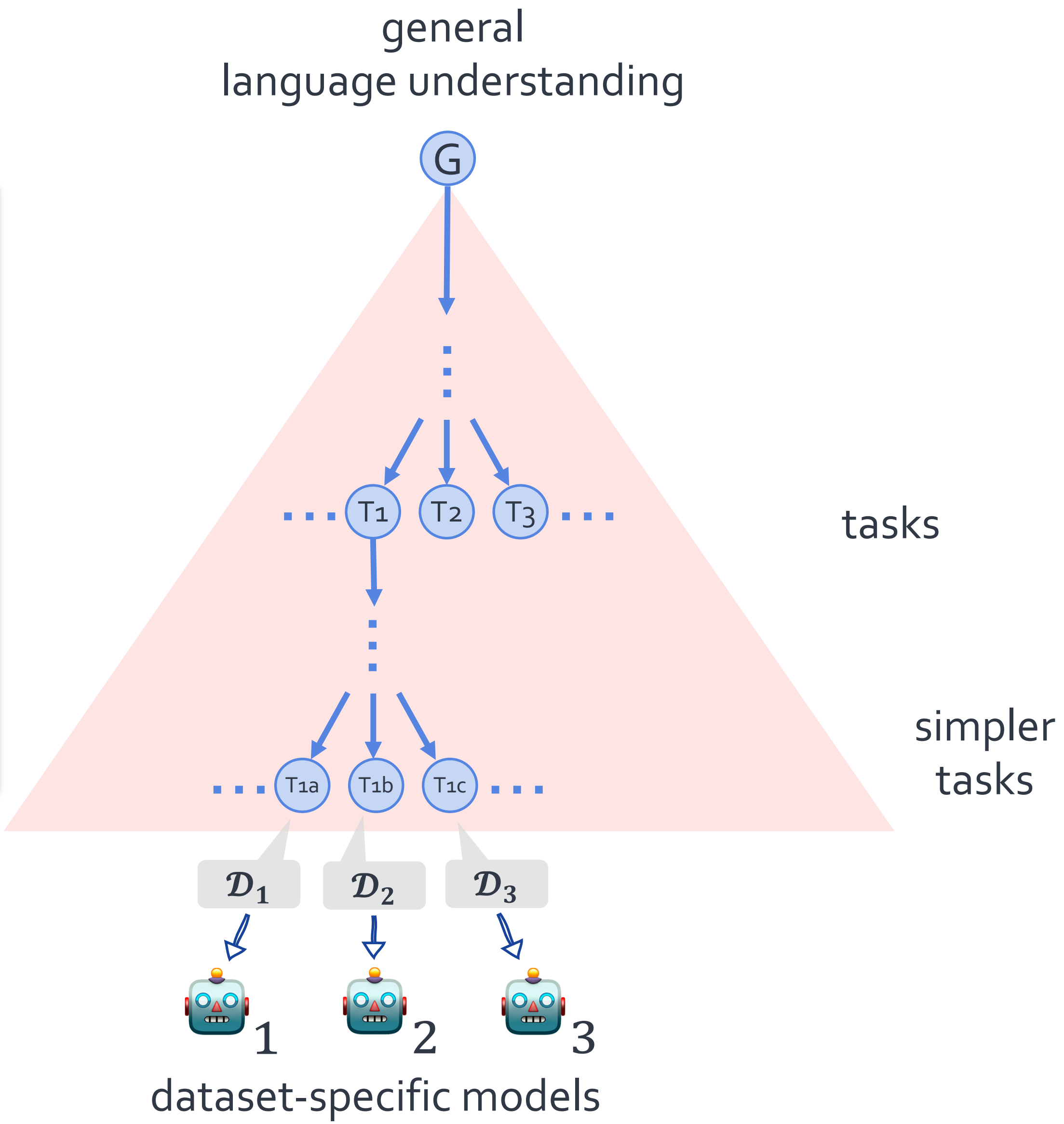
**GPT-3**
OpenAI



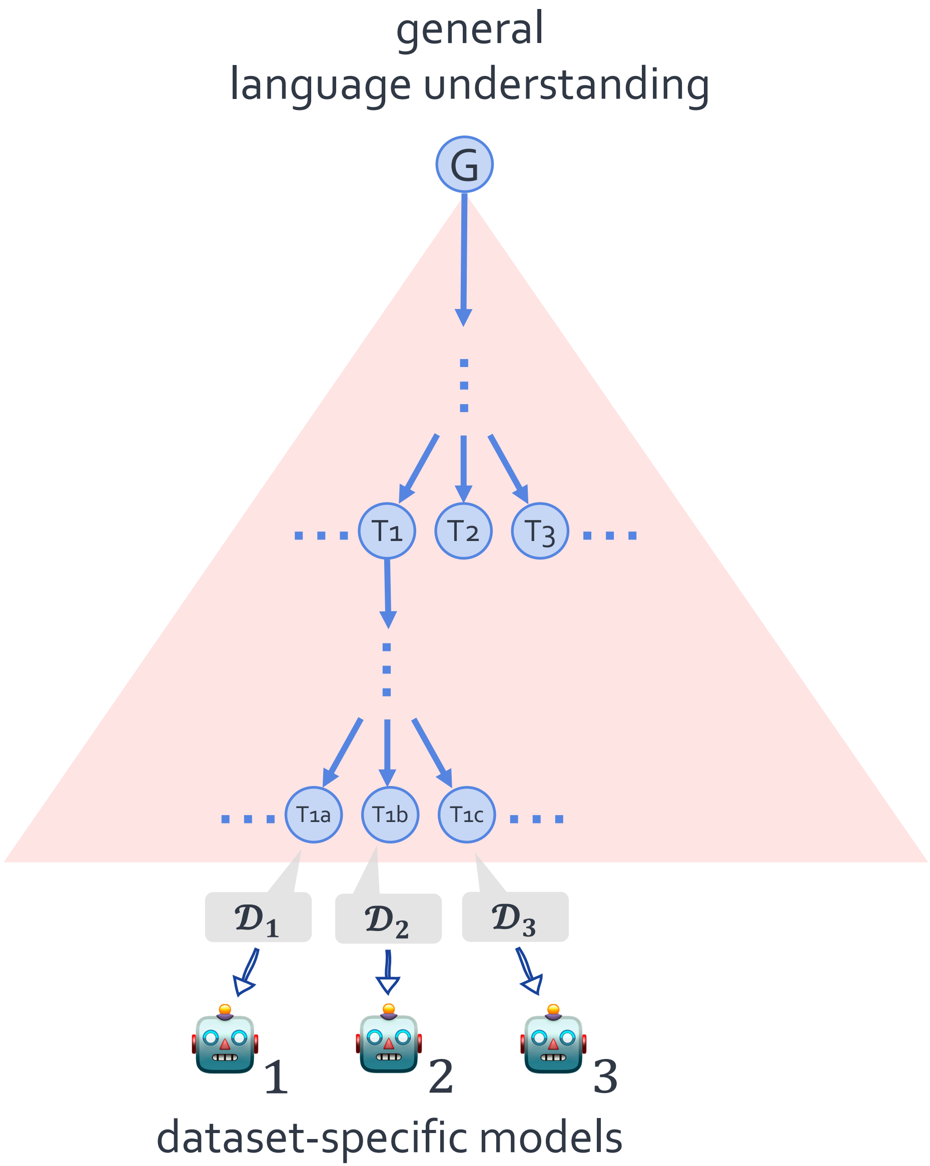
# Success at Dataset Level




progress over time
on a question answering benchmark
[Rajpurkar et al. '16]

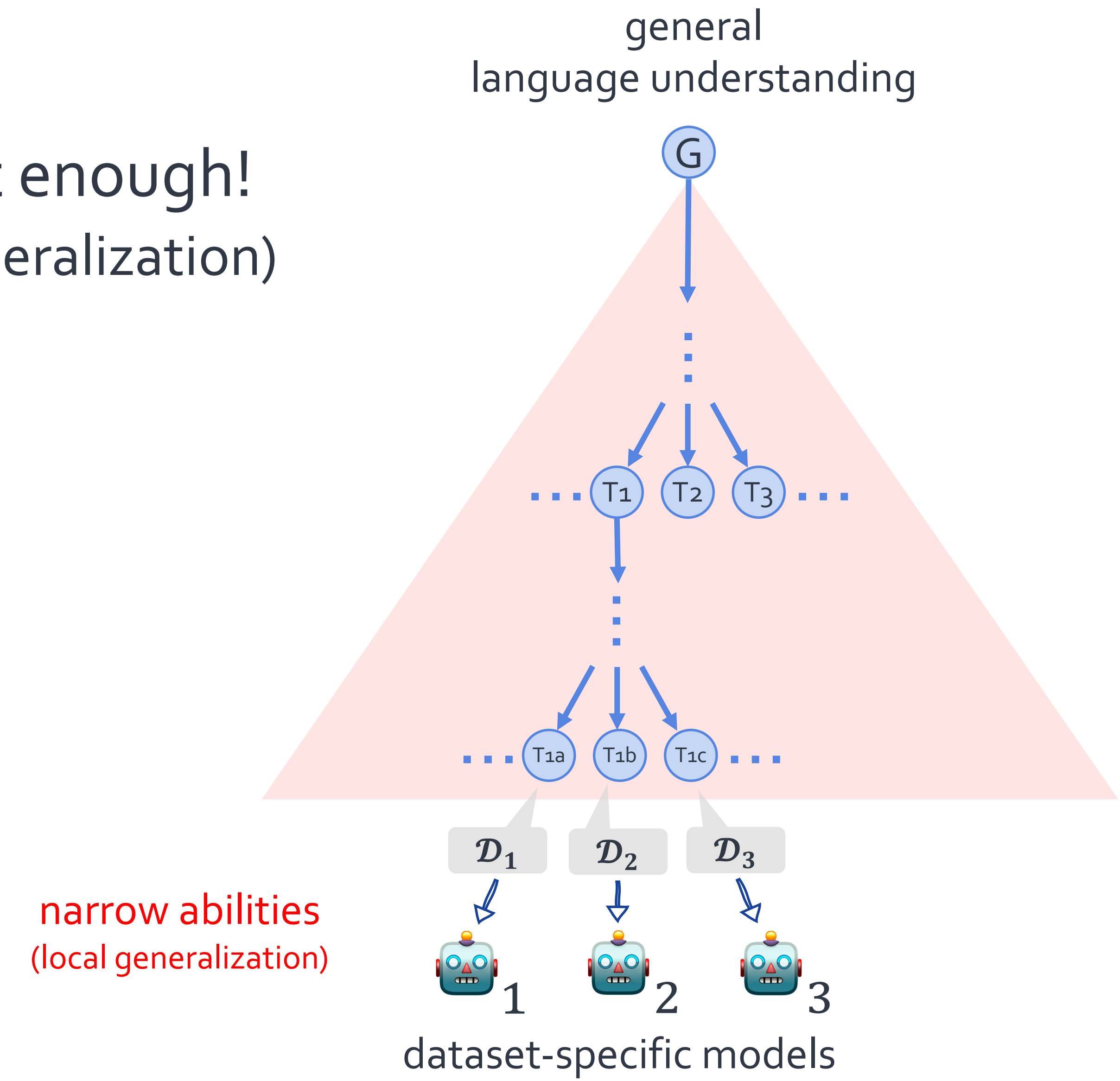


# Limits of Success at Dataset Level



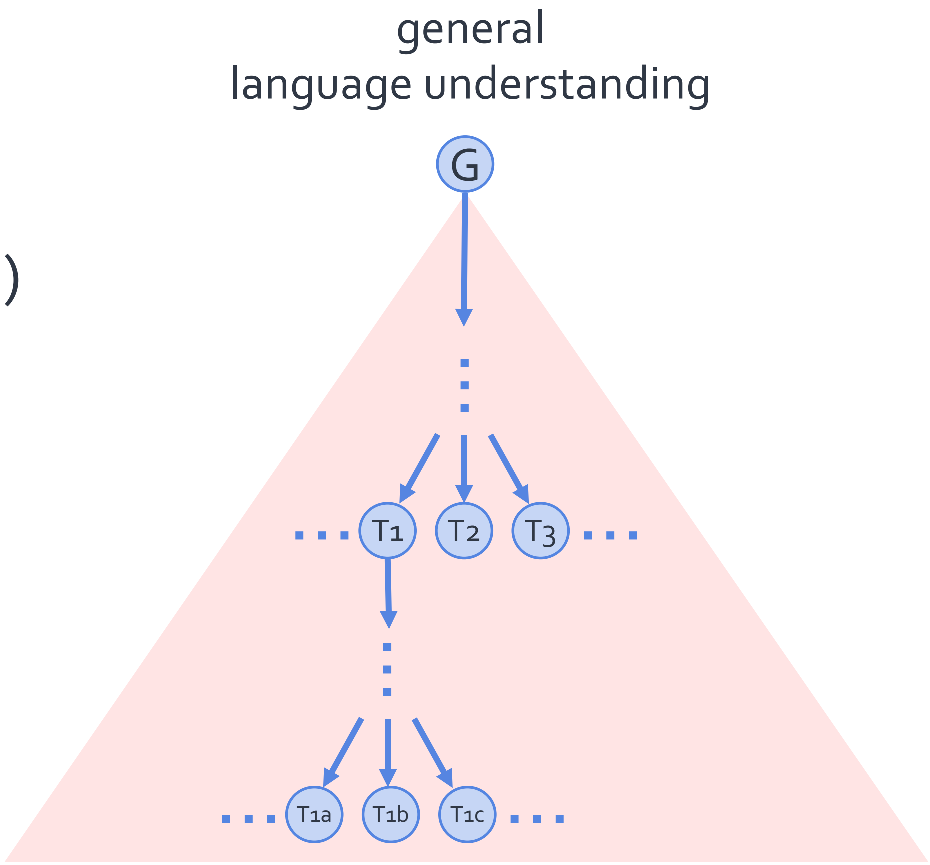
# Limits of Success at Dataset Level

- Success at dataset level is not enough!
  - Limited to the scope (local generalization)



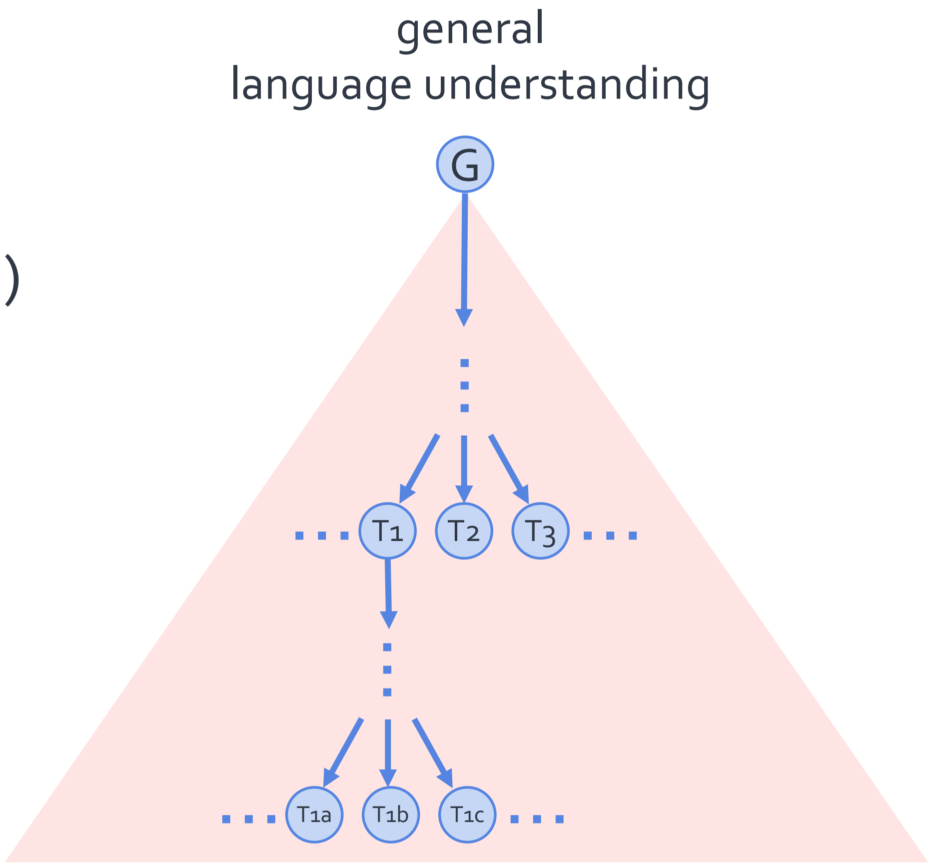
# Quest Toward "Generality"

- Success at dataset level is not enough!
    - Limited to the scope (local generalization)



# Quest Toward "Generality"

- Success at dataset level is not enough!
  - Limited to the scope (local generalization)
- "Generality" necessitates models that capture broader range of abilities.

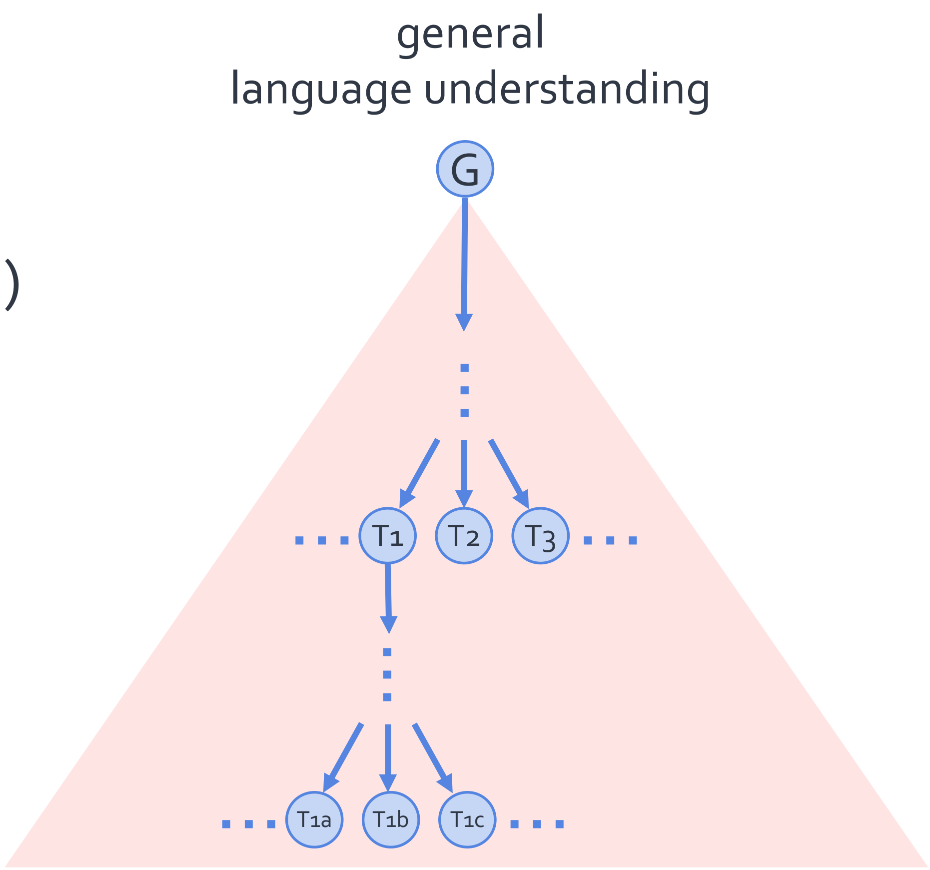


# Quest Toward “Generality”

- Success at dataset level is not enough!
  - Limited to the scope (local generalization)

- “Generality” necessitates models that capture broader range of abilities.

- Progress on:

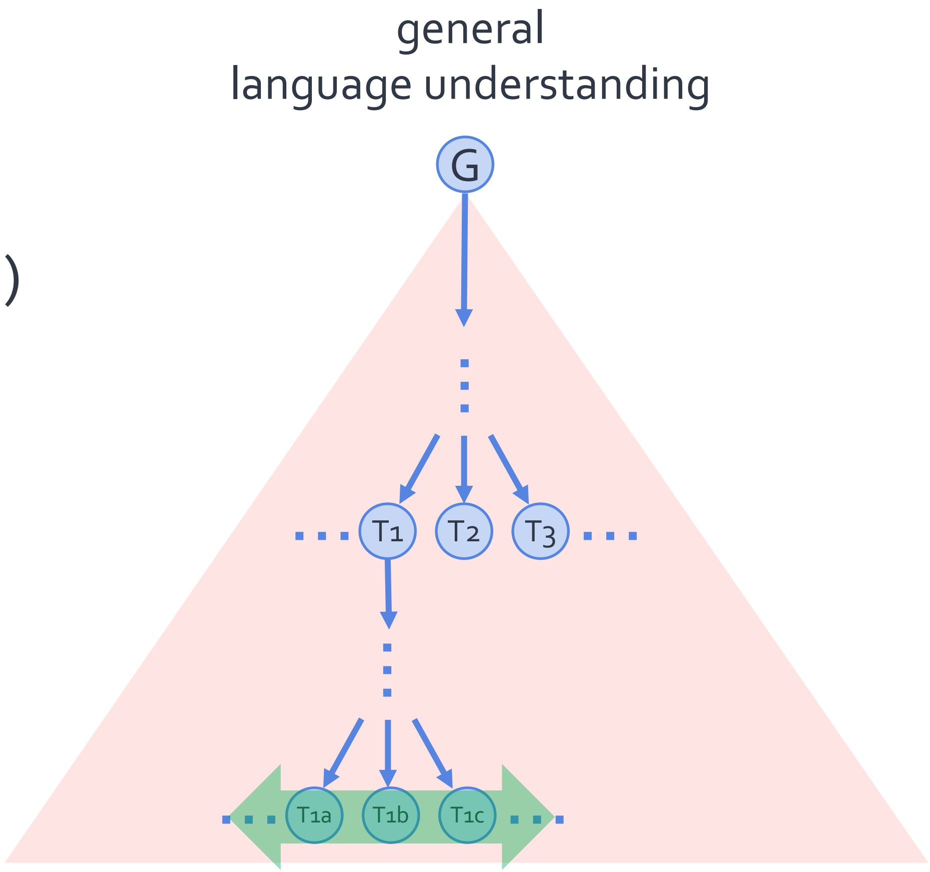


# Quest Toward "Generality"

- Success at dataset level is not enough!
  - Limited to the scope (local generalization)

- "Generality" necessitates models that capture broader range of abilities.

- Progress on:
  1. "breadth" — diverse abilities.

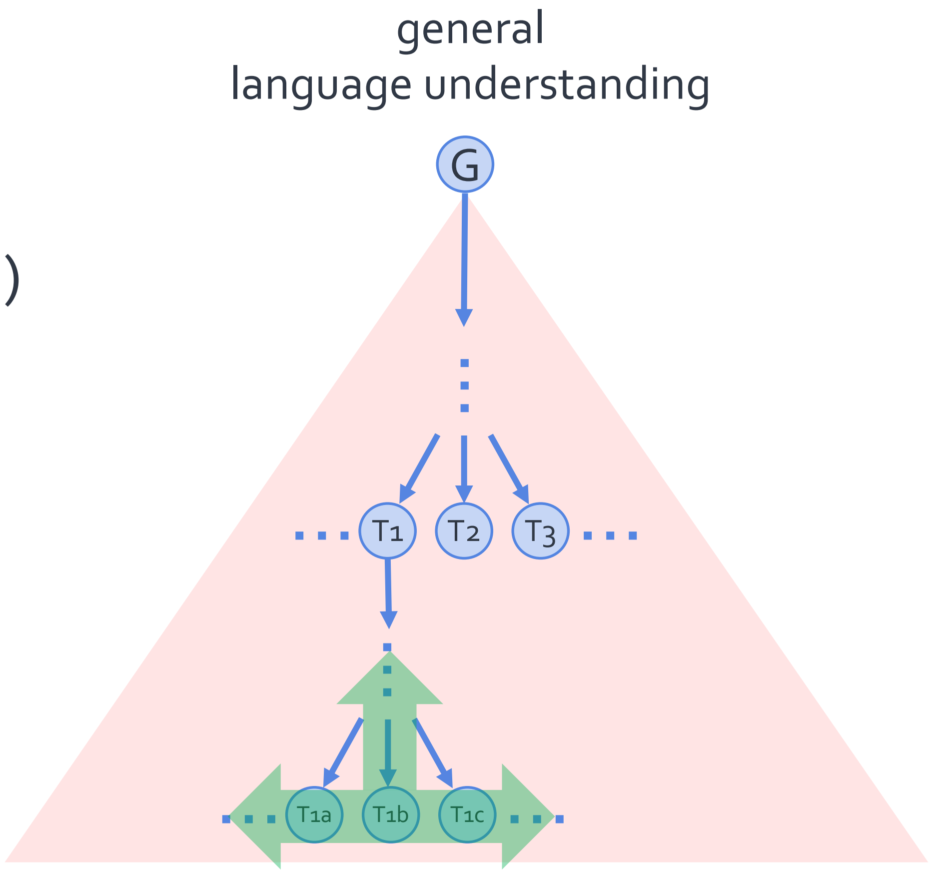


# Quest Toward "Generality"

- Success at dataset level is not enough!
  - Limited to the scope (local generalization)

- "Generality" necessitates models that capture broader range of abilities.

- Progress on:
  1. "breadth" — diverse abilities.
  2. "depth" — complex abilities.

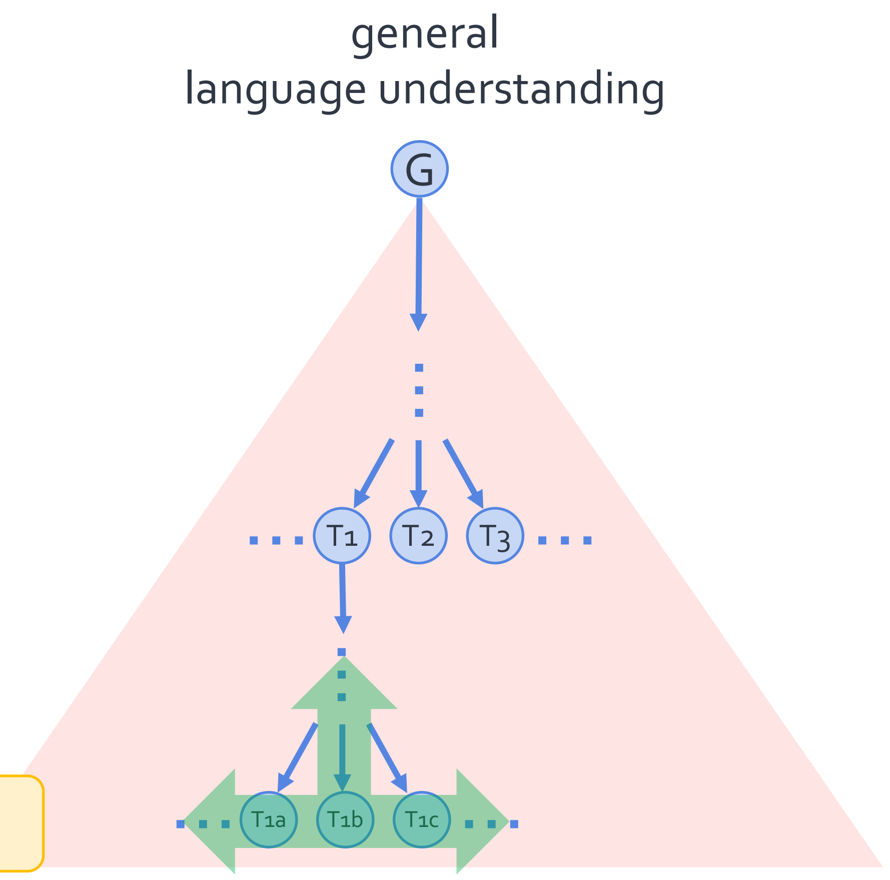


# Quest Toward "Generality"

- Success at dataset level is not enough!
  - Limited to the scope (local generalization)
- "Generality" necessitates models that capture broader range of abilities.
- Progress on:
  1. "breadth" — diverse abilities.
  2. "depth" — complex abilities.

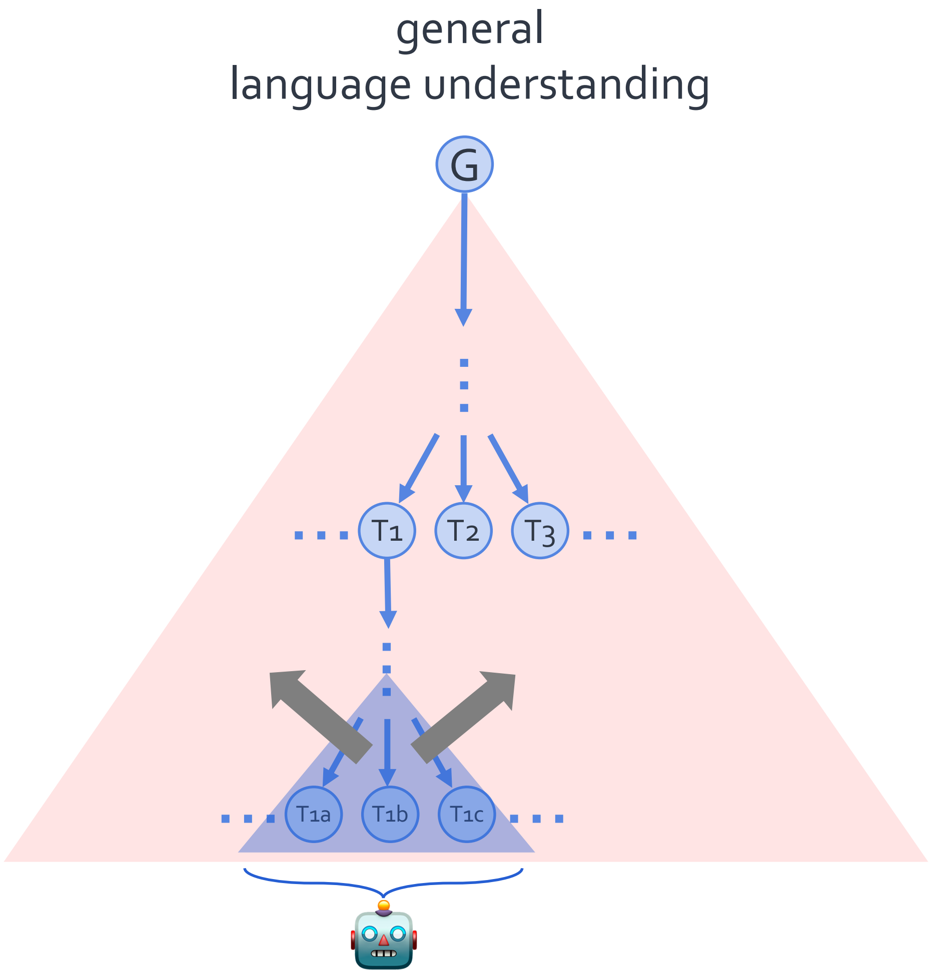
First part

Second part



# Research Goal

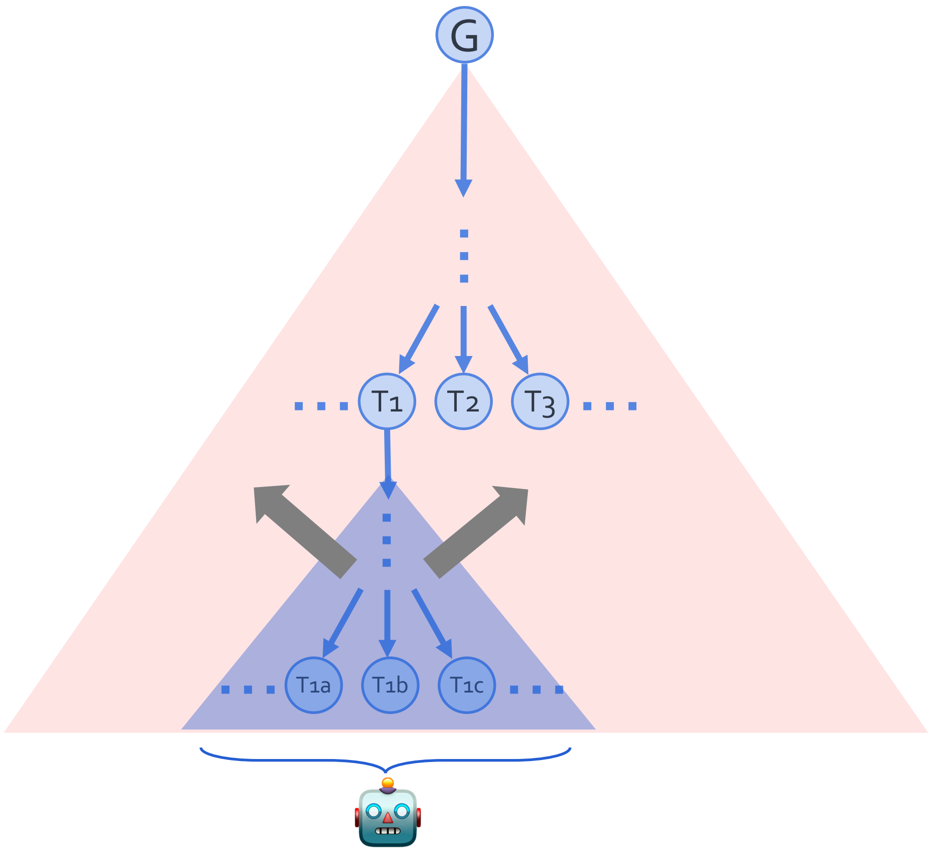
**Long-term goal:** more general natural language processing (NLP) systems through unified algorithms and theories.



# Research Goal

**Long-term goal:** more general natural language processing (NLP) systems through unified algorithms and theories.

general
language understanding



# Research Goal

**Long-term goal:** more general natural language processing (NLP) systems through unified algorithms and theories.

# Research Goal

**Long-term goal:** more general natural language processing (NLP) systems through unified algorithms and theories.

**Why?** AI-driven language interfaces that increasingly integrate in our life need to be versatile.



**Long-term goal:** more general natural language processing (NLP) systems through unified algorithms and theories.

**Long-term goal:** more general natural language processing (NLP) systems through unified algorithms and theories.

Generalization in "breadth"

Natural Instructions
*ACL '22*

UnifiedQA
*EMNLP Findings '20*

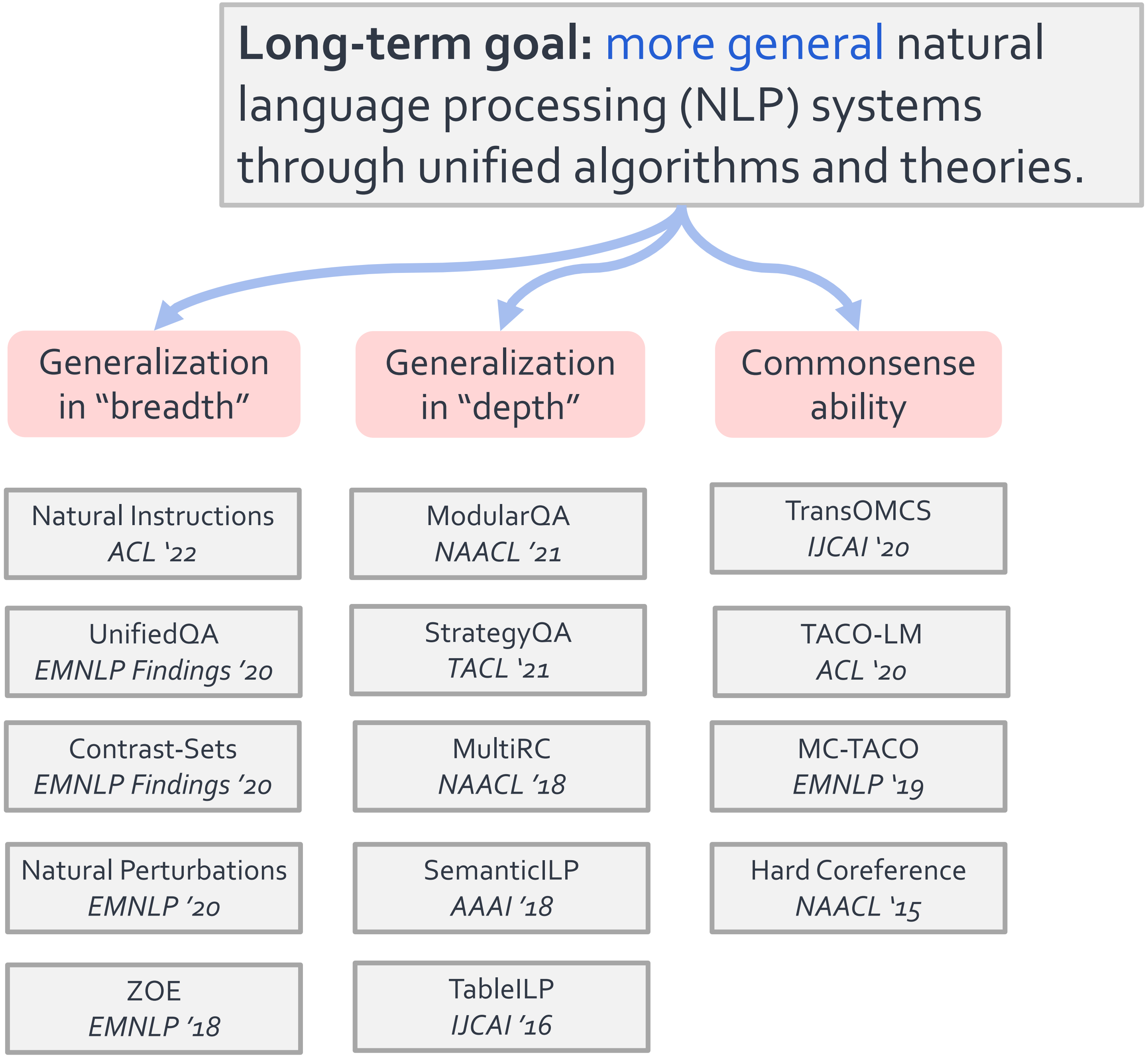
Contrast-Sets
*EMNLP Findings '20*

Natural Perturbations
*EMNLP '20*

ZOE
*EMNLP '18*

**Long-term goal:** more general natural language processing (NLP) systems through unified algorithms and theories.

- Generalization in “breadth”
  - Natural Instructions *ACL ‘22*
  - UnifiedQA *EMNLP Findings ’20*
  - Contrast-Sets *EMNLP Findings ’20*
  - Natural Perturbations *EMNLP ’20*
  - ZOE *EMNLP ’18*
- Generalization in “depth”
  - ModularQA *NAACL ’21*
  - StrategyQA *TACL ‘21*
  - MultiRC *NAACL ’18*
  - SemanticILP *AAAI ’18*
  - TableILP *IJCAI ’16*


**Long-term goal:** more general natural language processing (NLP) systems through unified algorithms and theories.
Generalization in "breadth"
Generalization in "depth"
Commonsense ability
Natural Instructions *ACL '22*
UnifiedQA *EMNLP Findings '20*
Contrast-Sets *EMNLP Findings '20*
Natural Perturbations *EMNLP '20*
ZOE *EMNLP '18*
ModularQA *NAACL '21*
StrategyQA *TACL '21*
MultiRC *NAACL '18*
SemanticILP *AAAI '18*
TableILP *IJCAI '16*
TransOMCS *IJCAI '20*
TACO-LM *ACL '20*
MC-TACO *EMNLP '19*
Hard Coreference *NAACL '15*

**Long-term goal:** more general natural language processing (NLP) systems through unified algorithms and theories.

| Generalization in "breadth" | Generalization in "depth" | Commonsense ability | Alignment with societal values |
|---|---|---|---|
| Natural Instructions *ACL '22* | ModularQA *NAACL '21* | TransOMCS *IJCAI '20* | Ethical Advice Taker *ACL Findings '21* |
| UnifiedQA *EMNLP Findings '20* | StrategyQA *TACL '21* | TACO-LM *ACL '20* | UnQover *EMNLP Findings '20* |
| Contrast-Sets *EMNLP Findings '20* | MultiRC *NAACL '18* | MC-TACO *EMNLP '19* | Perspectrum *NAACL '19* |
| Natural Perturbations *EMNLP '20* | SemanticILP *AAAI '18* | Hard Coreference *NAACL '15* | |
| ZOE *EMNLP '18* | TableILP *IJCAI '16* | | |

**Long-term goal:** more general natural language processing (NLP) systems through unified algorithms and theories.

| Generalization in "breadth" | Generalization in "depth" | Commonsense ability | Alignment with societal values |
|---|---|---|---|
| Natural Instructions *ACL '22* | ModularQA *NAACL '21* | TransOMCS *IJCAI '20* | Ethical Advice Taker *ACL Findings '21* |
| UnifiedQA *EMNLP Findings '20* | StrategyQA *TACL '21* | TACO-LM *ACL '20* | UnQover *EMNLP Findings '20* |
| Contrast-Sets *EMNLP Findings '20* | MultiRC *NAACL '18* | MC-TACO *EMNLP '19* | Perspectrum *NAACL '19* |
| Natural Perturbations *EMNLP '20* | SemanticILP *AAAI '18* | Hard Coreference *NAACL '15* | |
| ZOE *EMNLP '18* | TableILP *IJCAI '16* | | |

**Diagnosis**

**Long-term goal:** more general natural language processing (NLP) systems through unified algorithms and theories.

| Generalization in "breadth" | Generalization in "depth" | Commonsense ability | Alignment with societal values |
|---|---|---|---|
| Natural Instructions *ACL '22* | ModularQA *NAACL '21* | TransOMCS *IJCAI '20* | Ethical Advice Taker *ACL Findings '21* |
| UnifiedQA *EMNLP Findings '20* | StrategyQA *TACL '21* | TACO-LM *ACL '20* | UnQover *EMNLP Findings '20* |
| Contrast-Sets *EMNLP Findings '20* | MultiRC *NAACL '18* | MC-TACO *EMNLP '19* | Perspectrum *NAACL '19* |
| Natural Perturbations *EMNLP '20* | SemanticILP *AAAI '18* | Hard Coreference *NAACL '15* | |
| ZOE *EMNLP '18* | TableILP *IJCAI '16* | | |

**Diagnosis**

**Problem Definition**

**Long-term goal:** more general natural language processing (NLP) systems through unified algorithms and theories.

| Generalization in "breadth" | Generalization in "depth" | Commonsense ability | Alignment with societal values |
|---|---|---|---|
| Natural Instructions<br>*ACL '22* | ModularQA<br>*NAACL '21* | TransOMCS<br>*IJCAI '20* | Ethical Advice Taker<br>*ACL Findings '21* |
| UnifiedQA<br>*EMNLP Findings '20* | StrategyQA<br>*TACL '21* | TACO-LM<br>*ACL '20* | UnQover<br>*EMNLP Findings '20* |
| Contrast-Sets<br>*EMNLP Findings '20* | MultiRC<br>*NAACL '18* | MC-TACO<br>*EMNLP '19* | Perspectrum<br>*NAACL '19* |
| Natural Perturbations<br>*EMNLP '20* | SemanticILP<br>*AAAI '18* | Hard Coreference<br>*NAACL '15* | |
| ZOE<br>*EMNLP '18* | TableILP<br>*IJCAI '16* | | |

**Diagnosis**

**Problem Definition**

**Solution**

**Long-term goal:** more general natural language processing (NLP) systems through unified algorithms and theories.

## Generalization in "breadth"

- Natural Instructions — *ACL '22*
- UnifiedQA — *EMNLP Findings '20*
- Contrast-Sets — *EMNLP Findings '20*
- Natural Perturbations — *EMNLP '20*
- ZOE — *EMNLP '18*

## Generalization in "depth"

- ModularQA — *NAACL '21*
- StrategyQA — *TACL '21*
- MultiRC — *NAACL '18*
- SemanticILP — *AAAI '18*
- TableILP — *IJCAI '16*

## Commonsense ability

- TransOMCS — *IJCAI '20*
- TACO-LM — *ACL '20*
- MC-TACO — *EMNLP '19*
- Hard Coreference — *NAACL '15*

## Alignment with societal values

- Ethical Advice Taker — *ACL Findings '21*
- UnQover — *EMNLP Findings '20*
- Perspectrum — *NAACL '19*

**Diagnosis**

**Problem Definition**

**Solution**

# Talk Outline

Generality in "breadth" — tackling a variety of tasks

Generality in "depth" — tackling more complex tasks

*Future work:* Toward broad, interactive reasoning

UnifiedQA
*EMNLP Findings '20*

Natural Instructions
*ACL '22*

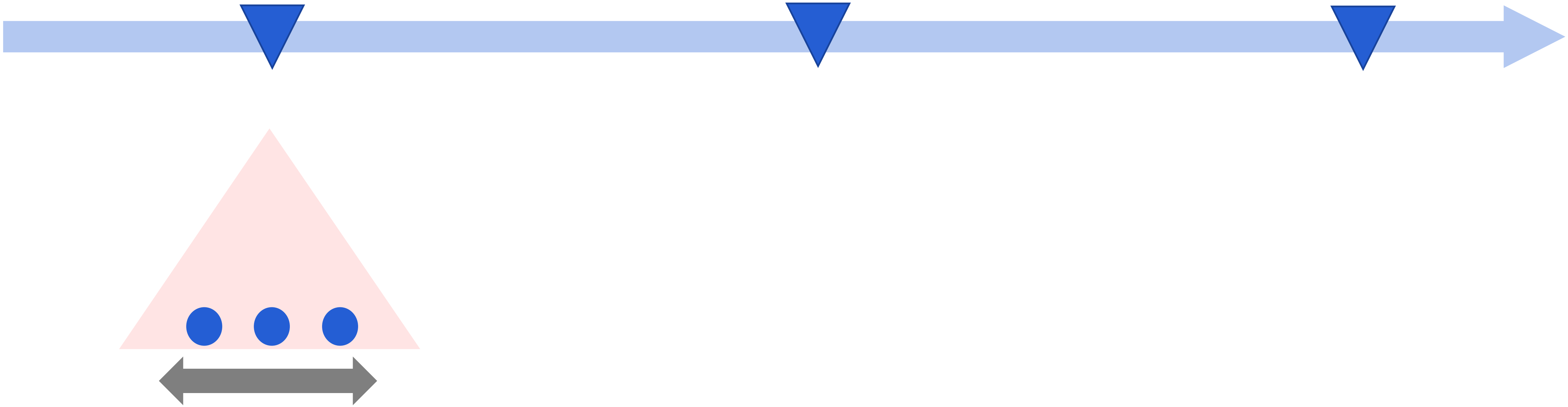
ModularQA
*NAACL '21*

# Talk Outline

Generality in “breadth” — tackling a variety of tasks

Generality in “depth” — tackling more complex tasks

*Future work:* Toward broad, interactive reasoning



UnifiedQA
*EMNLP Findings ’20*

Natural Instructions
*ACL ‘22*

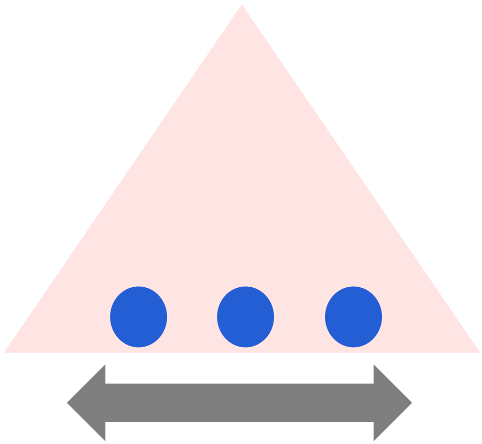
ModularQA
*NAACL ’21*

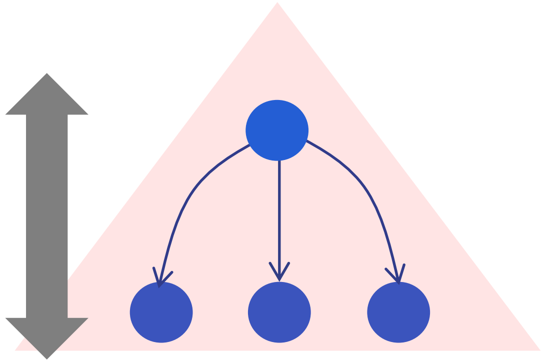
# Talk Outline

Generality in "breadth" — tackling a variety of tasks

Generality in "depth" — tackling more complex tasks

*Future work:*
Toward broad, interactive reasoning





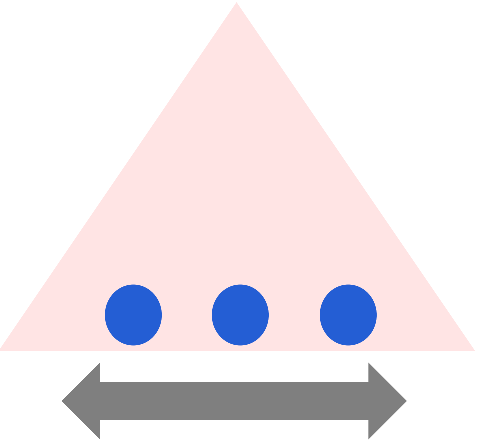
UnifiedQA
*EMNLP Findings '20*

Natural Instructions
*ACL '22*

ModularQA
*NAACL '21*

# Talk Outline

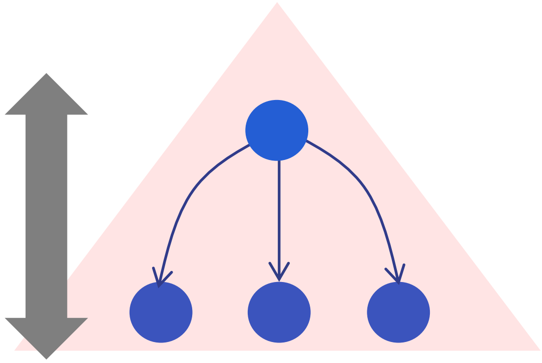
Generality in "breadth" — tackling a variety of tasks



UnifiedQA
*EMNLP Findings '20*

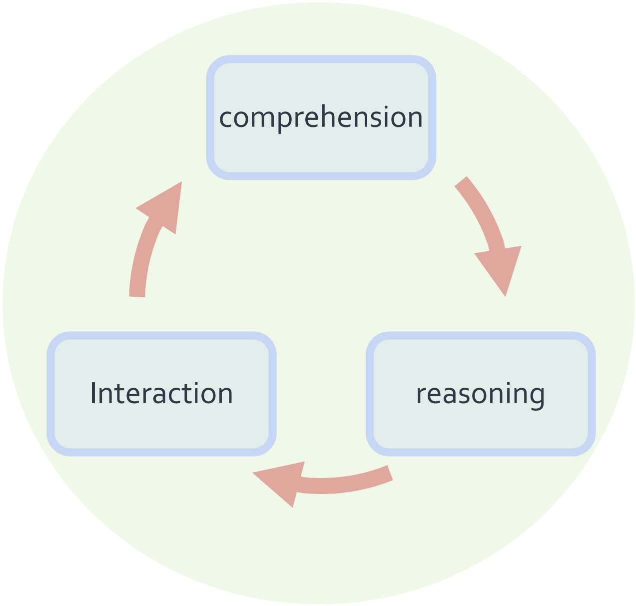
Natural Instructions
*ACL '22*

Generality in "depth" — tackling more complex tasks



ModularQA
*NAACL '21*

*Future work:*
Toward broad, interactive reasoning

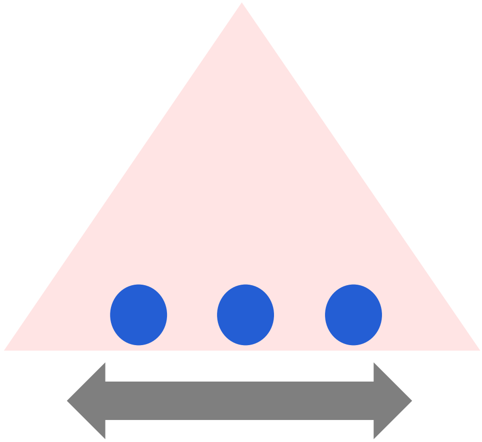


# Talk Outline

Generality in “breadth” — tackling a variety of tasks

Generality in “depth” — tackling more complex tasks

*Future work:* Toward broad, interactive reasoning

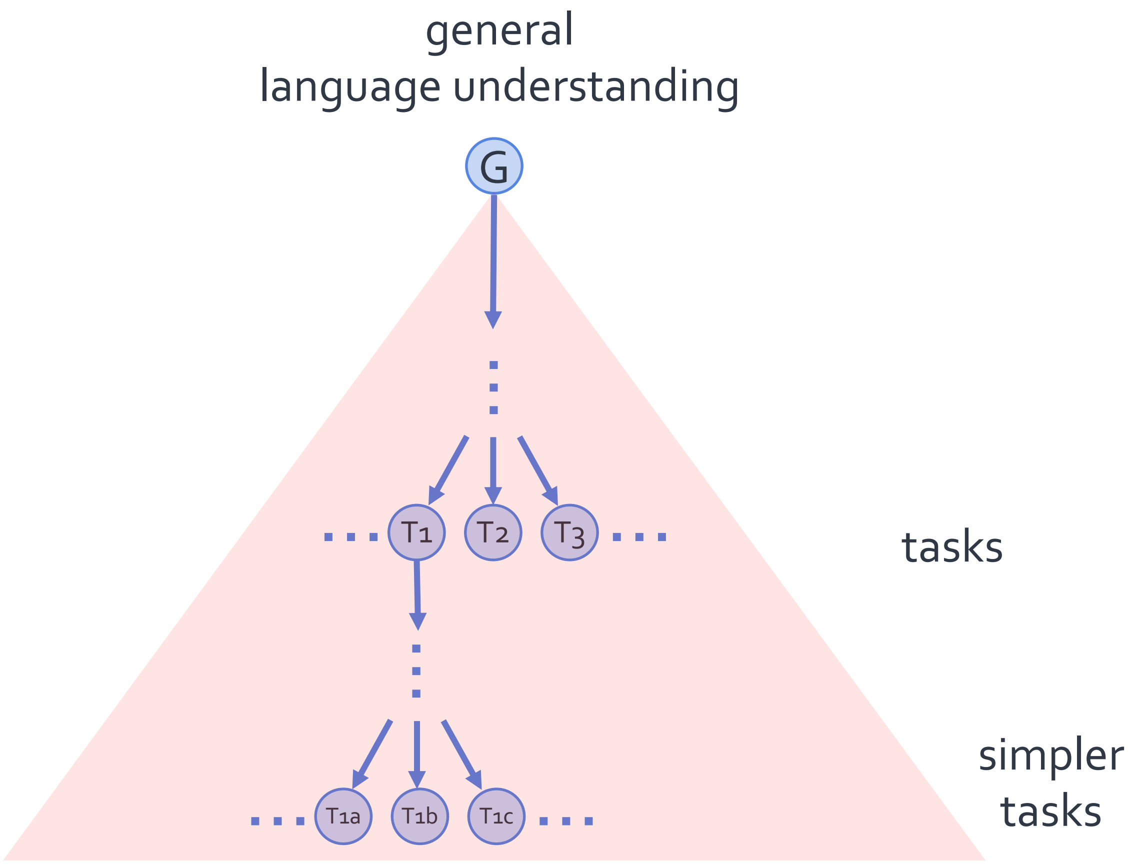


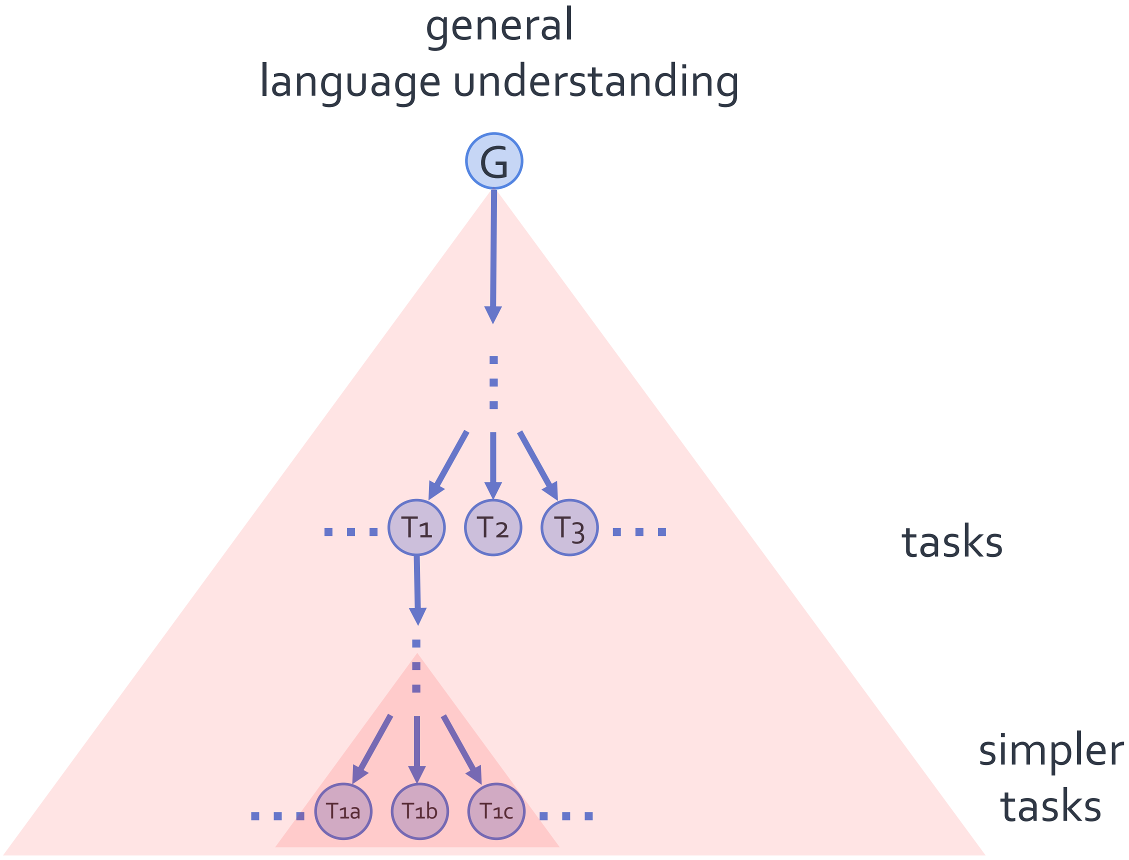
UnifiedQA
*EMNLP Findings ’20*

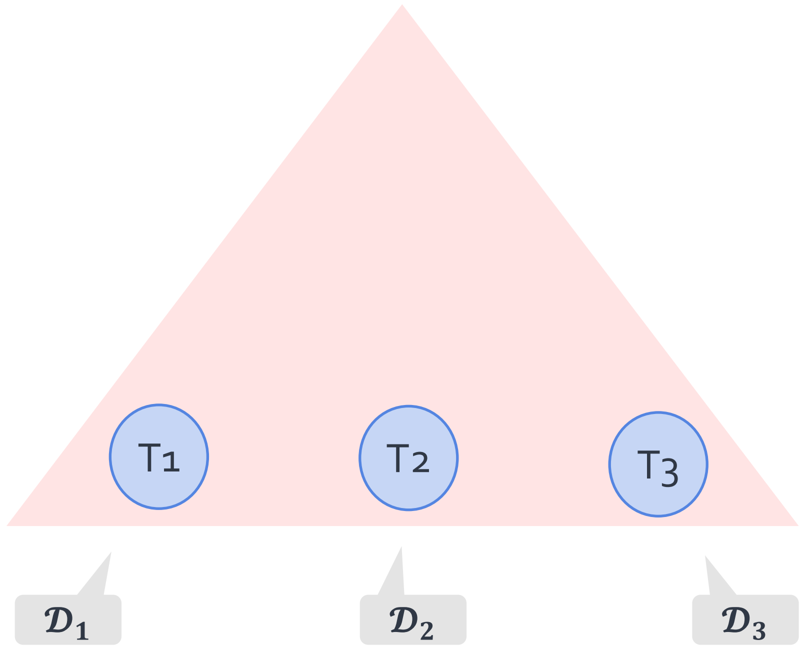
Natural Instructions
*ACL ‘22*

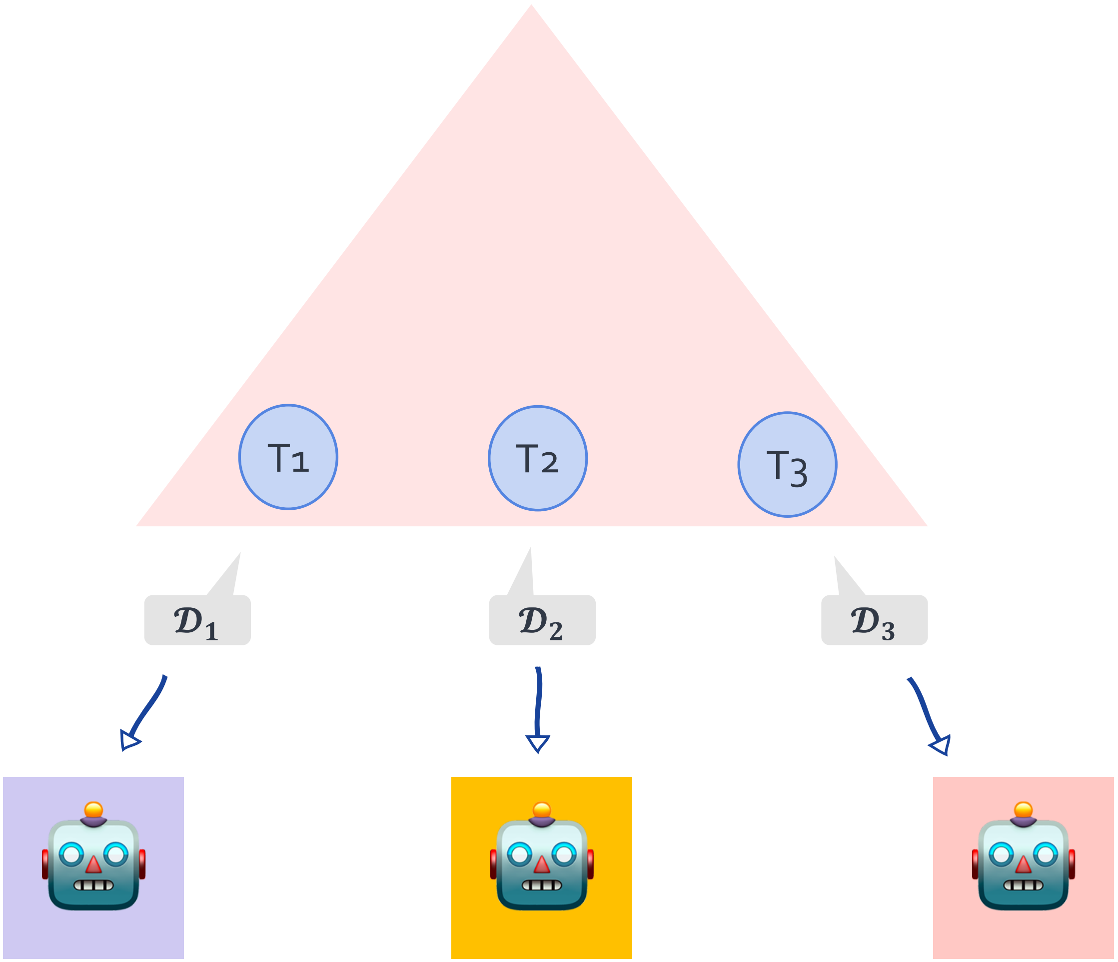


ModularQA
*NAACL ’21*

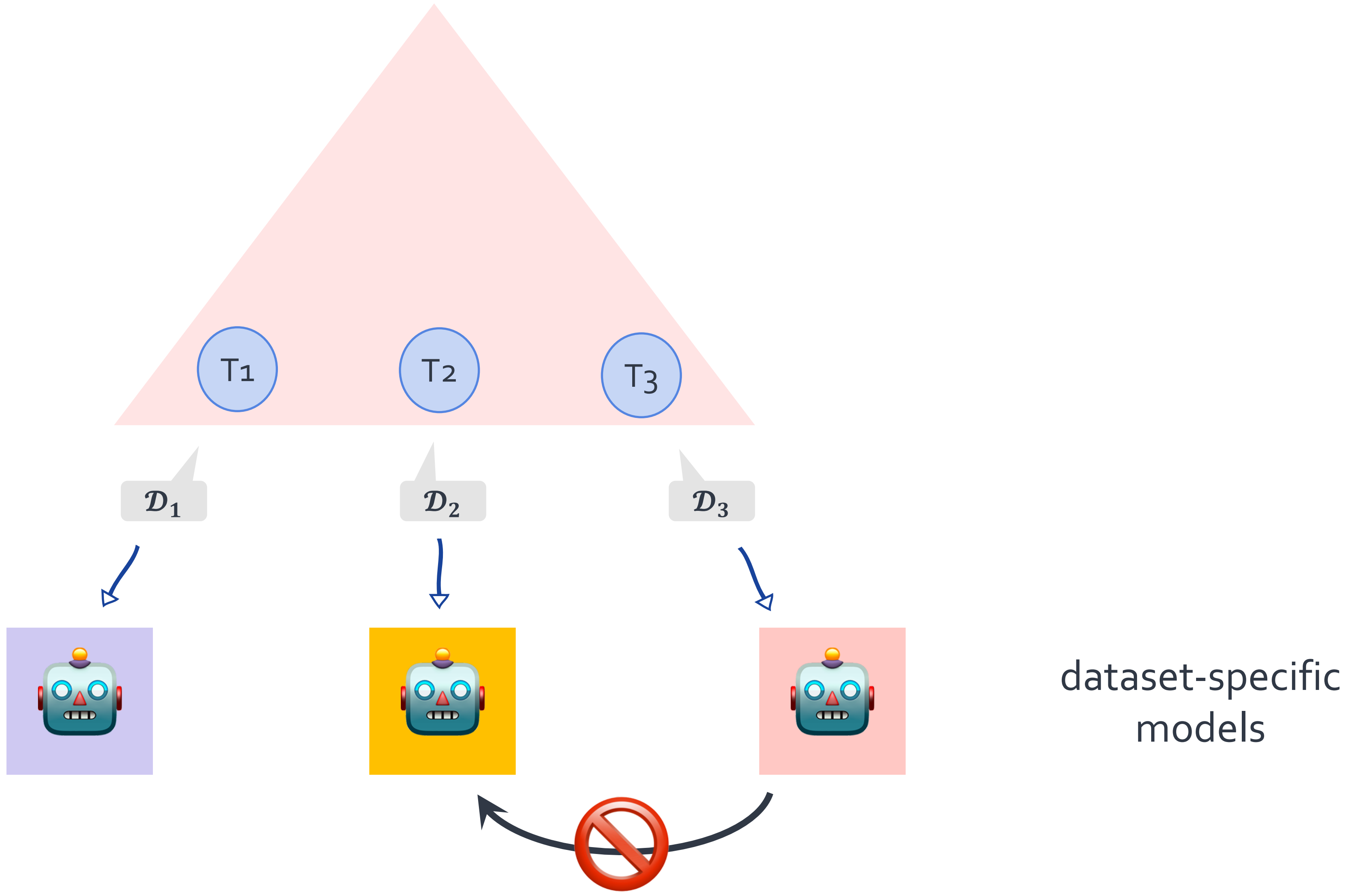

general
language understanding
G
T1
T2
T3
tasks
T1a
T1b
T1c
simpler
tasks


general
language understanding
G
T1
T2
T3
tasks
T1a
T1b
T1c
simpler
tasks

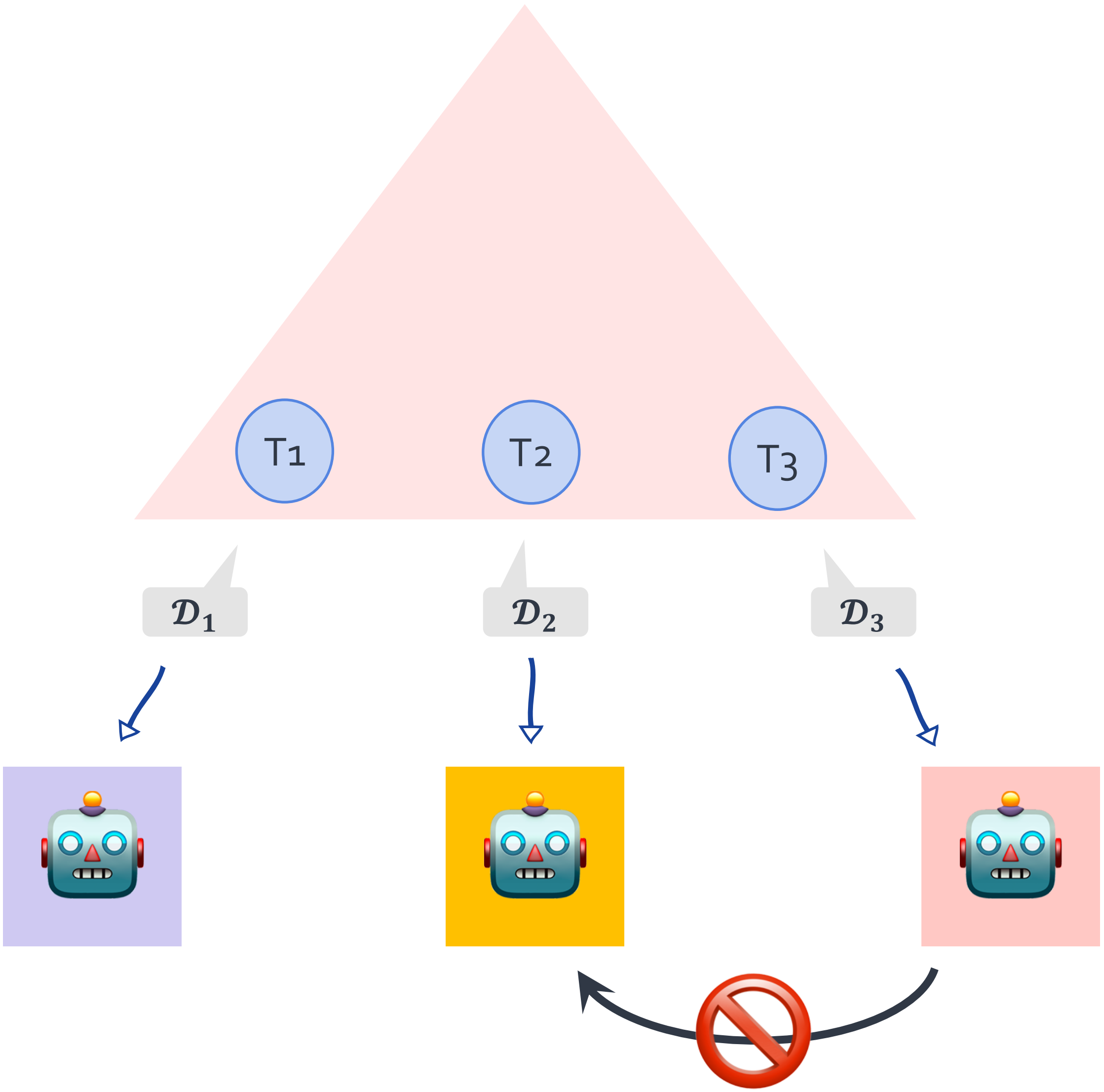

T1
T2
T3
$\mathcal{D}_1$
$\mathcal{D}_2$
$\mathcal{D}_3$




dataset-specific
models



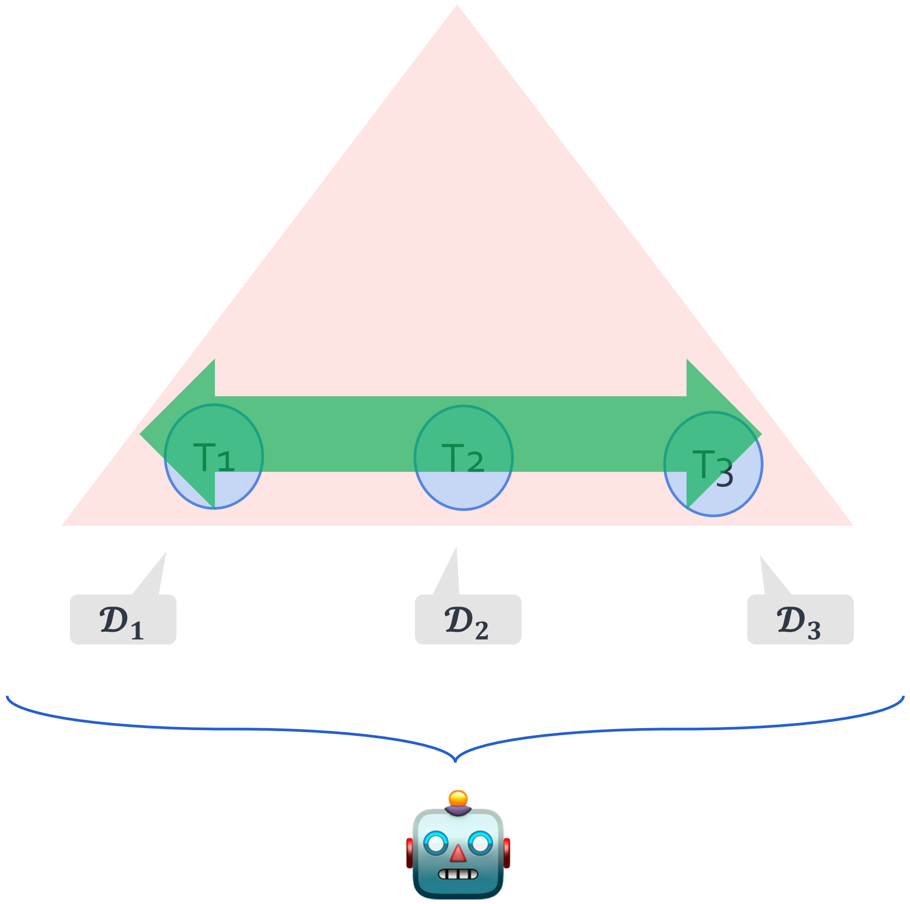

Task specific assumptions prevent generalization!



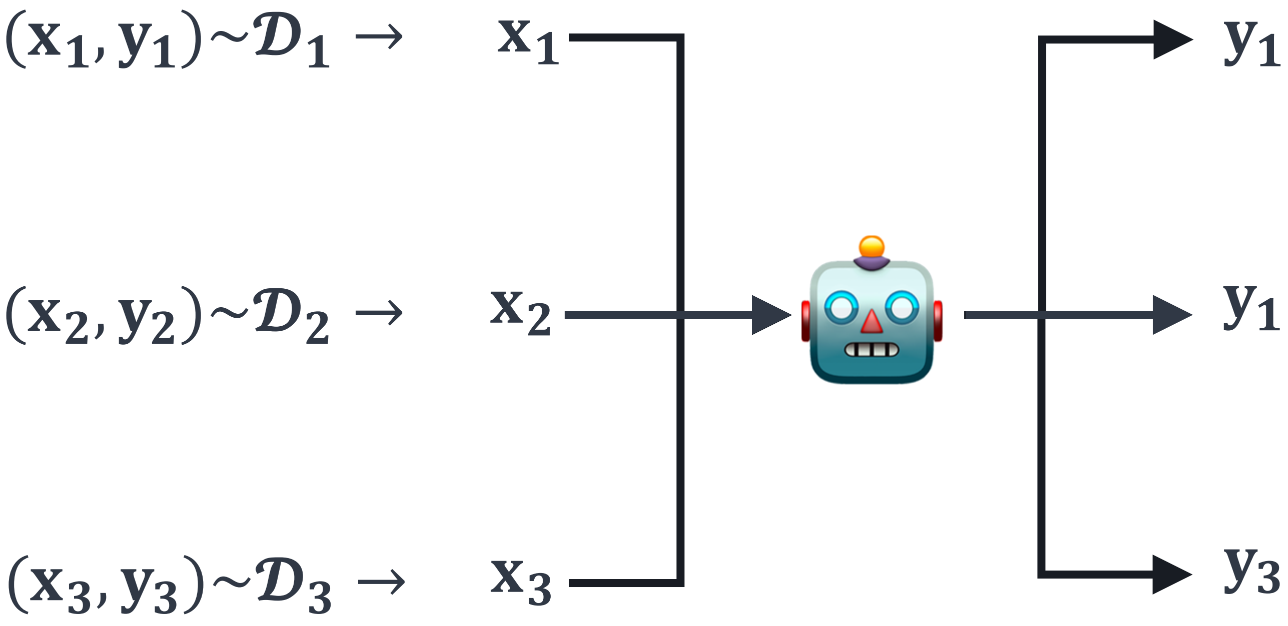

dataset-specific
models

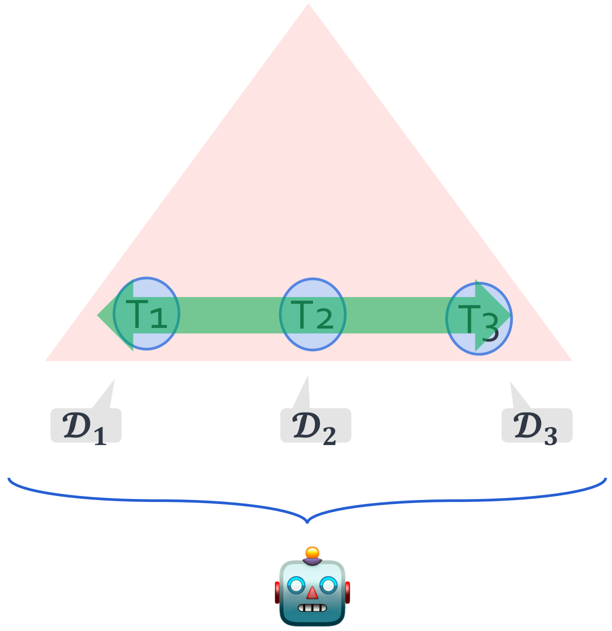
There are MANY tasks
— this is not scalable!

Task specific assumptions
prevent generalization!



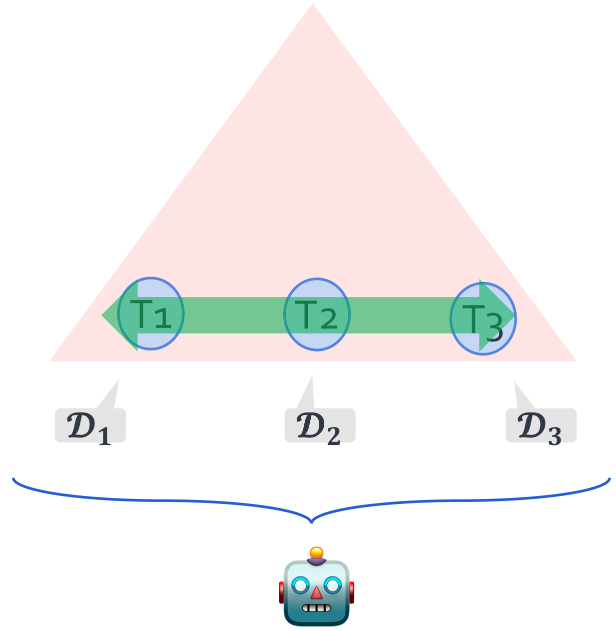

**Research questions:** How can we build a system that tackles a variety of language tasks?

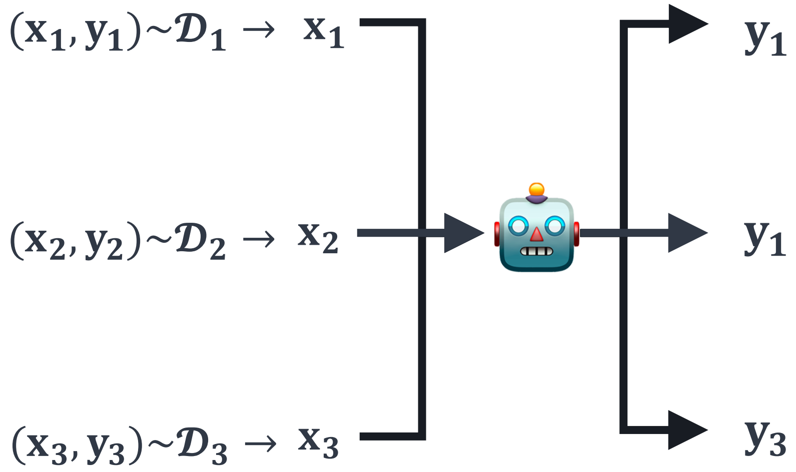



**Research questions:** How can we build a system that tackles a variety of language tasks?

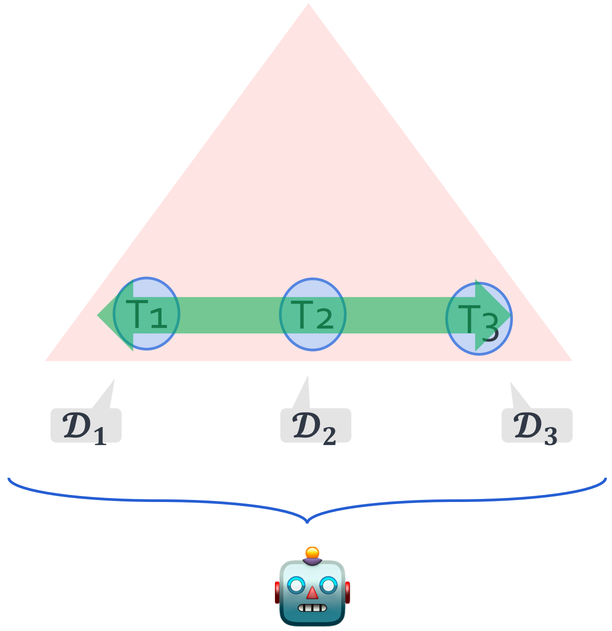
- Multi-task learning [Caruana '97; McCann et al. '18]

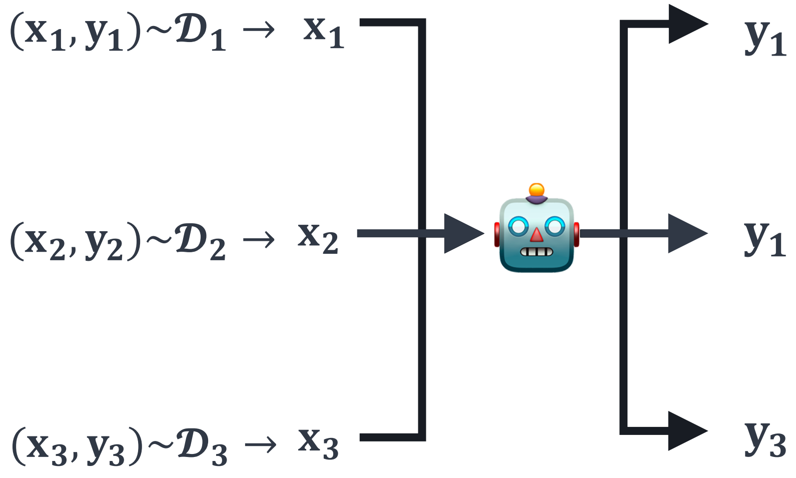


- Multi-task learning [Caruana '97; McCann et al. '18]

***Solving multiple learning tasks at the same time,***
*while exploiting commonalities across tasks.*

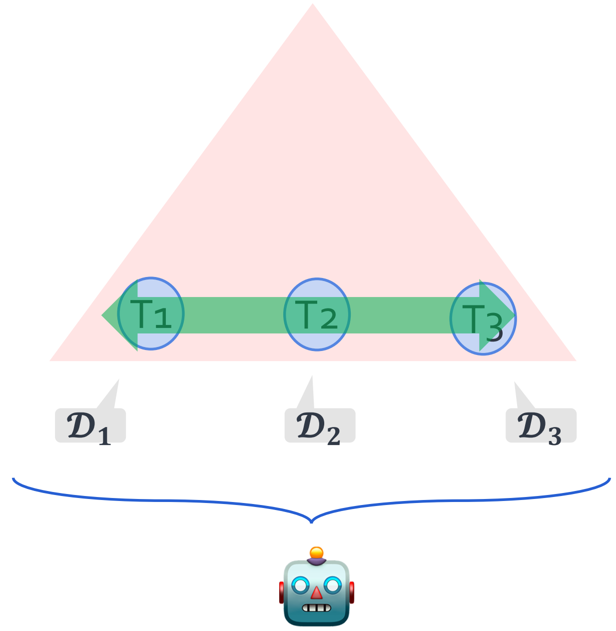


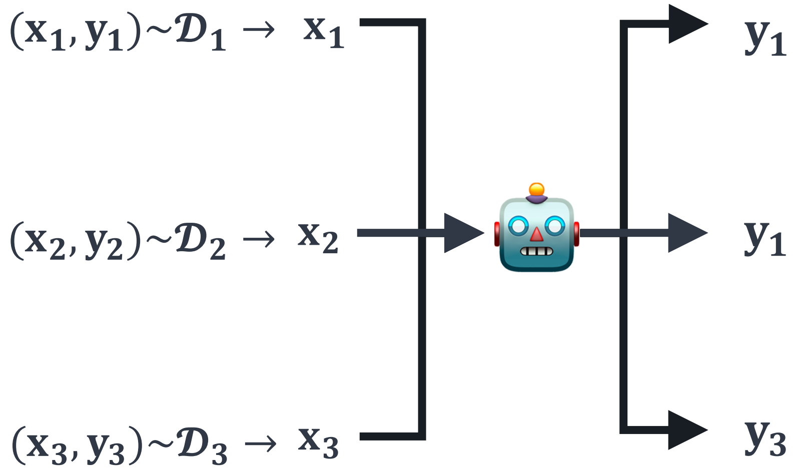



- Multi-task learning [Caruana '97; McCann et al. '18]

> ***Solving multiple learning tasks at the same time,***
> *while exploiting commonalities across tasks.*

- **Challenge:** negative transfer:
  - multi-tasking can hurt, if there is **not enough commonalities** among the tasks.

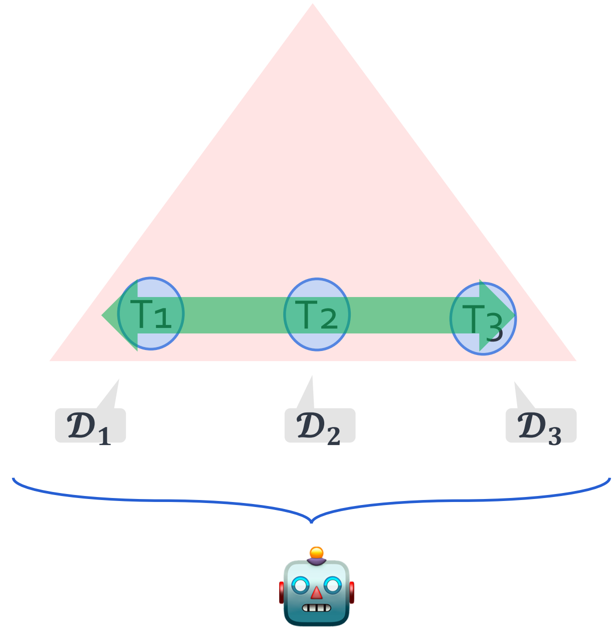


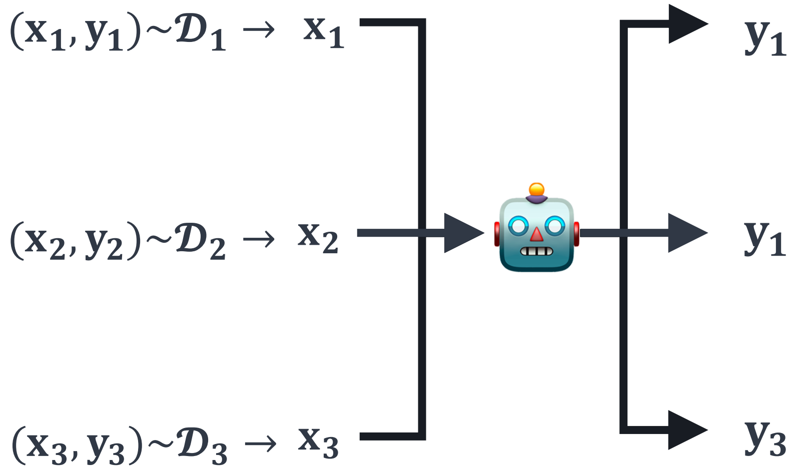



- Multi-task learning [Caruana '97; McCann et al. '18]

***Solving multiple learning tasks at the same time,***
*while exploiting* ***commonalities*** *across tasks.*

- **Challenge:** negative transfer:
  - multi-tasking can hurt, if there is **not enough commonalities** among the tasks.



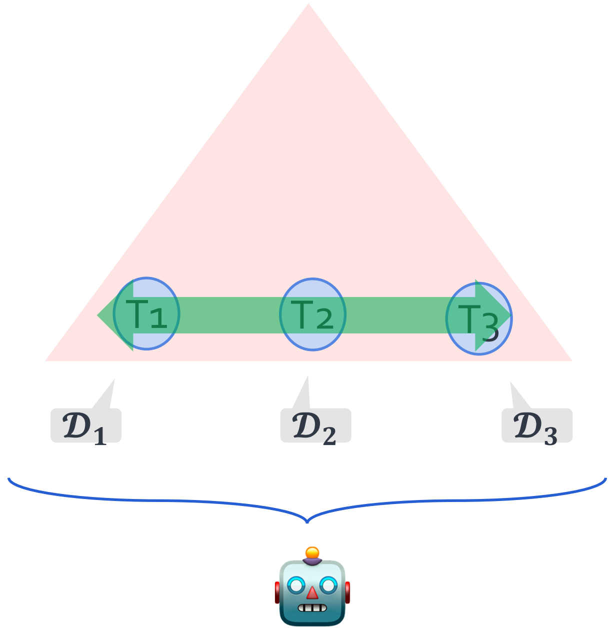



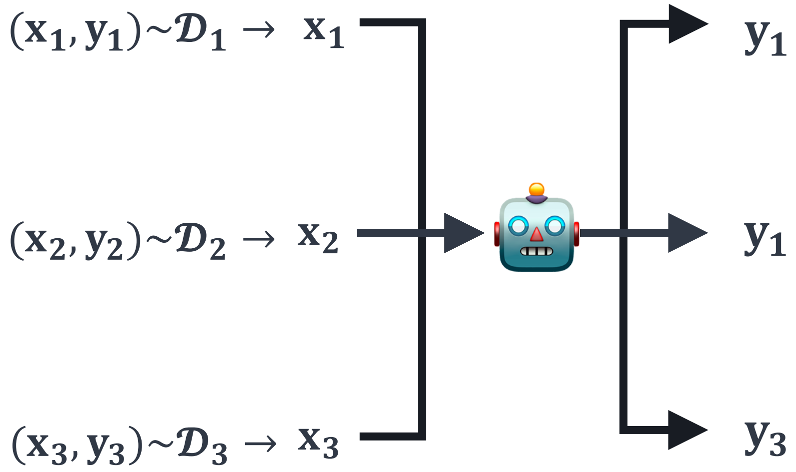
- Multi-task learning [Caruana '97; McCann et al. '18]

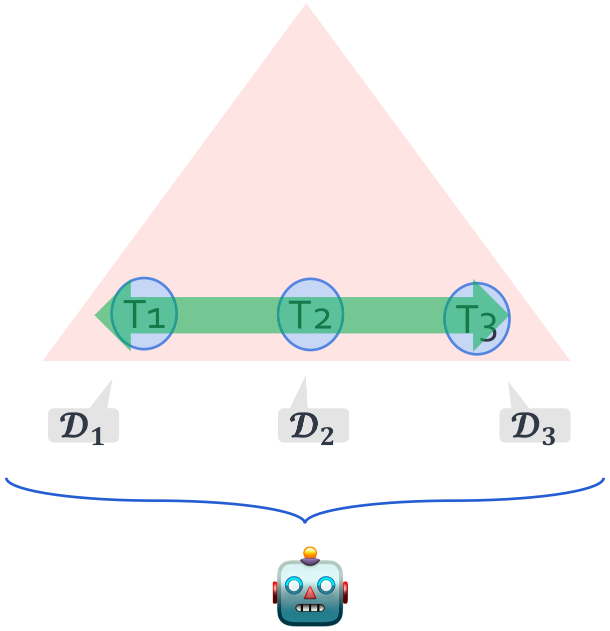
***Solving multiple learning tasks at the same time,***
*while exploiting **commonalities** across tasks.*

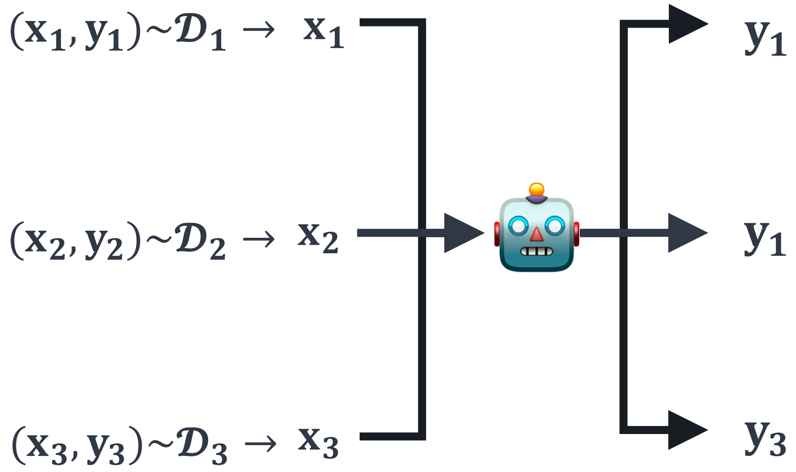
- **Challenge:** negative transfer:
  - multi-tasking can hurt, if there is **not enough commonalities** among the tasks.

**Multi-Task Learning in NLP:**
How can we leverage (or induce) commonalities among language tasks?





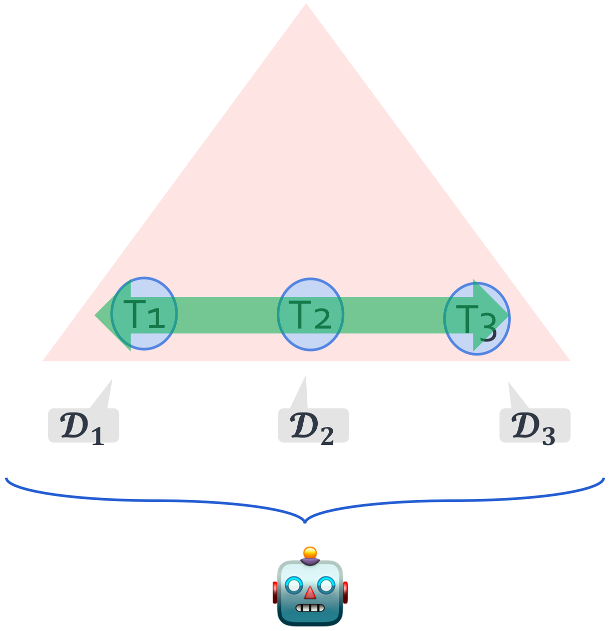



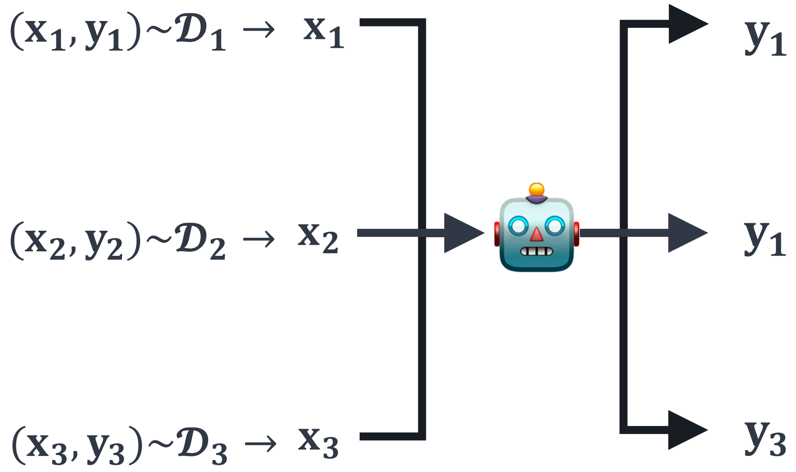

**Multi-Task Learning in NLP:**
How can we leverage (or induce) commonalities among language tasks?


Answering Questions

**Multi-Task Learning in NLP:**

How can we leverage (or induce) <u>commonalities</u> among language tasks?



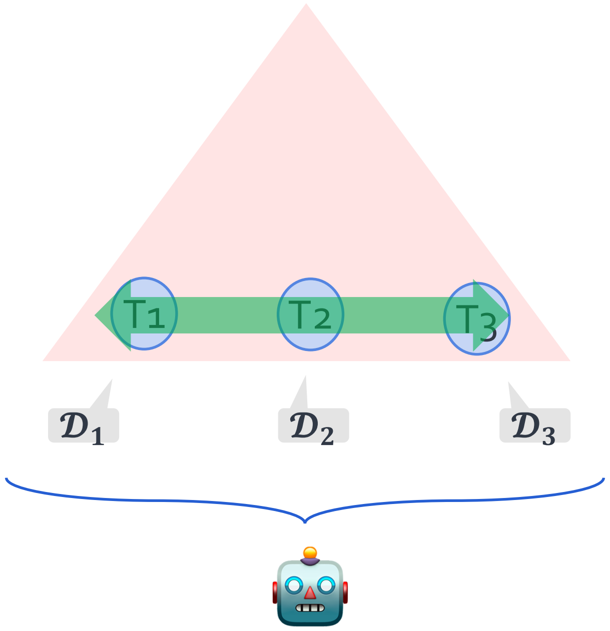



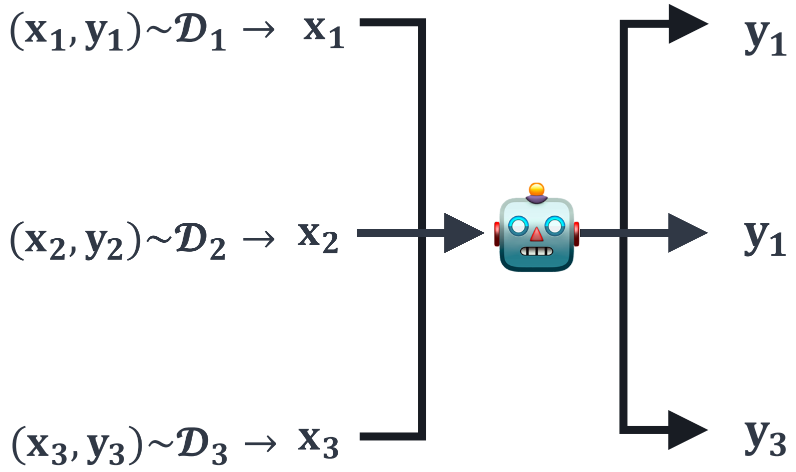

Answering Questions

Common NLP tasks

**Multi-Task Learning in NLP:**

How can we leverage (or induce) <u>commonalities</u> among language tasks?



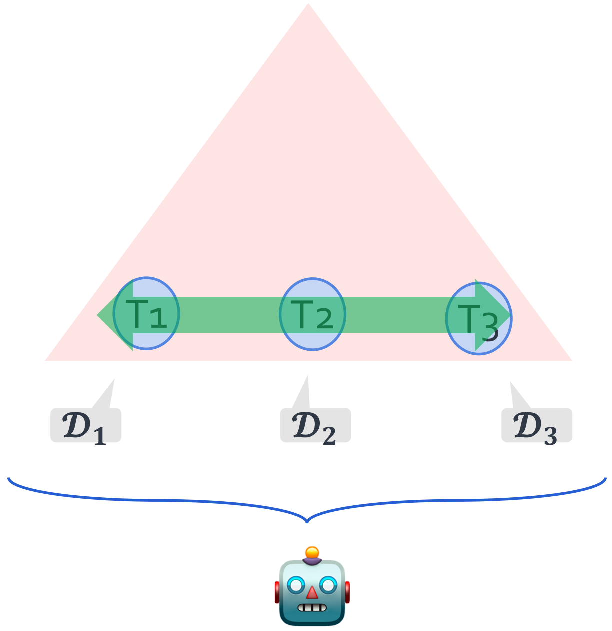



[Raffel et al. 2020]

**Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer**

**T5**

Google AI



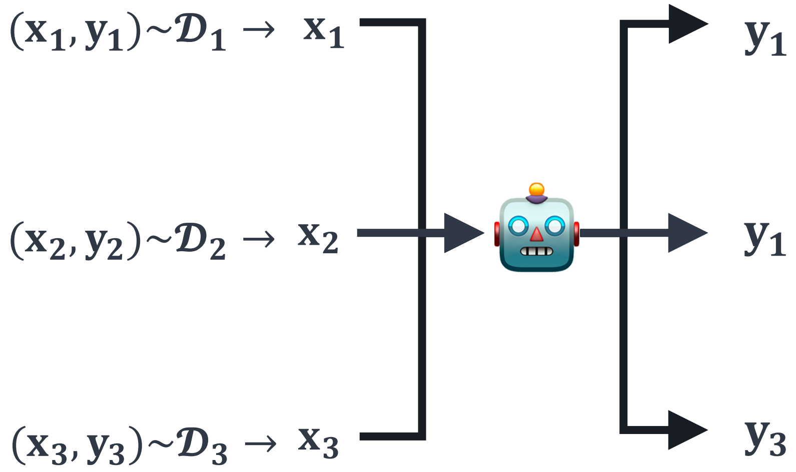


Answering Questions

Common NLP tasks

**Multi-Task Learning in NLP:**

How can we leverage (or induce) <u>commonalities</u> among language tasks?



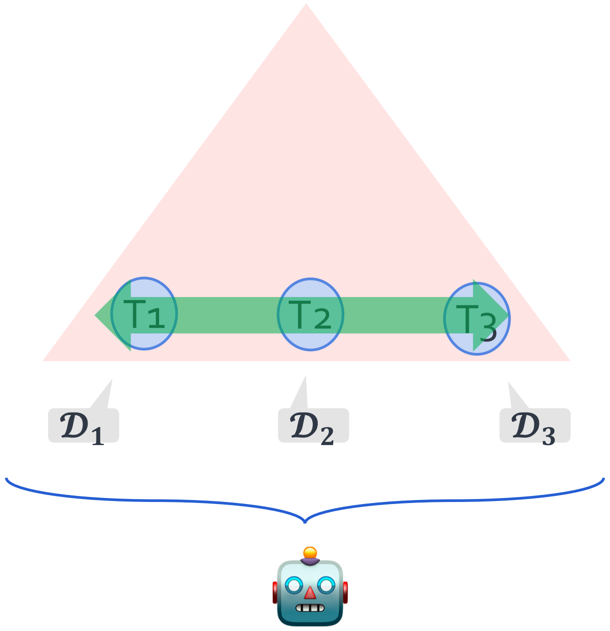
"…. *we find that* ***multi-task training*** ***underperforms*** ***fine-tuning on most tasks***"

[Raffel et al. 2020]

**Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer**

**T5**

Google AI

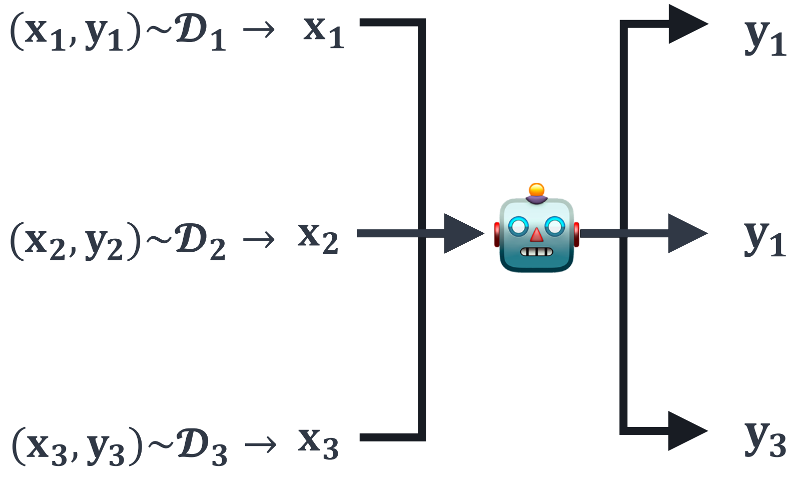




Answering Questions

Common NLP tasks

**Multi-Task Learning in NLP:**
How can we leverage (or induce) <u>commonalities</u> among language tasks?



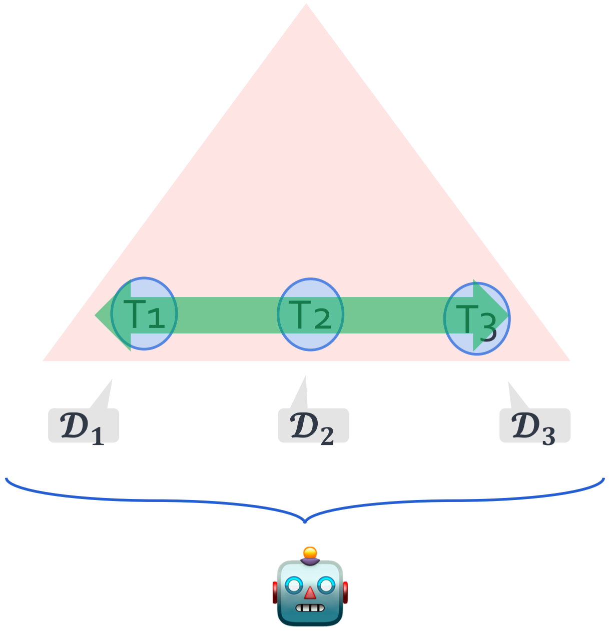
*"…. we find that* ***multi-task training*** ***underperforms*** ***fine-tuning on most tasks"***

[Raffel et al. 2020]

**Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer**

**T5**

Google AI



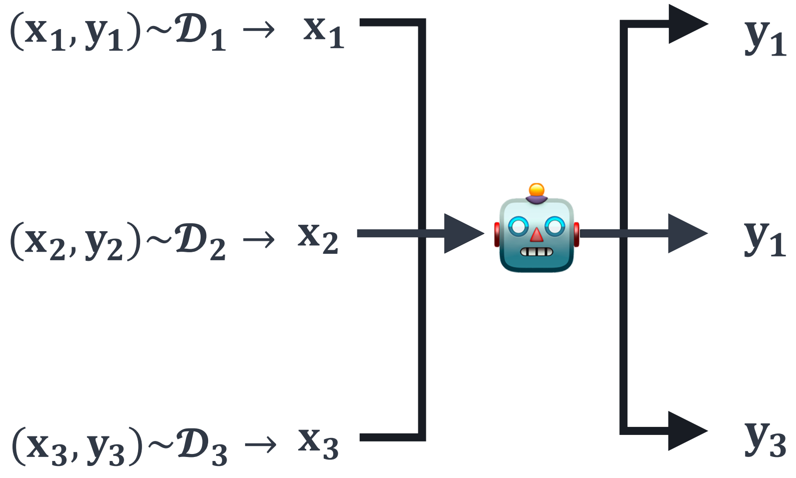


Answering Questions

Common NLP tasks

**Multi-Task Learning in NLP:**

How can we leverage (or induce) <u>commonalities</u> among language tasks?



# UnifiedQA

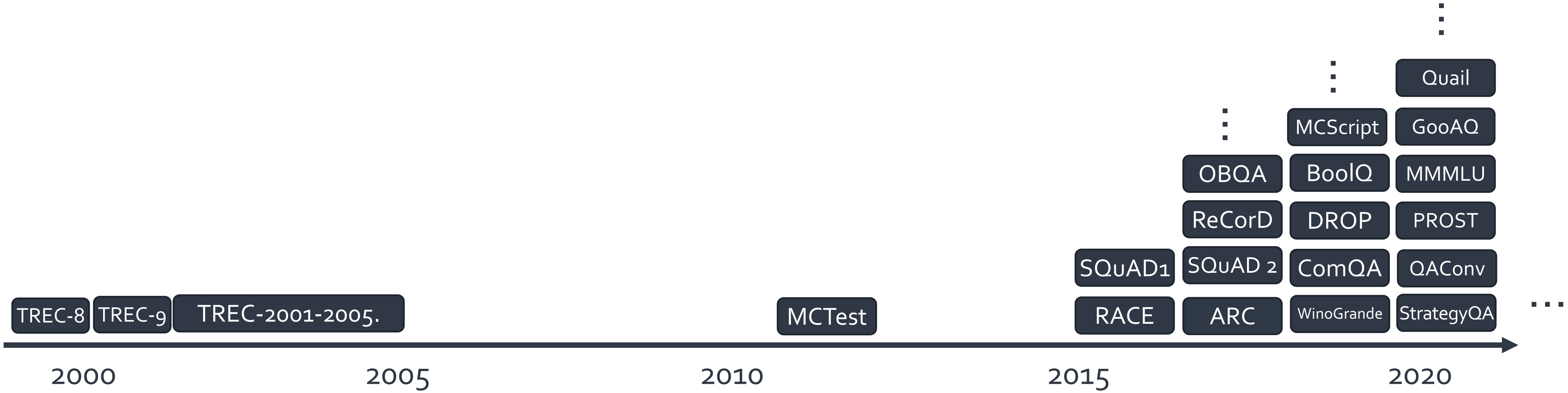
## Answering a broad range of questions with a single system

**Daniel Khashabi**, Sewon Min, Tushar Khot, Ashish Sabharwal
Oyvind Tafjord, Peter Clark and Hannaneh Hajishirzi

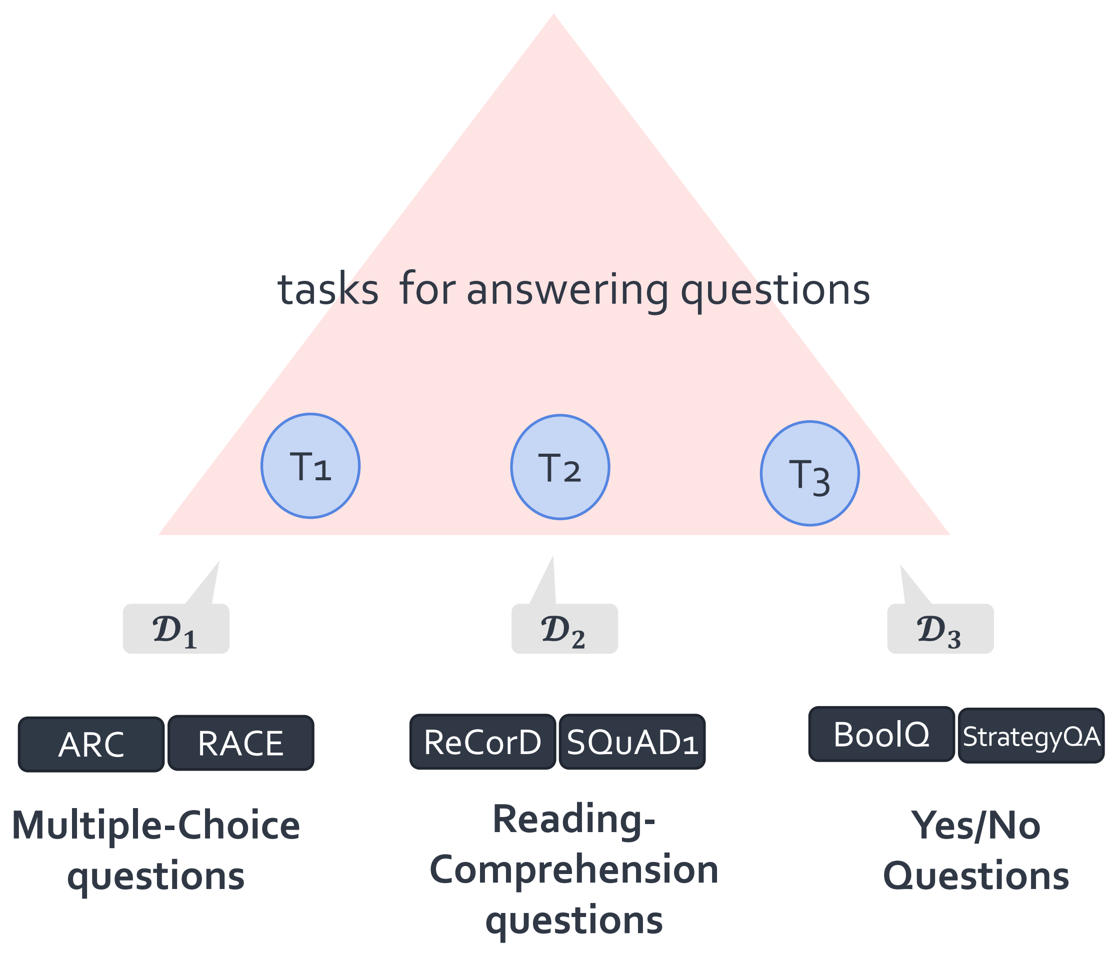
*EMNLP Findings 2020*

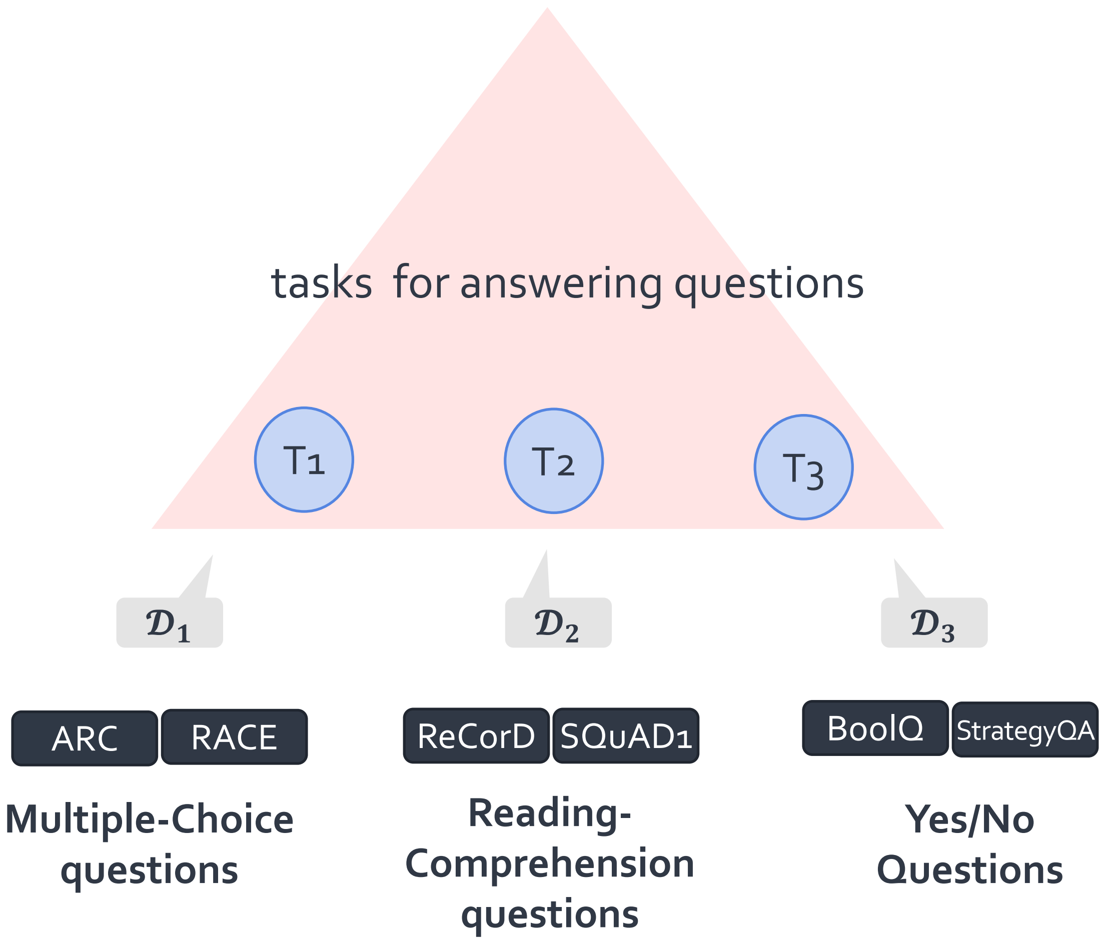
# Answering Questions: Sub-tasks



# Answering Questions: Sub-tasks



# Answering Questions: Sub-tasks



# Answering Questions: Sub-tasks

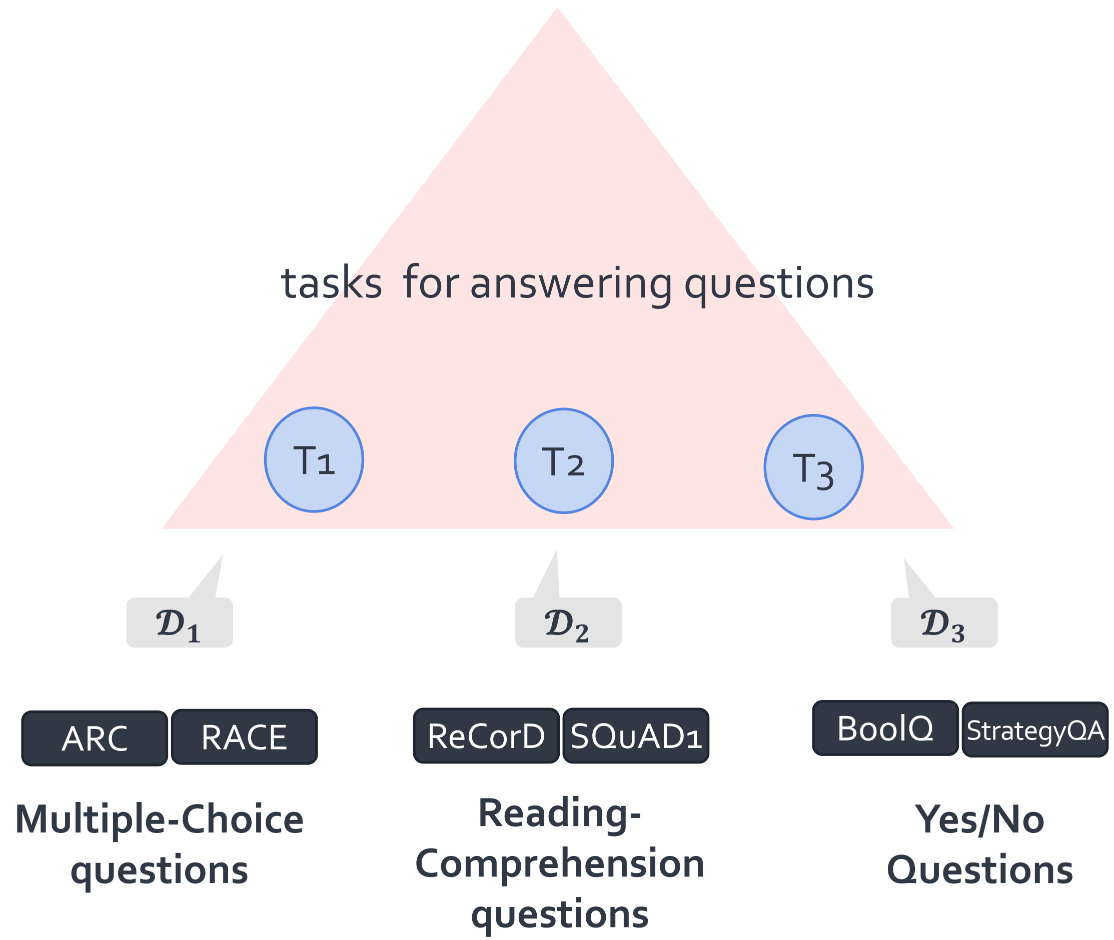
$\mathbf{x} =$ **Question:** *"What does photosynthesis produce that helps plants grow?"*

**Candidates:** *(A) water* *(B) oxygen* *(C) protein* *(D) sugar*

$\rightarrow$ $\mathbf{y} =$*"oxygen"*

multiple-choice

[Clark et al. 18]



# Answering Questions: Sub-tasks

$\mathbf{x} =$ **Question:** *"What does photosynthesis produce that helps plants grow?"*

**Candidates:** *(A) water* *(B) oxygen* *(C) protein* *(D) sugar*

$\rightarrow$ $\mathbf{y} =$ *"oxygen"*

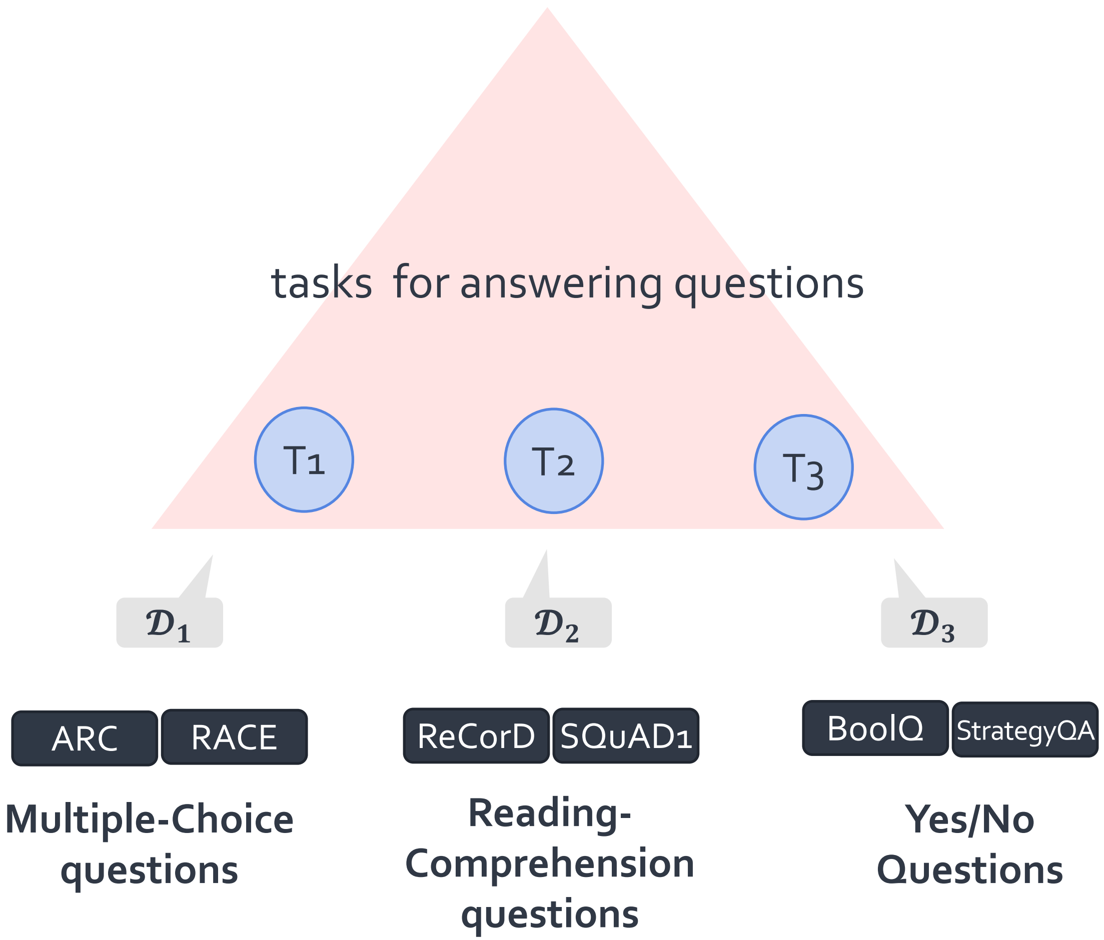
multiple-choice
[Clark et al. 18]

$\mathbf{x} =$ **Question:** *"At what speed did the turbine operate?"*

**Context:** *On his 50th birthday in 1906, Tesla demonstrated ....*

$\rightarrow$ $\mathbf{y} =$ *"16,000 rpm"*

reading-comprehension
[Rajpurkar et al. '16]



# Answering Questions: Sub-tasks

$\mathbf{x} =$ **Question:** "*What does photosynthesis produce that helps plants grow?*"

**Candidates:** *(A) water* *(B) oxygen* *(C) protein* *(D) sugar*

→ $\mathbf{y} =$"*oxygen*"

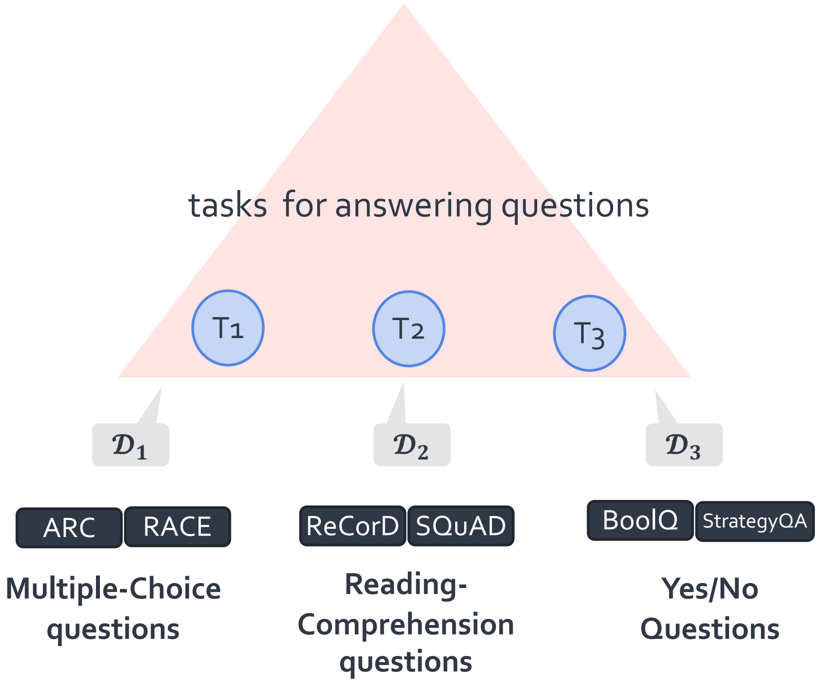
multiple-choice

$\mathbf{x} =$ **Question:** "*At what speed did the turbine operate?*"

**Context:** *On his 50th birthday in 1906, Tesla demonstrated ....*

→ $\mathbf{y} =$"*16,000 rpm*"

reading-comprehension



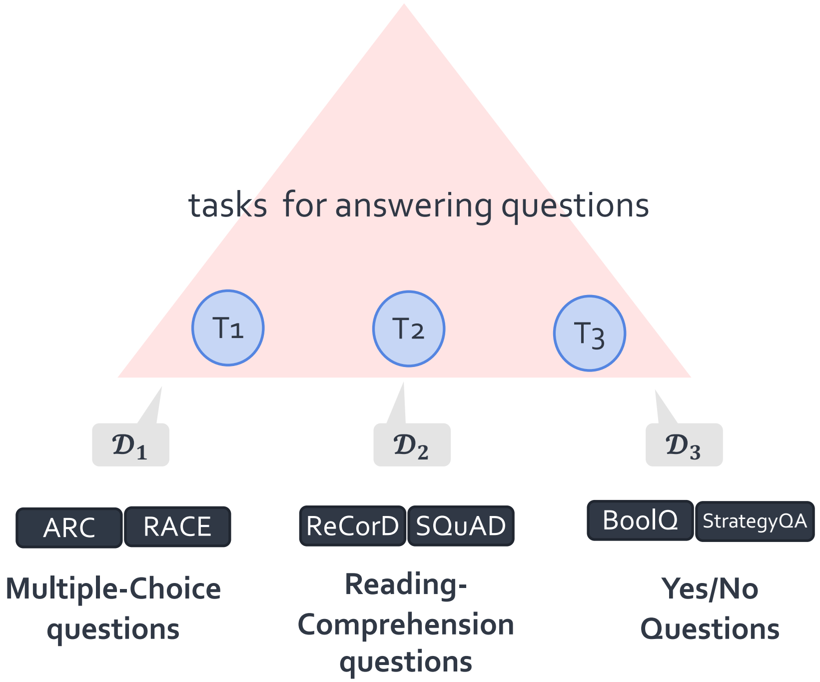
# Toward Unified Question Answering

$\mathbf{x} =$ **Question:** *"What does photosynthesis produce that helps plants grow?"*

**Candidates:** *(A) water* *(B) oxygen* *(C) protein* *(D) sugar*

multiple-choice

$\mathbf{y} =$ *"oxygen"*



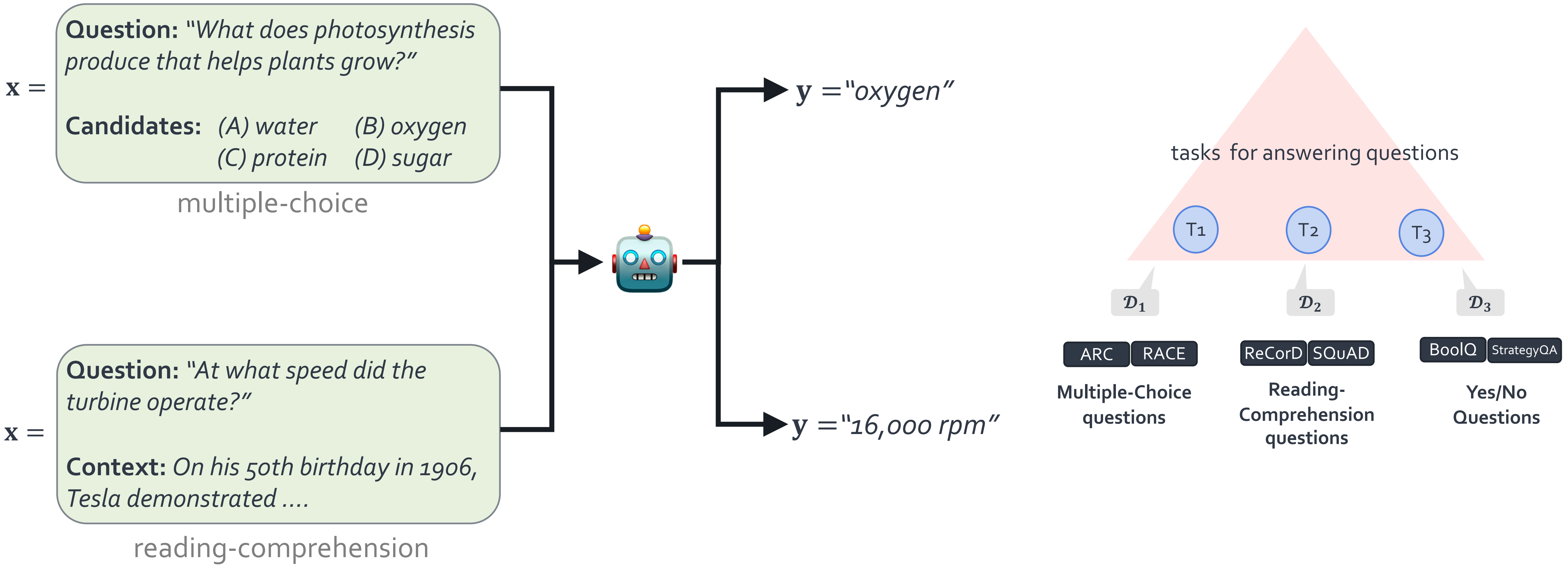

$\mathbf{x} =$ **Question:** *"At what speed did the turbine operate?"*

**Context:** *On his 50th birthday in 1906, Tesla demonstrated ….*

reading-comprehension

$\mathbf{y} =$ *"16,000 rpm"*

# Toward Unified Question Answering



# Toward Unified Question Answering

$\mathbf{x}$ = **Question:** *"What does photosynthesis produce that helps plants grow?"*

**Candidates:** *(A) water* *(B) oxygen* *(C) protein* *(D) sugar*

multiple-choice

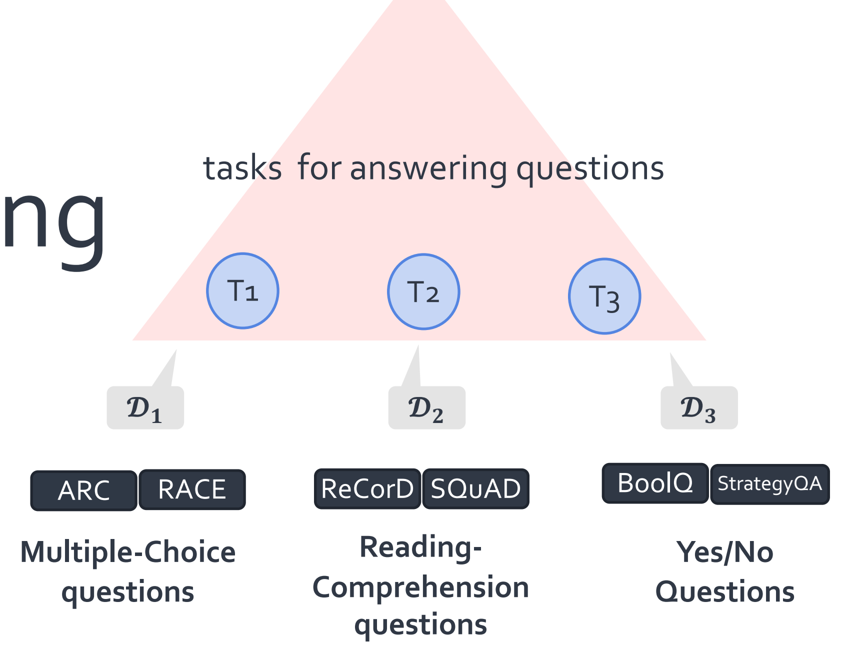
$\mathbf{x}$ = **Question:** *"At what speed did the turbine operate?"*

**Context:** *On his 50th birthday in 1906, Tesla demonstrated ….*

reading-comprehension

$\mathbf{y}$ =*"oxygen"*

$\mathbf{y}$ =*"16,000 rpm"*




**Multi-Task Learning in NLP:**
How can we leverage (induce) commonalities among ~~language~~ tasks?

*Question Answering*

# Toward Unified Question Answering

$\mathbf{x} =$ **Question:** *"What does photosynthesis produce that helps plants grow?"*

**Candidates:** *(A) water* *(B) oxygen* *(C) protein* *(D) sugar*

multiple-choice

**different but relevant**

$\mathbf{x} =$ **Question:** *"At what speed did the turbine operate?"*

**Context:** *On his 50th birthday in 1906, Tesla demonstrated ….*

reading-comprehension

$\mathbf{y} =$ *"oxygen"*

$\mathbf{y} =$ *"16,000 rpm"*

tasks for answering questions

T1 T2 T3

$\mathcal{D}_1$ $\mathcal{D}_2$ $\mathcal{D}_3$

ARC RACE

**Multiple-Choice questions**

ReCorD SQuAD

**Reading-Comprehension questions**

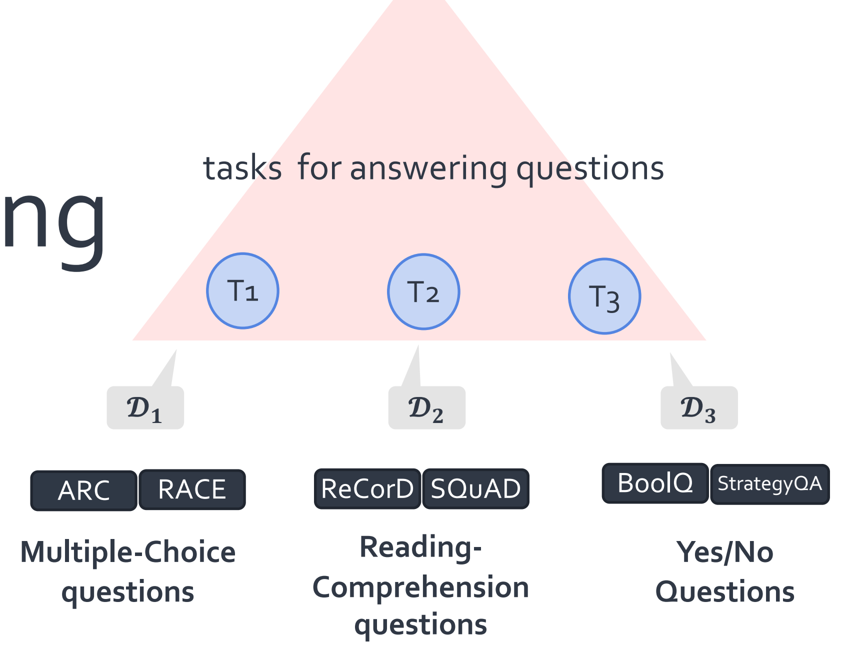
BoolQ StrategyQA

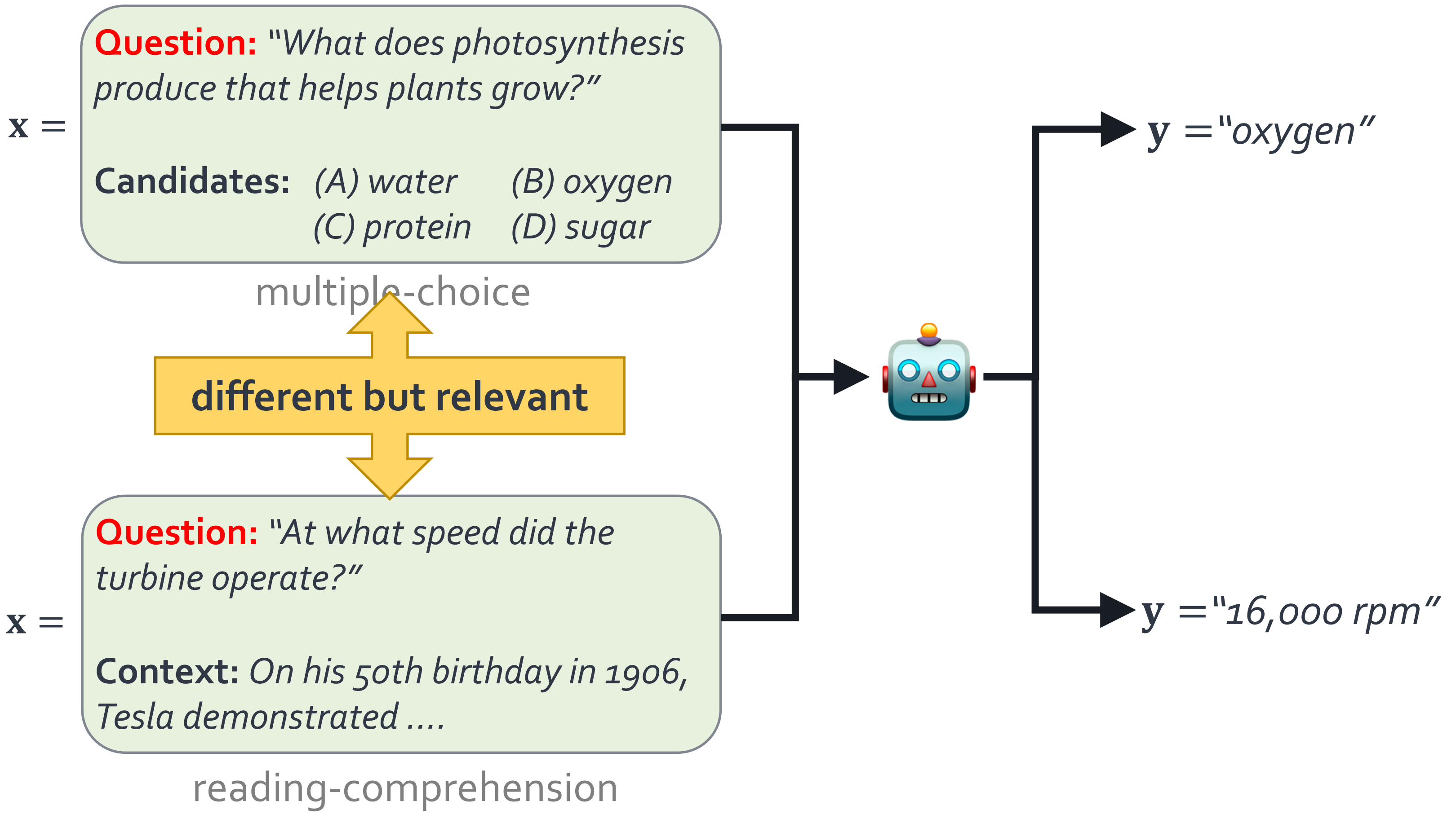
**Yes/No Questions**

**Multi-Task Learning in NLP:**
How can we leverage (induce) commonalities among language tasks?

*Question Answering*

# Toward Unified Question Answering



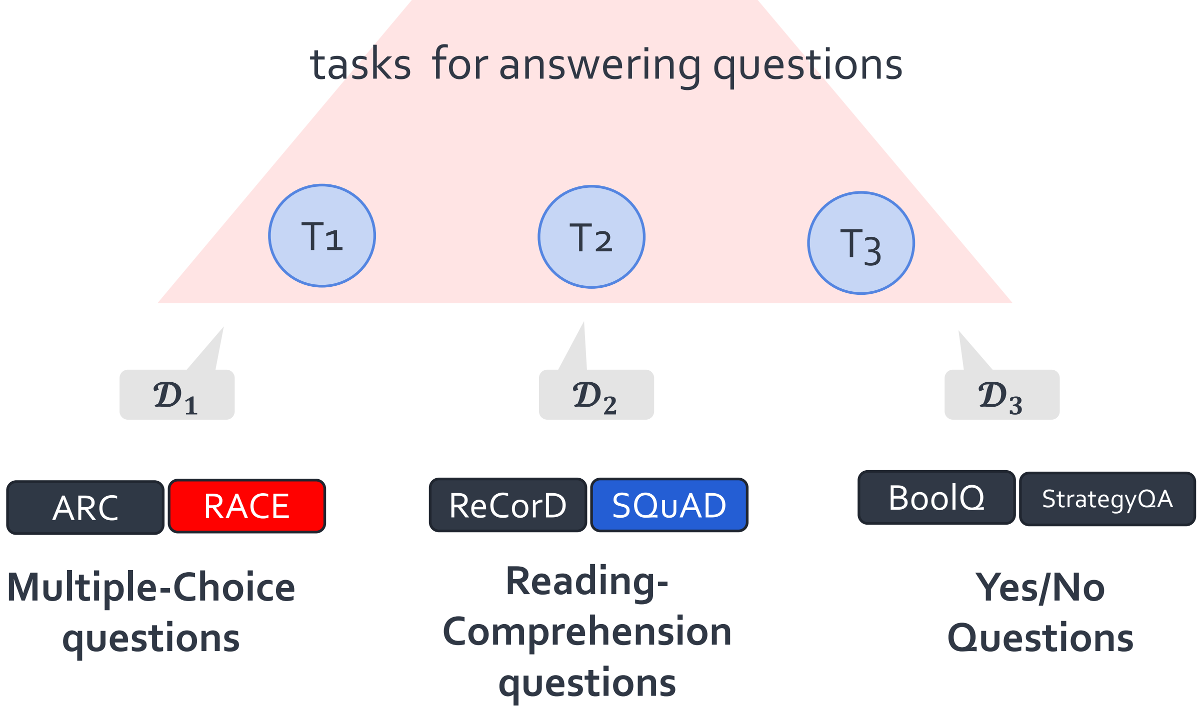




**Hypothesis: "questions"** induce sharedness among QA subtasks.

**Multi-Task Learning in NLP:**
How can we leverage (induce) commonalities among ~~language~~ tasks?

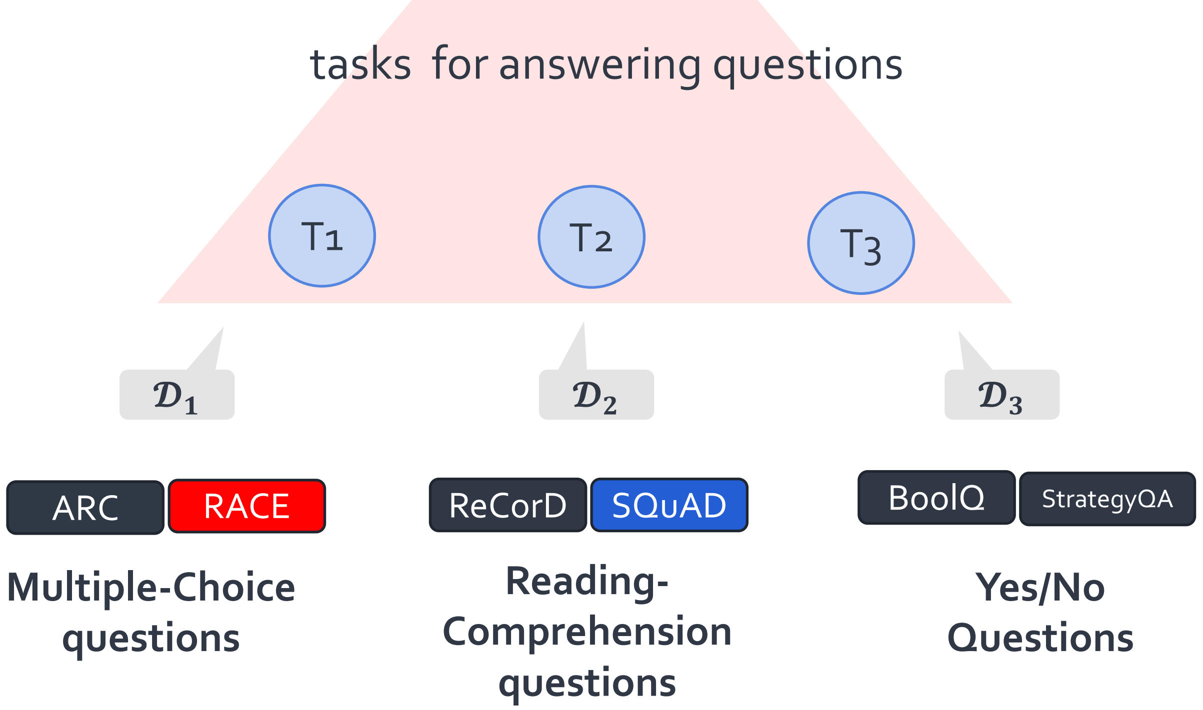
*Question Answering*

**Hypothesis:** "questions" induce sharedness among QA subtasks.



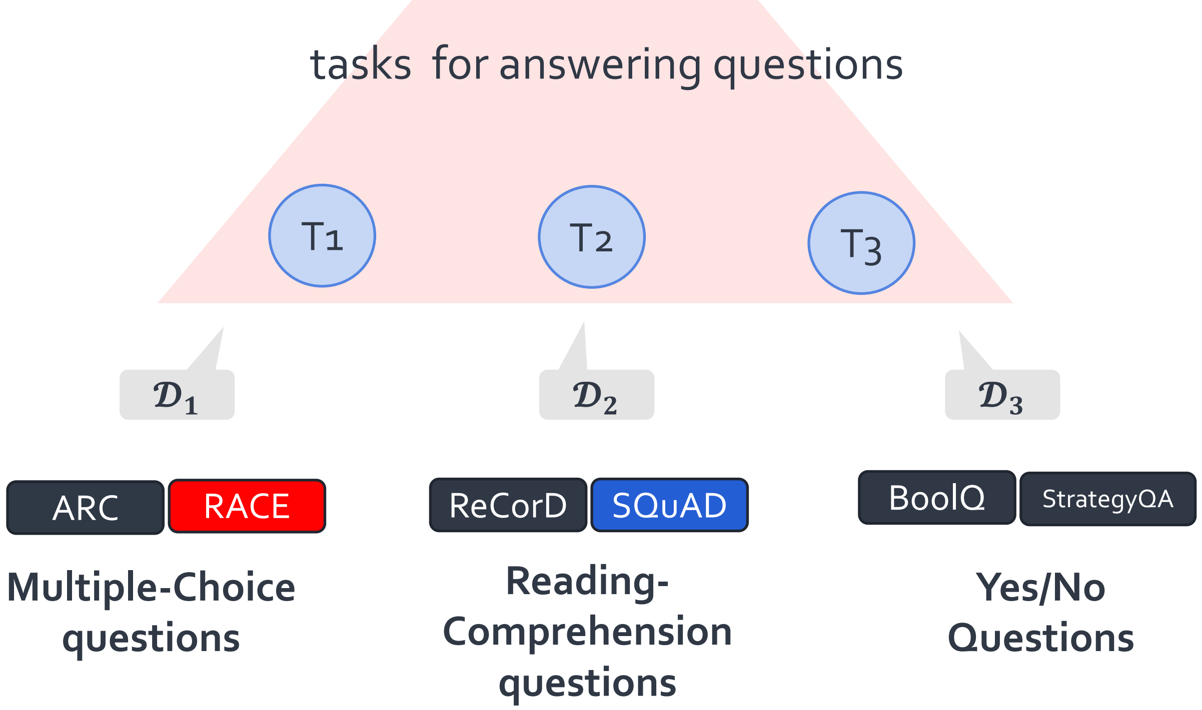
**Hypothesis:** "questions" induce sharedness among QA subtasks.

**Pairwise transferability:**
gain in mixing pairs of QA tasks?



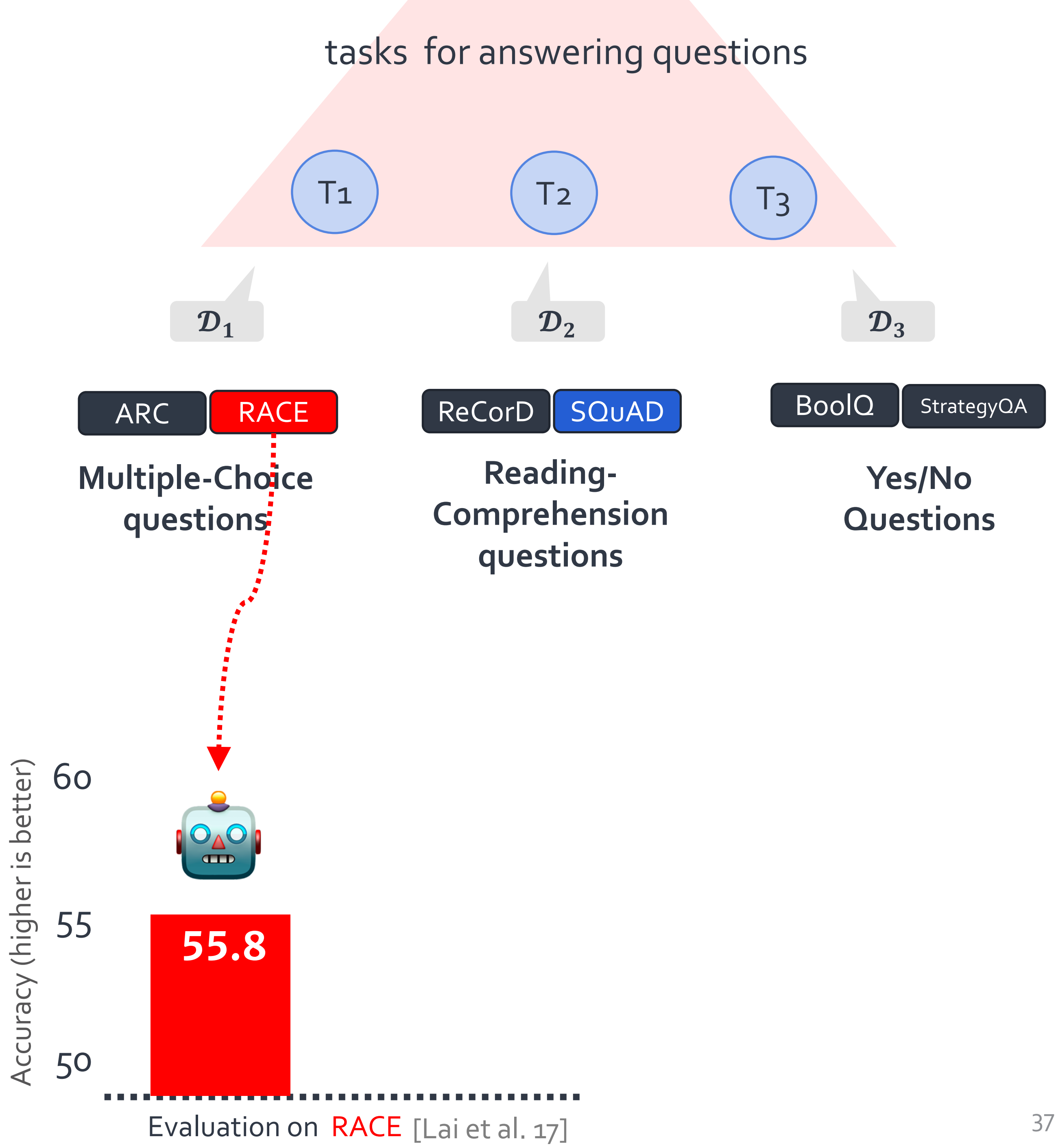
**Hypothesis:** "questions" induce sharedness among QA subtasks.

**Pairwise transferability:**
gain in mixing pairs of QA tasks?



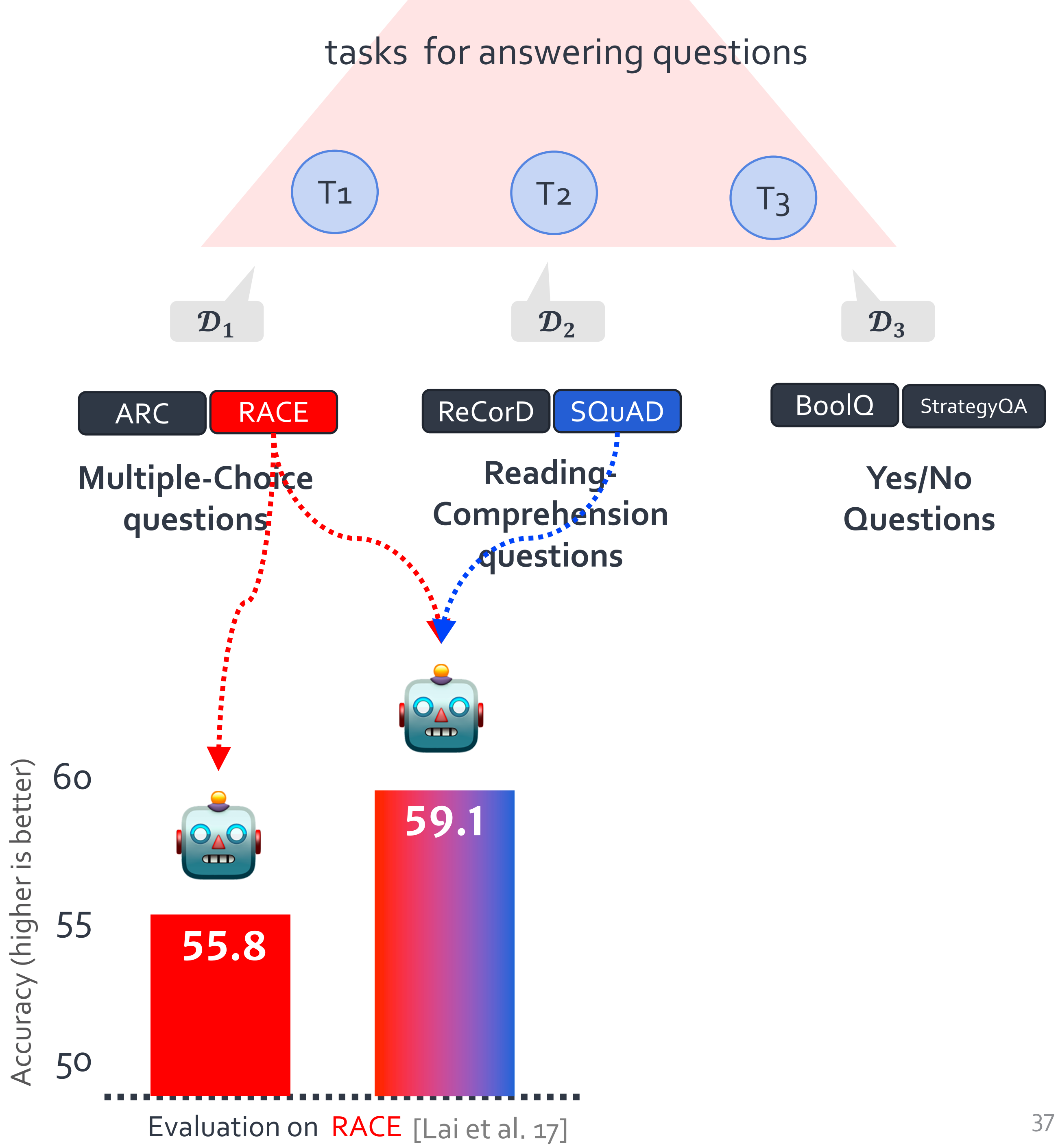

**Hypothesis:** "questions" induce sharedness among QA subtasks.

**Pairwise transferability:** gain in mixing pairs of QA tasks?



**Hypothesis:** "questions" induce sharedness among QA subtasks.

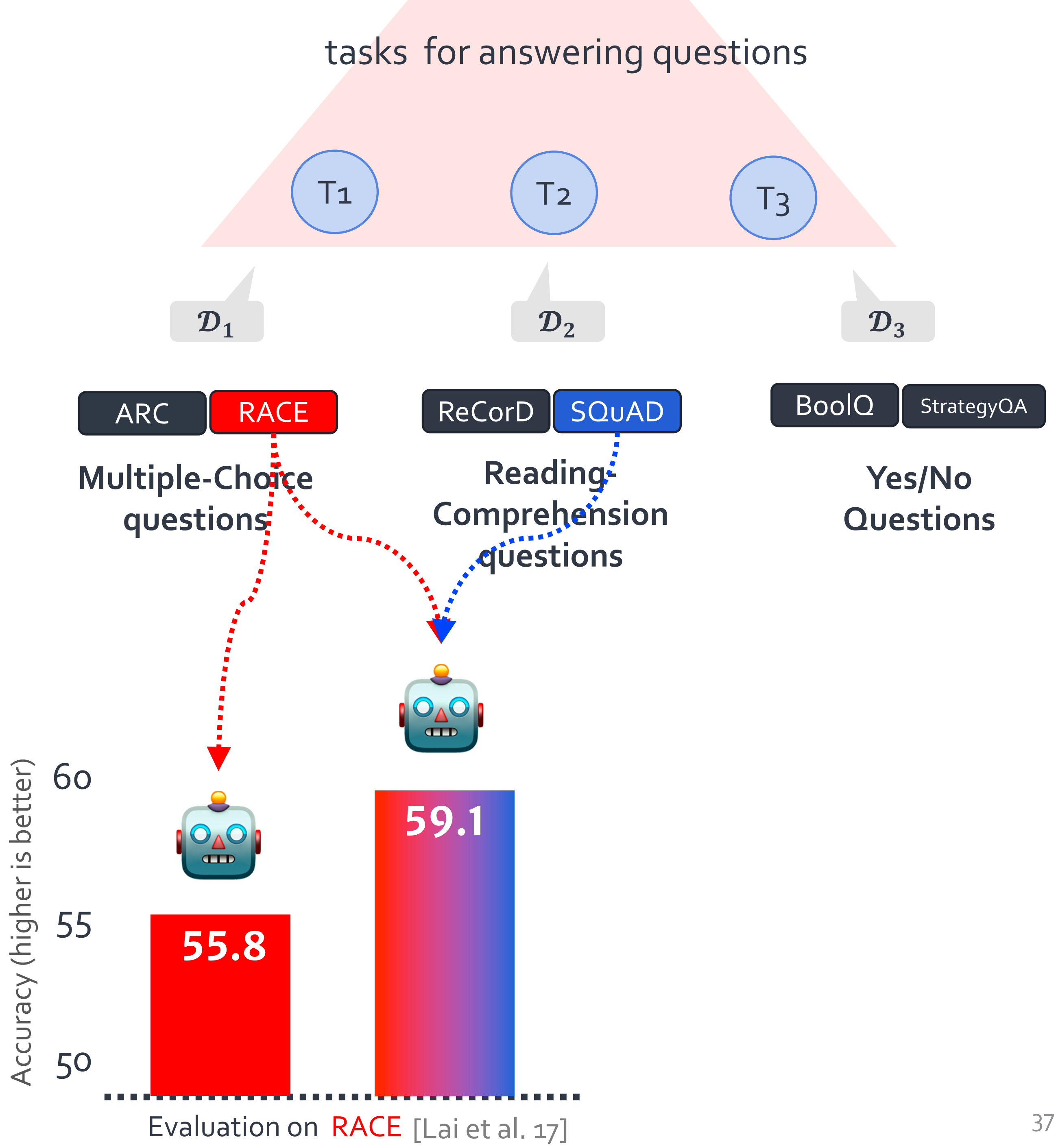
**Pairwise transferability:**
gain in mixing pairs of QA tasks?



**Hypothesis:** "questions" induce sharedness among QA subtasks.

**Pairwise transferability:**
gain in mixing pairs of QA tasks?

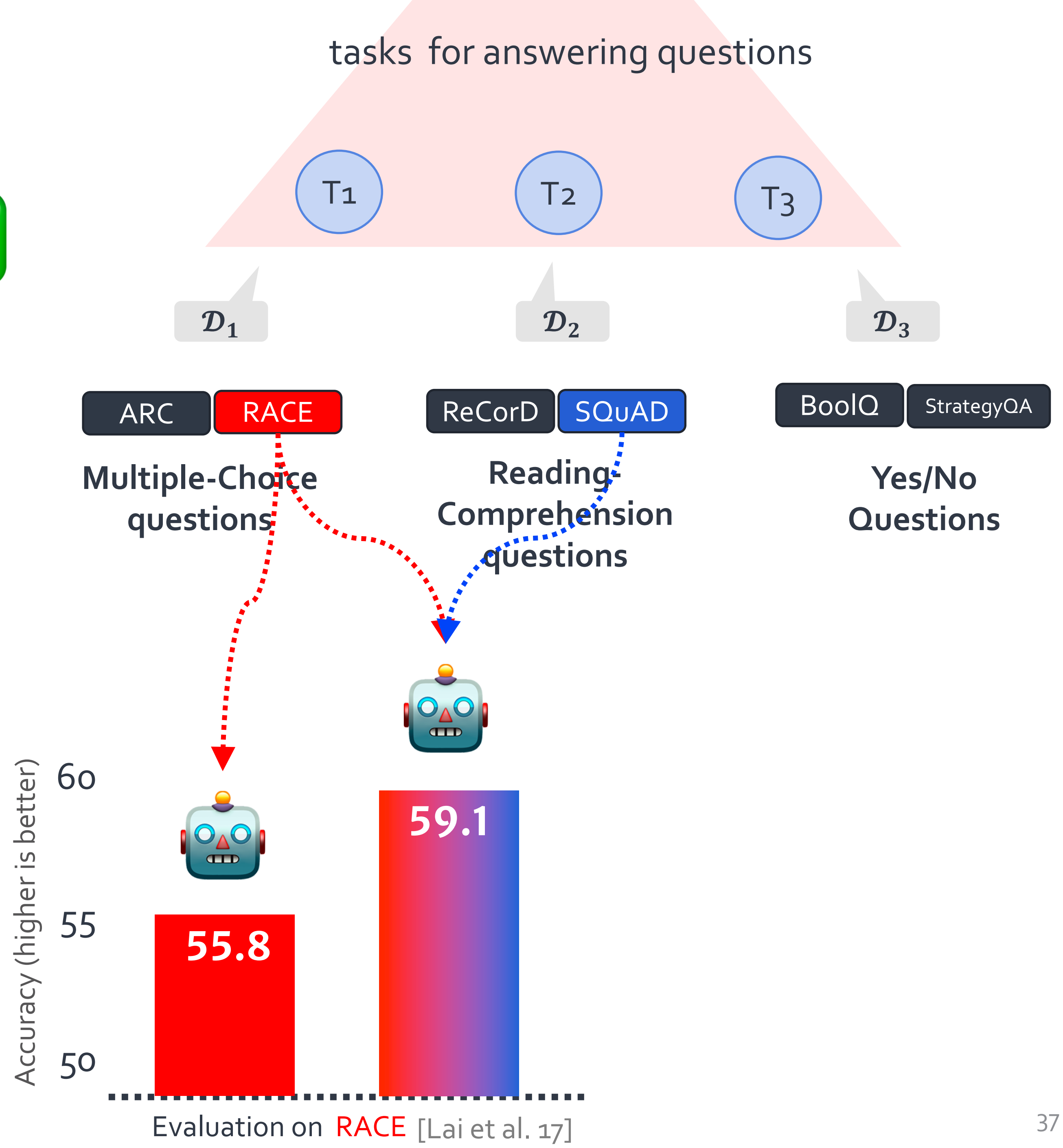
Yes, mixing datasets of different QA subtasks often leads to **positive transfer**.



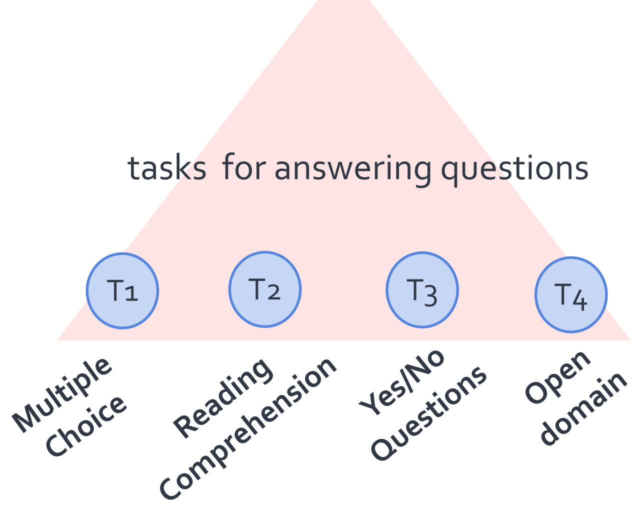
**Hypothesis:** "questions" induce sharedness among QA subtasks.

**Pairwise transferability:**
gain in mixing pairs of QA tasks?

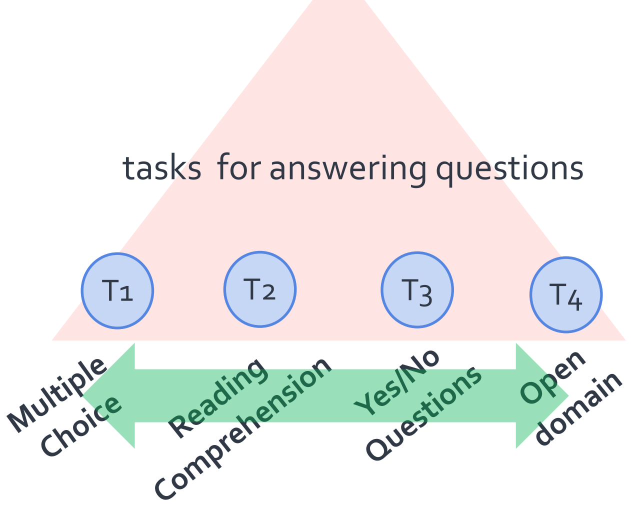
Yes, mixing datasets of different QA subtasks often leads to **positive transfer**.



# A Single Unified Model for QA

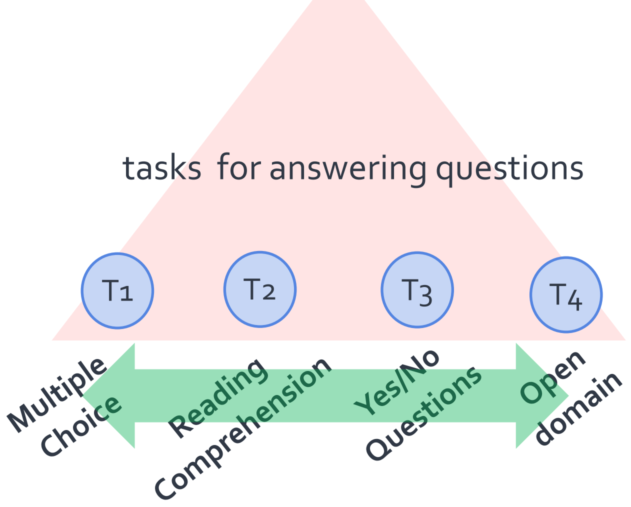


# A Single Unified Model for QA



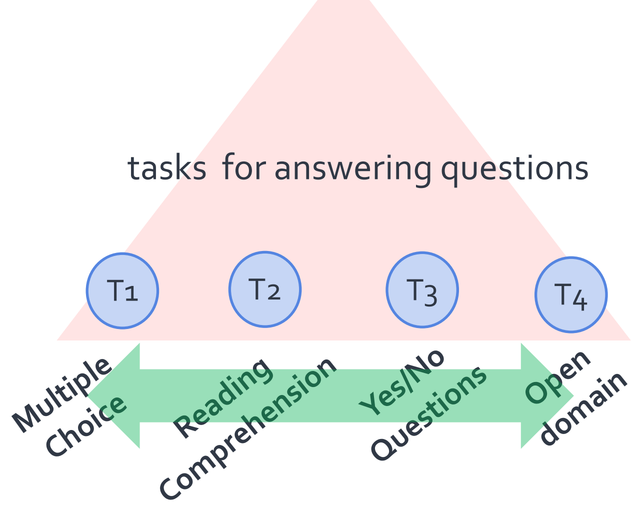

- **UnifiedQA:** a model trained on the union of datasets from four different QA tasks.

# A Single Unified Model for QA

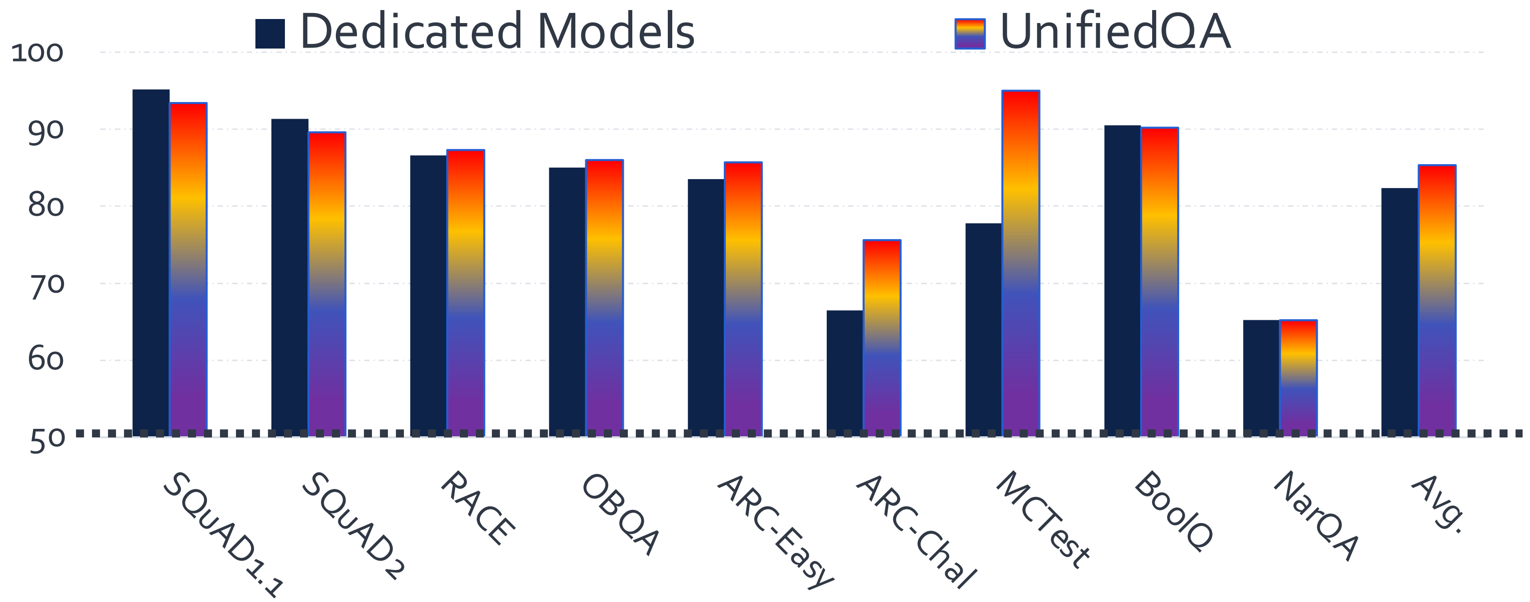



- **UnifiedQA:** a model trained on the union of datasets from four different QA tasks.
- Summary of empirical results:

# A Single Unified Model for QA

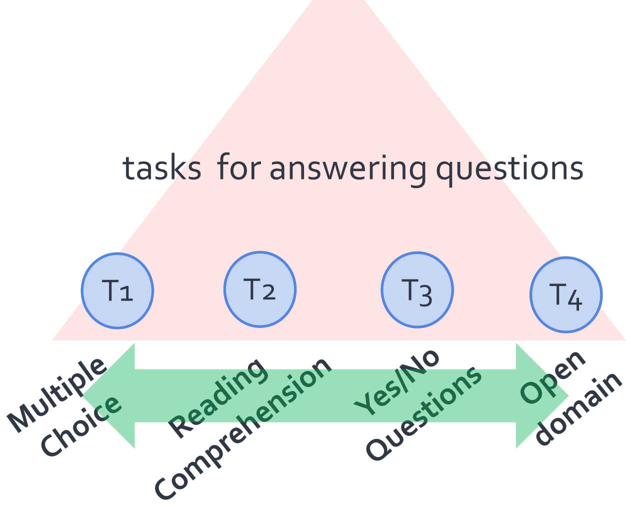



- **UnifiedQA:** a model trained on the union of datasets from four different QA tasks.
- Summary of empirical results:
  - Outperforms dataset-specific models




"UnifiedQA: Crossing Format Boundaries With a Single QA System." **Khashabi** and others EMNLP-Findings '20

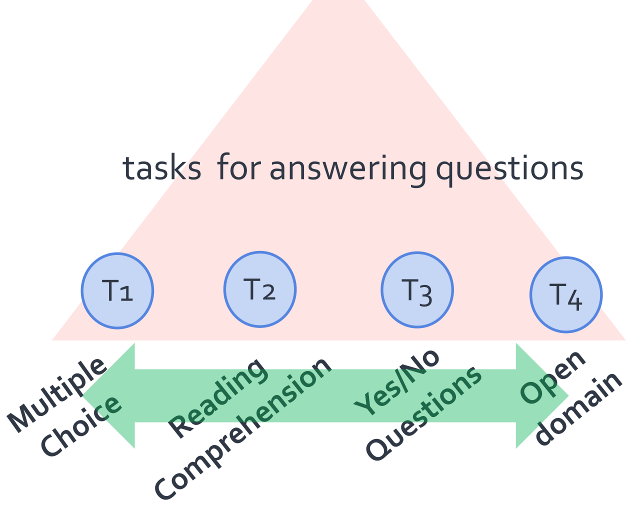
# A Single Unified Model for QA




- **UnifiedQA:** a model trained on the union of datasets from four different QA tasks.
- Summary of empirical results:
  - Outperforms dataset-specific models

"UnifiedQA: Crossing Format Boundaries With a Single QA System." **Khashabi** and others EMNLP-Findings '20

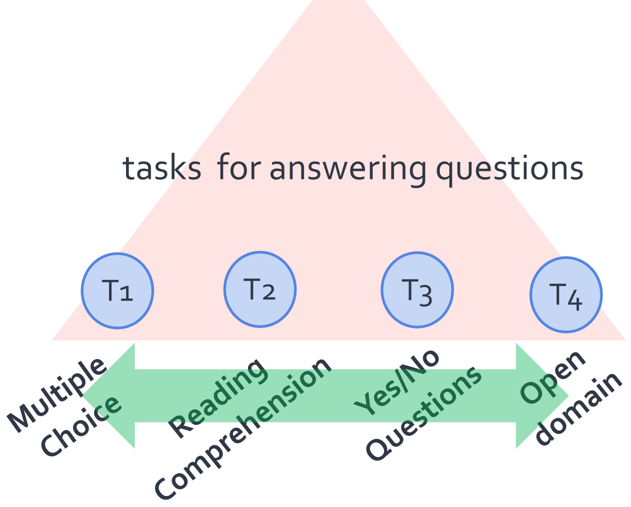
# A Single Unified Model for QA



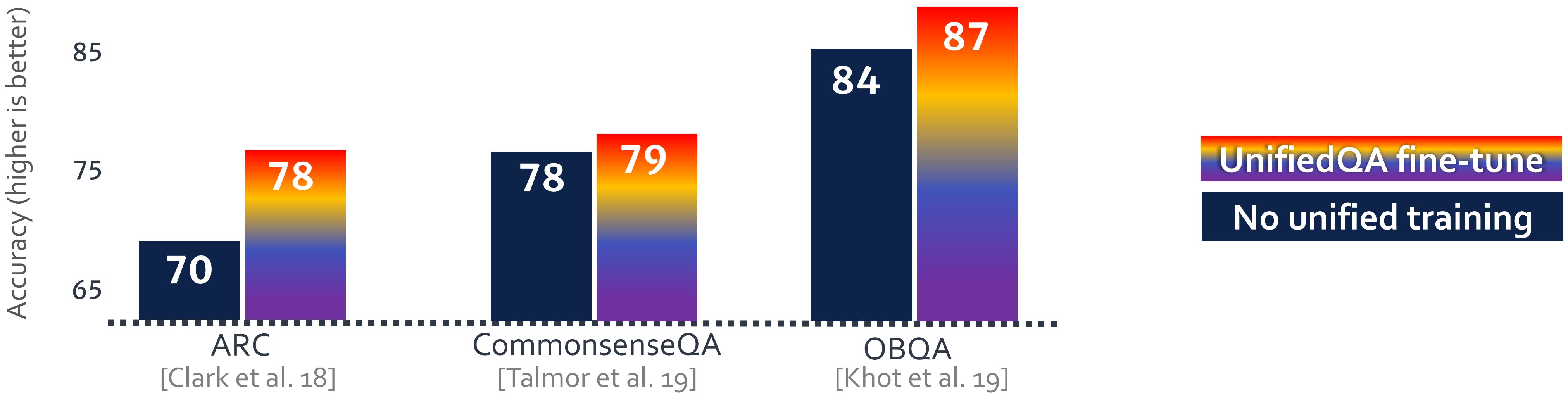

- **UnifiedQA:** a model trained on the union of datasets from four different QA tasks.
- Summary of empirical results:
  - Outperforms dataset-specific models
  - Improved state-of-art results on 10 datasets.

"UnifiedQA: Crossing Format Boundaries With a Single QA System." **Khashabi** and others EMNLP-Findings '20

# A Single Unified Model for QA

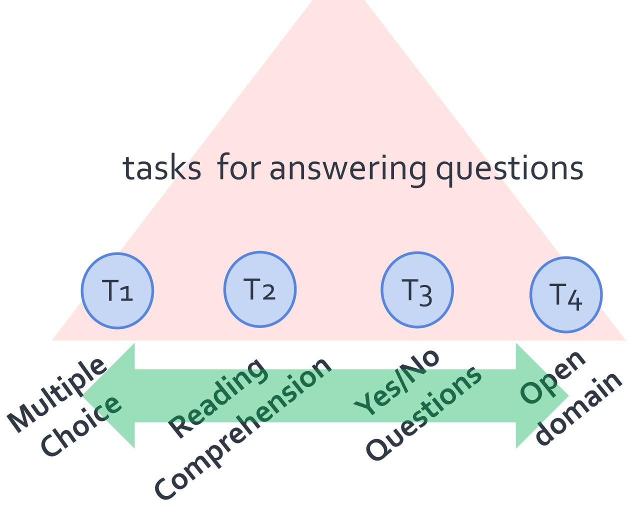



- **UnifiedQA:** a model trained on the union of datasets from four different QA tasks.
- Summary of empirical results:
  - Outperforms dataset-specific models
  - Improved state-of-art results on 10 datasets.




"UnifiedQA: Crossing Format Boundaries With a Single QA System." **Khashabi** and others EMNLP-Findings '20

# A Single Unified Model for QA




- **UnifiedQA:** a model trained on the union of datasets from four different QA tasks.
- Summary of empirical results:
  - Outperforms dataset-specific models
  - Improved state-of-art results on 10 datasets.
  - Strong generalization to *unseen* datasets.

"UnifiedQA: Crossing Format Boundaries With a Single QA System." **Khashabi** and others EMNLP-Findings '20

# UnifiedQA: Impact

- Its empirical success was reproduced on new datasets.

[Bragg et al. '21; Wu et al. '21; Zhong et al. '21, …]

# UnifiedQA: Impact

- Its empirical success was reproduced on new datasets.

[Bragg et al. '21; Wu et al. '21; Zhong et al. '21, ...]

| Model | Answer $F_1$ Span | Answer $F_1$ Abstractive | Answer $F_1$ Overall |
|---|---|---|---|
| LED-base | 54.20 | 24.95 | 44.96 |
| T5-large | 65.59 | 29.11 | 60.03 |
| UnifiedQA-large | 67.23 | 28.92 | 61.39 |

Qasper [Dasigi et al. '21]

| Model | Average |
|---|---|
| Random Baseline | 25.0 |
| RoBERTa | 27.9 |
| ALBERT | 27.1 |
| GPT-2 | 32.4 |
| UnifiedQA | 48.9 |
| GPT-3 Small (few-shot) | 25.9 |
| GPT-3 Medium (few-shot) | 24.9 |
| GPT-3 Large (few-shot) | 26.0 |
| GPT-3 X-Large (few-shot) | 43.9 |

MMMLU [Hendrycks et al. '21]

| | Zero-Shot | | |
|---|---|---|---|
| | **EM** | **F1** | **FZ-R** |
| Human Performance | 79.99 | 89.87 | 92.33 |
| T5-Base (UnifiedQA) | 57.75 | 69.90 | 76.31 |
| T5-Large (UnifiedQA) | 64.83 | 75.73 | 80.59 |
| T5-3B (UnifiedQA) | **66.77** | **76.98** | **81.77** |
| T5-11B (UnifiedQA) | 51.13 | 66.19 | 71.68 |
| GPT-3 | 53.72 | 67.45 | 72.94 |

QAConv [Wu et al. '21]

# UnifiedQA: Impact

- Its empirical success was reproduced on new datasets.

[Bragg et al. '21; Wu et al. '21; Zhong et al. '21, ...]

| Model | Answer $F_1$ Span | Abstractive | Overall |
|---|---|---|---|
| LED-base | 54.20 | 24.95 | 44.96 |
| T5-large | 65.59 | 29.11 | 60.03 |
| UnifiedQA-large | 67.23 | 28.92 | 61.39 |

Qasper [Dasigi et al. '21]

| Model | Average |
|---|---|
| Random Baseline | 25.0 |
| RoBERTa | 27.9 |
| ALBERT | 27.1 |
| GPT-2 | 32.4 |
| UnifiedQA | 48.9 |
| GPT-3 Small (few-shot) | 25.9 |
| GPT-3 Medium (few-shot) | 24.9 |
| GPT-3 Large (few-shot) | 26.0 |
| 16x larger GPT-3 X-Large (few-shot) | 43.9 |

MMMLU [Hendrycks et al. '21]

| | Zero-Shot EM | F1 | FZ-R |
|---|---|---|---|
| Human Performance | 79.99 | 89.87 | 92.33 |
| T5-Base (UnifiedQA) | 57.75 | 69.90 | 76.31 |
| T5-Large (UnifiedQA) | 64.83 | 75.73 | 80.59 |
| T5-3B (UnifiedQA) | **66.77** | **76.98** | **81.77** |
| T5-11B (UnifiedQA) | 51.13 | 66.19 | 71.68 |
| GPT-3 | 53.72 | 67.45 | 72.94 |

QAConv [Wu et al. '21]

# UnifiedQA: Impact

- Its empirical success was reproduced on new datasets.

[Bragg et al. ‘21; Wu et al. ‘21; Zhong et al. ‘21, …]

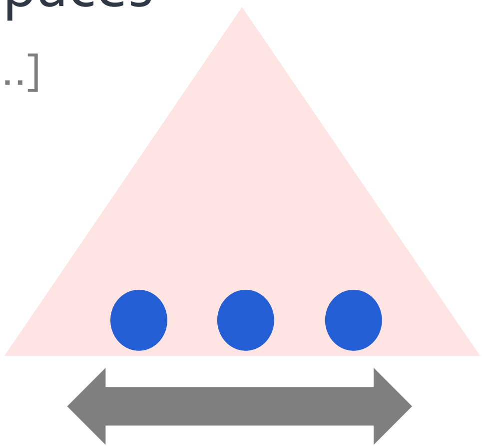
# UnifiedQA: Impact

- Its empirical success was reproduced on new datasets.

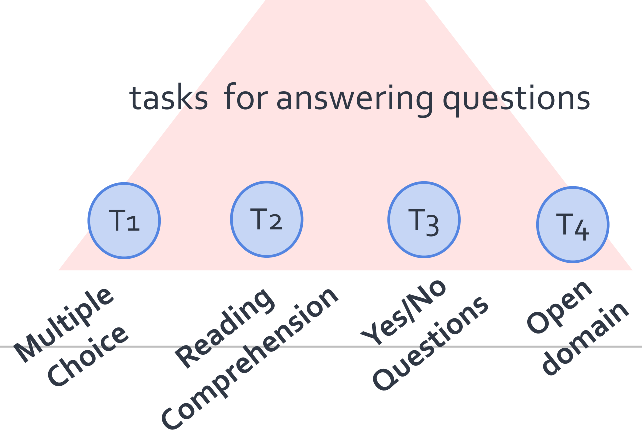
  [Bragg et al. ‘21; Wu et al. ‘21; Zhong et al. ‘21, …]

- Helped alleviated the conceptual barriers for building broader models.
  - Follow-ups works have applied these ideas to different problem spaces

    [Aghajanyan et al.'21, Gupta et al.'21, Jiang et al.21, Bragg et al. ‘21, Aribandi et al. 21, …]



# Summary So Far

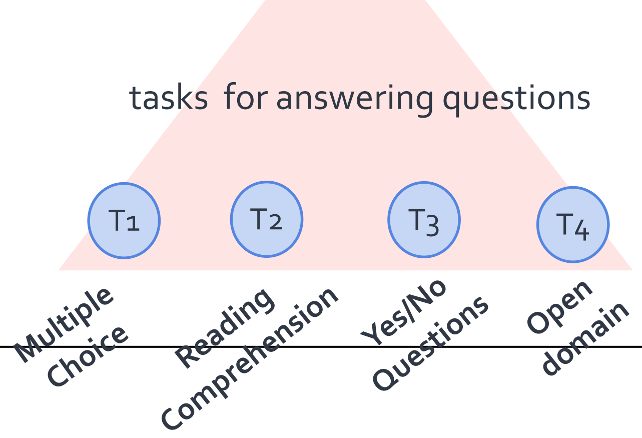



- **Motivating Question:** Can we build a more general individual system that can gains from tackling a variety of QA formats?

- Yes we can!
- Added incentive: there is value in mixing QA tasks.
- UnifiedQA: a single QA system working across four common QA types

- **Open questions:**
  - What about other non-QA tasks?

# Summary So Far

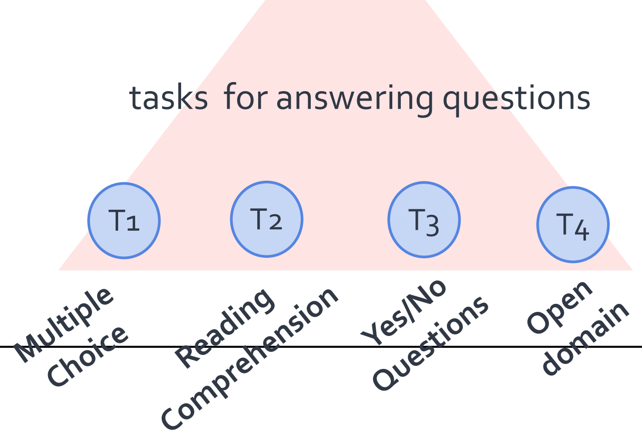



- **Motivating Question:** Can we build a more general individual system that can gains from tackling a variety of QA formats?

- Yes we can!
- Added incentive: there is value in mixing QA tasks.
- UnifiedQA: a single QA system working across four common QA types

- **Open questions:**
  - What about other non-QA tasks?

# Summary So Far

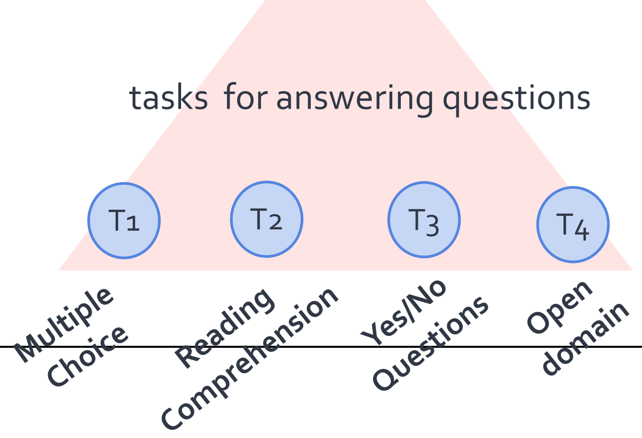



- **Motivating Question:** Can we build a more general individual system that can gains from tackling a variety of QA formats?

- Yes we can!
- Added incentive: there is value in mixing QA tasks.
- UnifiedQA: a single QA system working across four common QA types

- **Open questions:**
  - What about other non-QA tasks?

# Summary So Far

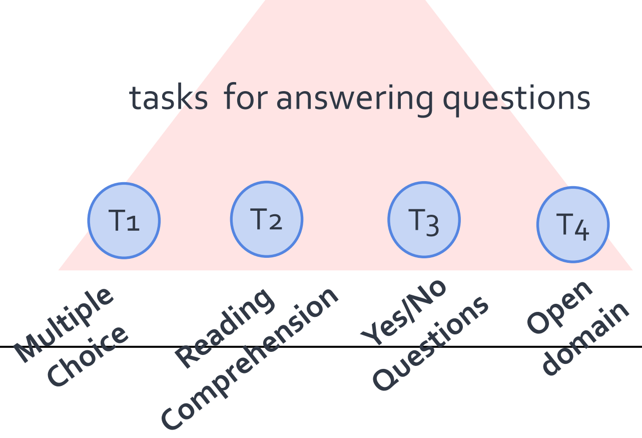



- **Motivating Question:** Can we build a more general individual system that can gains from tackling a variety of QA formats?

- Yes we can!
- Added incentive: there is value in mixing QA tasks.
- UnifiedQA: a single QA system working across four common QA types

- **Open questions:**
  - What about other non-QA tasks?

# Summary So Far

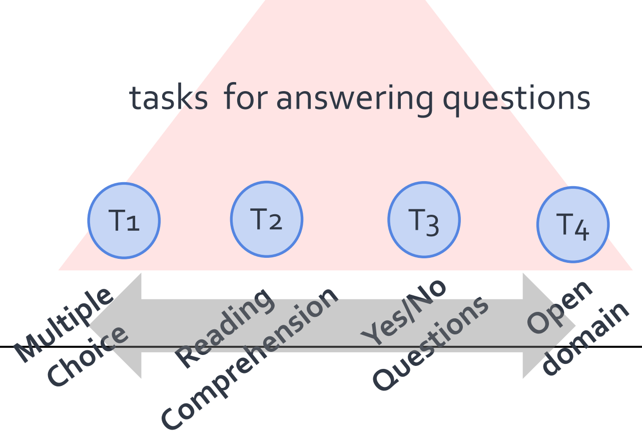



- **Motivating Question:** Can we build a more general individual system that can gains from tackling a variety of QA formats?

- Yes we can!
- Added incentive: there is value in mixing QA tasks.
- UnifiedQA: a single QA system working across four common QA types

- **Open questions:**
  - What about other non-QA tasks?

# Summary So Far




- **Motivating Question:** Can we build a more general individual system that can gains from tackling a variety of QA formats?

- Yes we can!
- Added incentive: there is value in mixing QA tasks.
- UnifiedQA: a single QA system working across four common QA types

- **Open questions:**
  - What about other non-QA tasks?

# Beyond Answering Questions

- There are many other jobs that we can accomplish via language.

# Beyond Answering Questions

- There are many other jobs that we can accomplish via language.

**Pronoun Resolution**

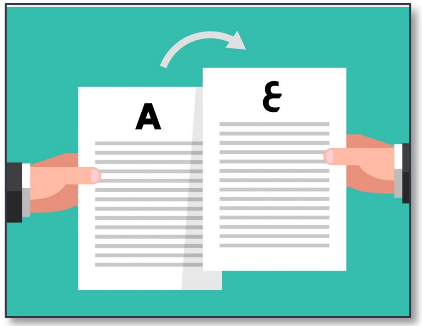
*"Jack fired James but he did not regret it."*

# Beyond Answering Questions

- There are many other jobs that we can accomplish via language.

**Pronoun Resolution**

*"Jack fired James but he did not regret it."*




**Grammar Check**

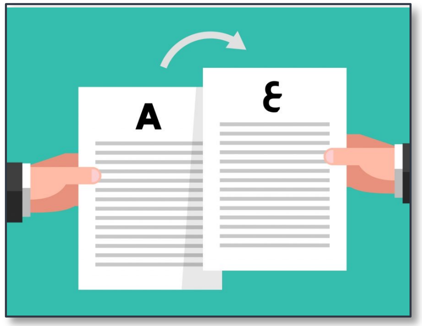
*" … he ~~does~~ not regret."*

*not grammatical*

# Beyond Answering Questions

- There are many other jobs that we can accomplish via language.

**Pronoun Resolution**




**Grammar Check**

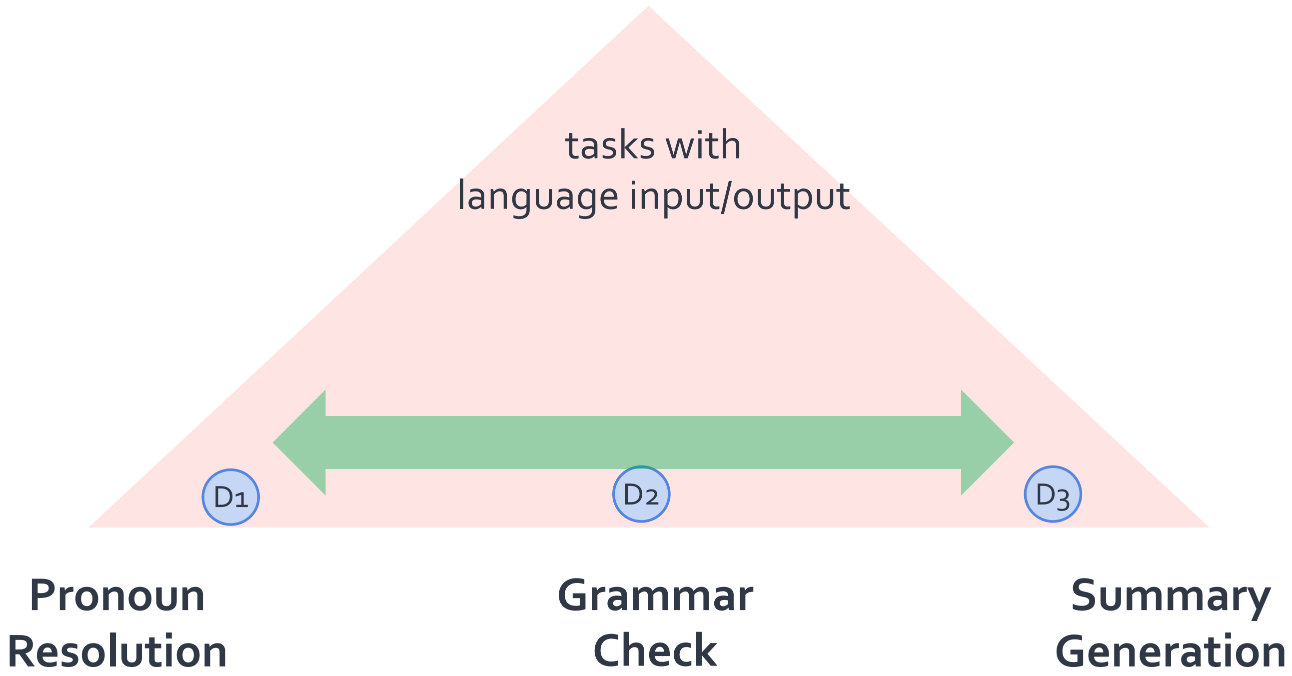
**Summary Generation**

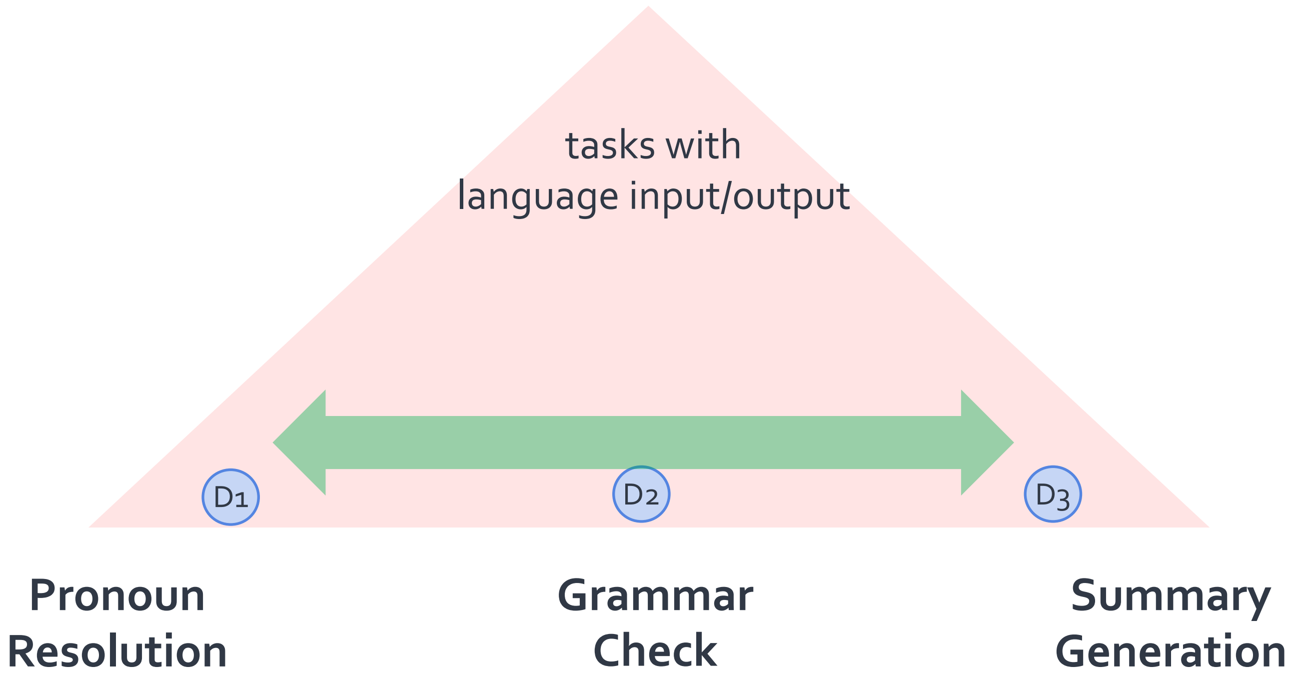
*"Jack fired James but he did not regret it."*

*" … he ~~does~~ not regret."*

*not grammatical*

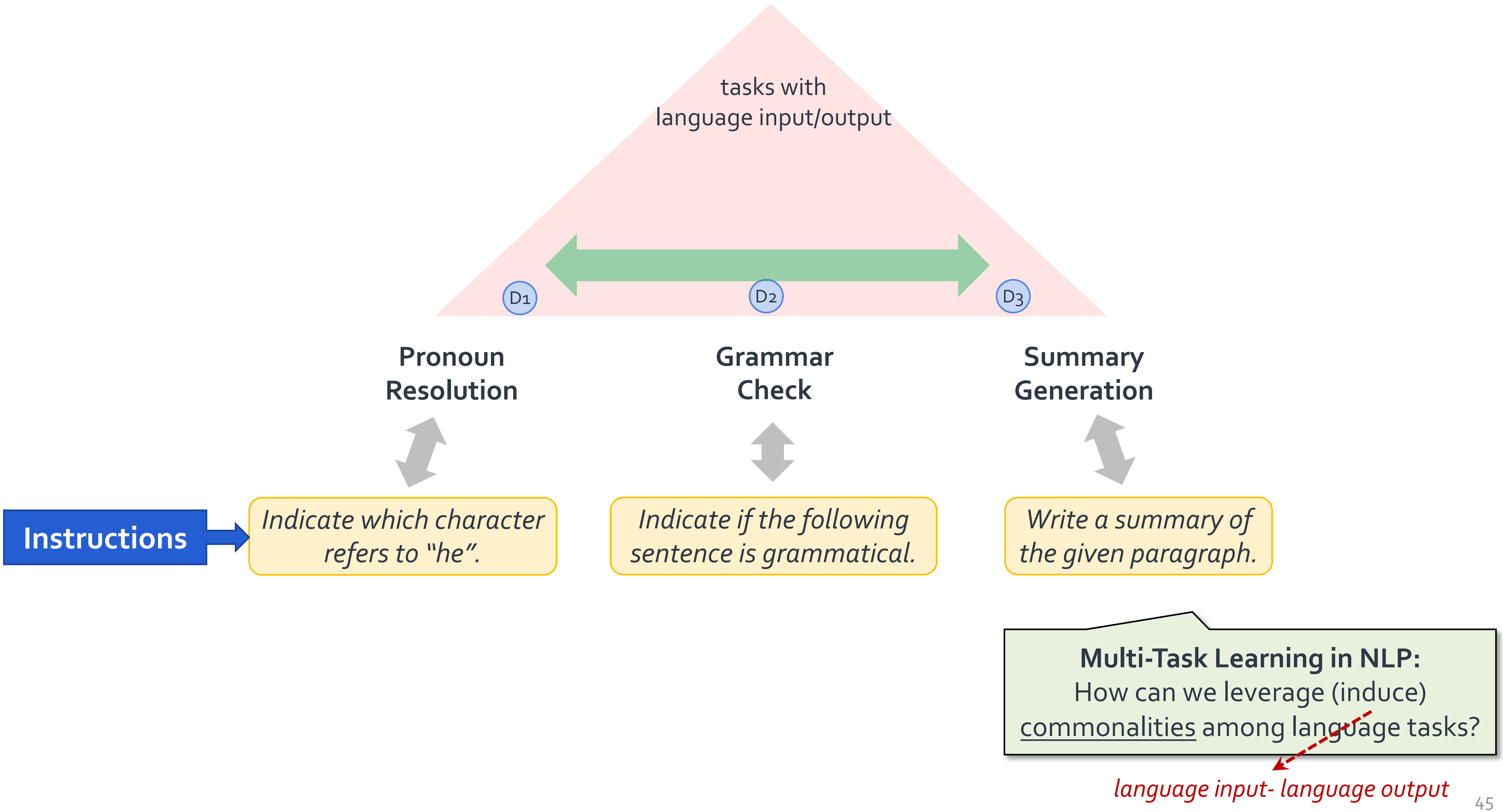
*"In summary, Jack …."*


tasks with
language input/output
D1
D2
D3
Pronoun
Resolution
Grammar
Check
Summary
Generation

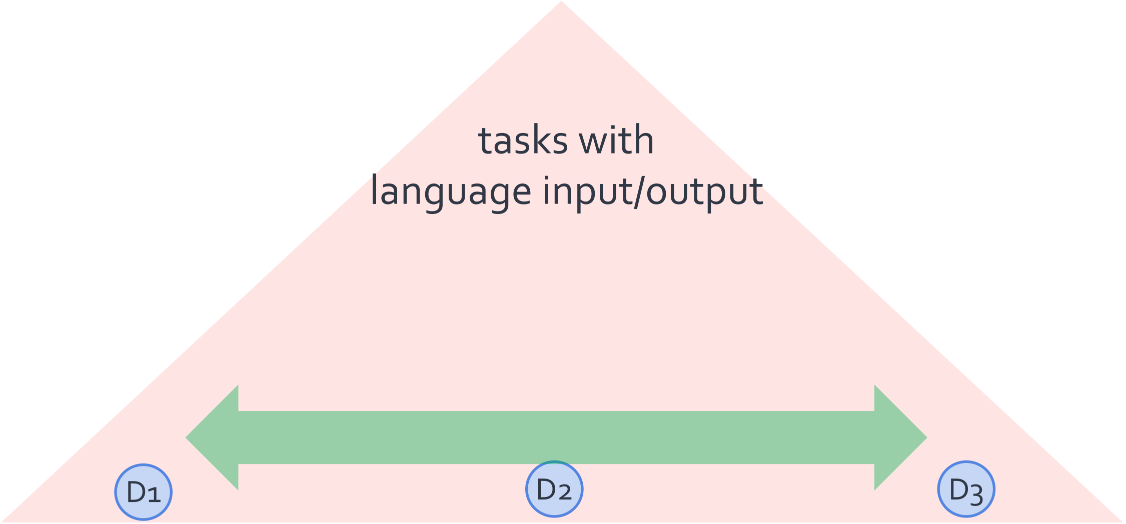



**Multi-Task Learning in NLP:**
How can we leverage (induce) commonalities among language tasks?

*language input- language output*



tasks with
language input/output
D1
D2
D3
Pronoun
Resolution
Grammar
Check
Summary
Generation
Instructions
Indicate which character refers to "he".
Indicate if the following sentence is grammatical.
Write a summary of the given paragraph.
Multi-Task Learning in NLP:
How can we leverage (induce) commonalities among language tasks?
language input- language output




**Pronoun Resolution**

**Grammar Check**

**Summary Generation**

**Instructions**

*Indicate which character refers to "he".*

*Indicate if the following sentence is grammatical.*

*Write a summary of the given paragraph.*

**Hypothesis:** Task "instructions" are enough to induce sharedness among them.

**Multi-Task Learning in NLP:**
How can we leverage (induce) commonalities among language tasks?

*language input- language output*

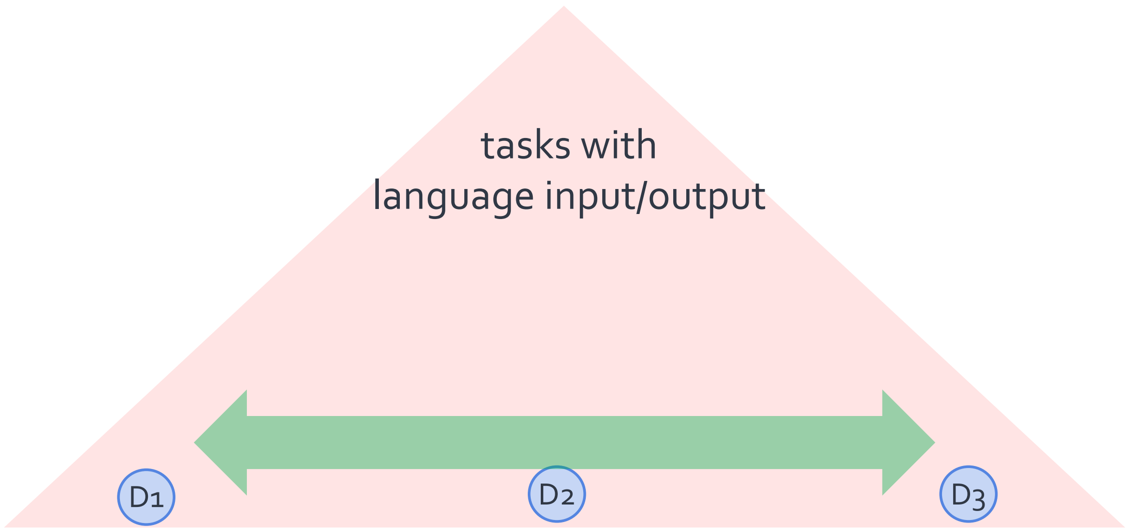
# Generalization via Task Instructions

solving language tasks
via language instructions

Swaroop Mishra, **Daniel Khashabi**, Chitta Baral, Hannaneh Hajishirzi

*ACL 2022*




**Pronoun Resolution**

*Indicate which character refers to "he".*

**Grammar Check**

*Indicate if the following sentence is grammatical.*

**Summary Generation**

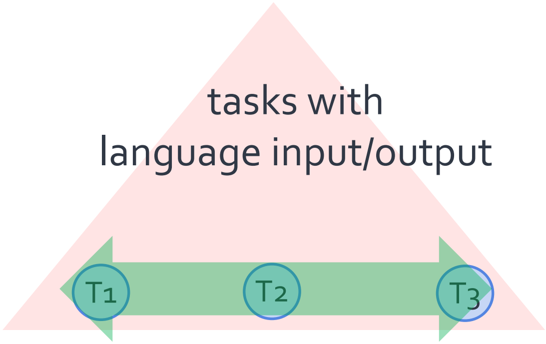
*Write a summary of the given paragraph.*

**Hypothesis:** Task "instructions" are enough to induce sharedness among them.

**Multi-Task Learning in NLP:**
How can we leverage (induce) <u>commonalities</u> among language tasks?

*language input- language output*

# Beyond Task-Specific Models

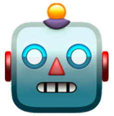



Pronoun Resolution

Grammar Check

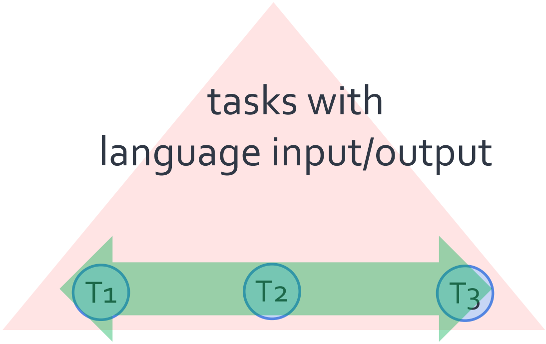
Summary Generation

*Input text*



*????*

# Beyond Task-Specific Models




human readable definitions;
fully define the task

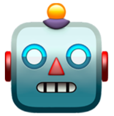
*Instructions*

**Pronoun Resolution** — *Indicate which character refers to "he".*

**Grammar Check**

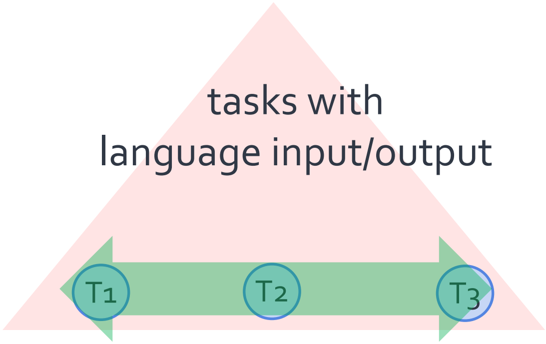
**Summary Generation**

***Input text***



*????*

# Beyond Task-Specific Models

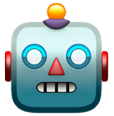



human readable definitions; fully define the task

*Instructions*

**Pronoun Resolution** — *Indicate which character refers to "he".*

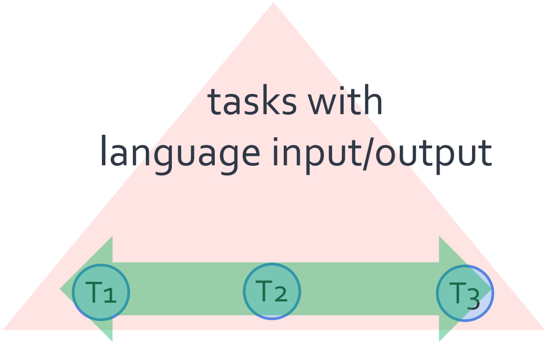
*Input text*



**Grammar Check**

**Summary Generation**

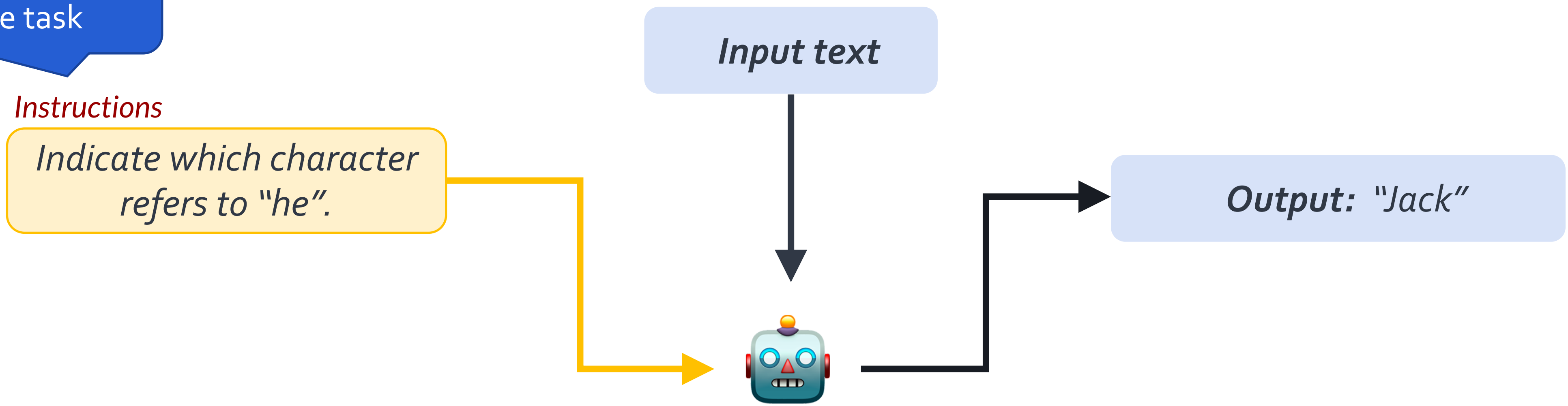
# Beyond Task-Specific Models




human readable definitions; fully define the task

*Instructions*

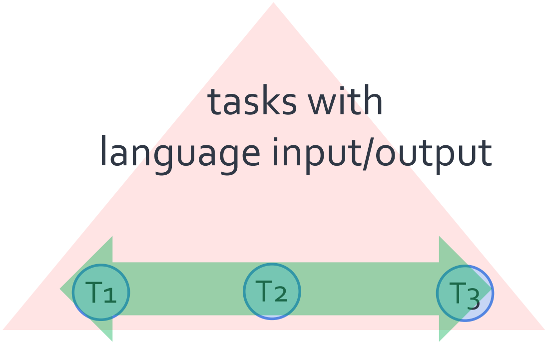
**Pronoun Resolution**




**Grammar Check**

**Summary Generation**

# Beyond Task-Specific Models




human readable definitions; fully define the task

**Pronoun Resolution**

*Instructions*

*Indicate which character refers to "he".*

**Grammar Check**

*Instructions*

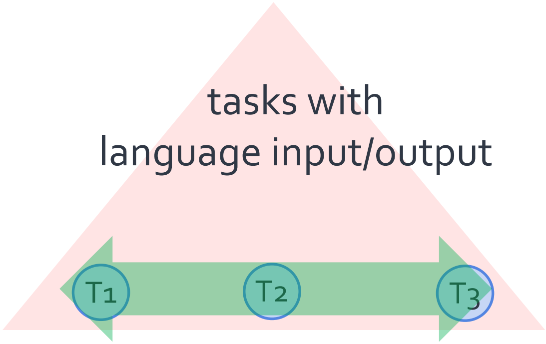
*Indicate if the following sentence is grammatical.*

**Summary Generation**

***Input text***

***Output:** "Jack"*

# Beyond Task-Specific Models




human readable definitions; fully define the task

Input text

**Pronoun Resolution**

*Instructions*

*Indicate which character refers to "he".*

*Output: "Jack"*

**Grammar Check**

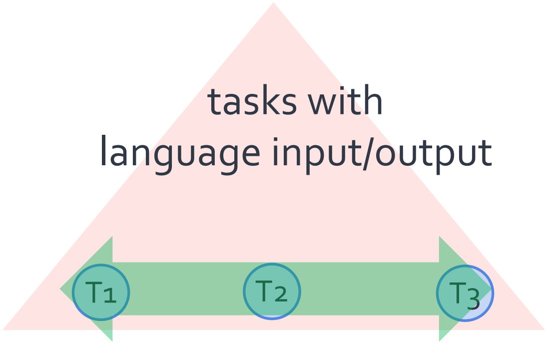
*Instructions*

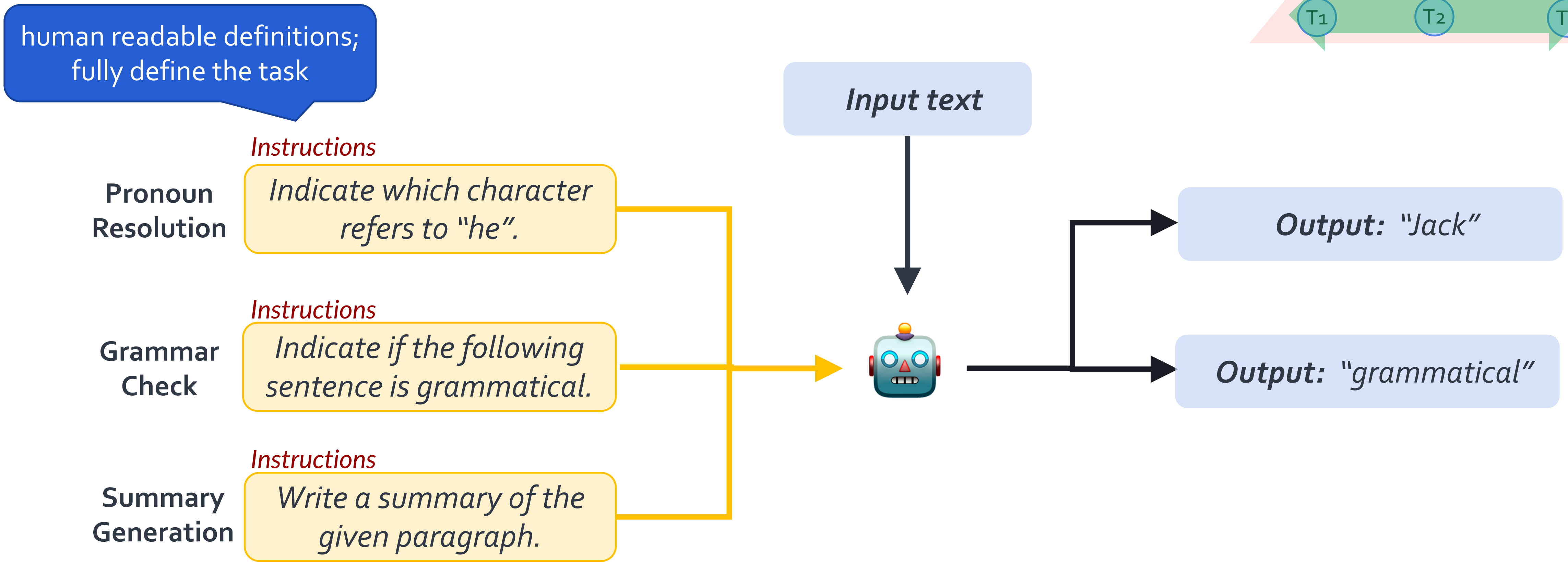
*Indicate if the following sentence is grammatical.*

*Output: "grammatical"*

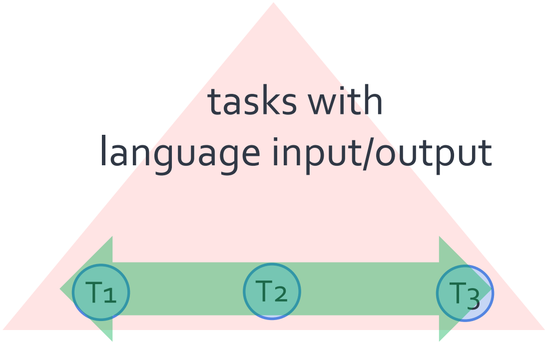
**Summary Generation**

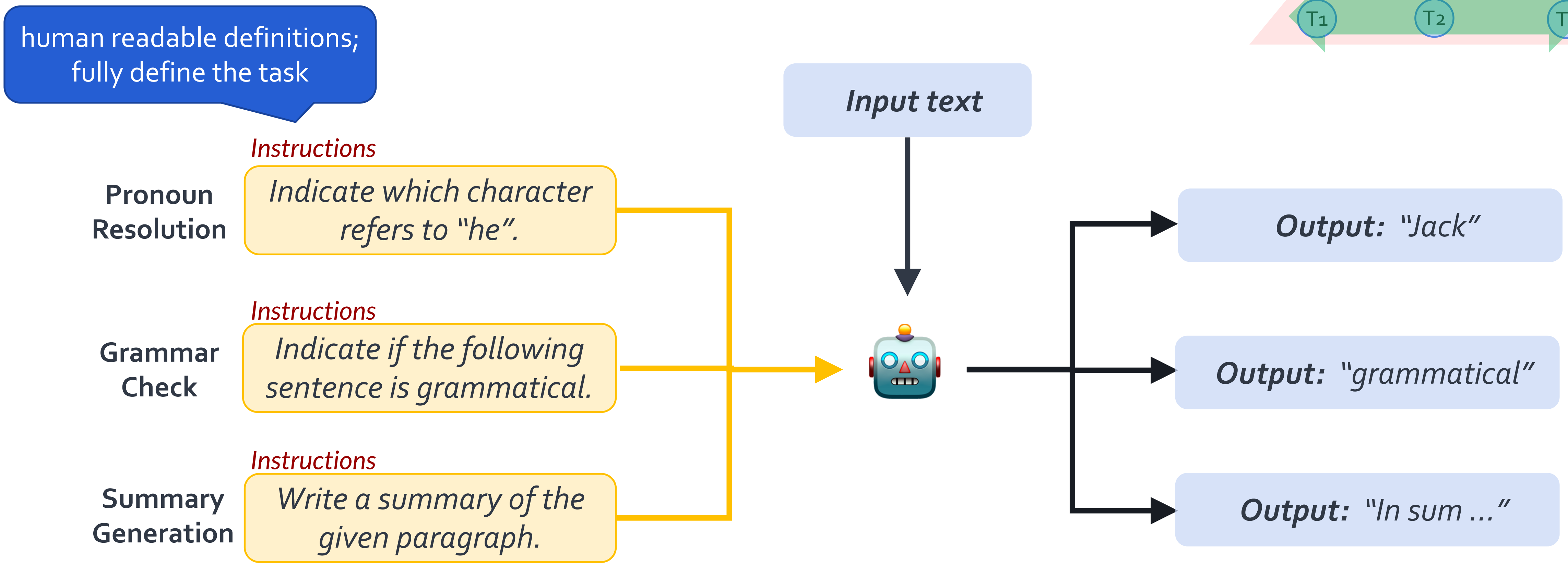
# Beyond Task-Specific Models



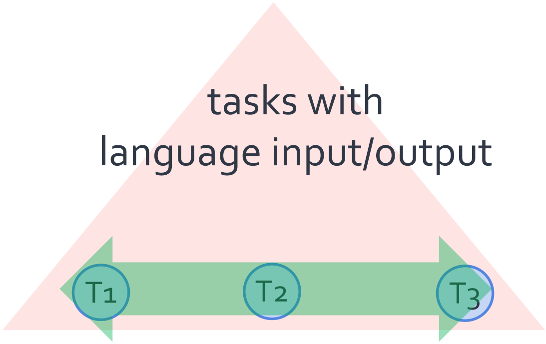



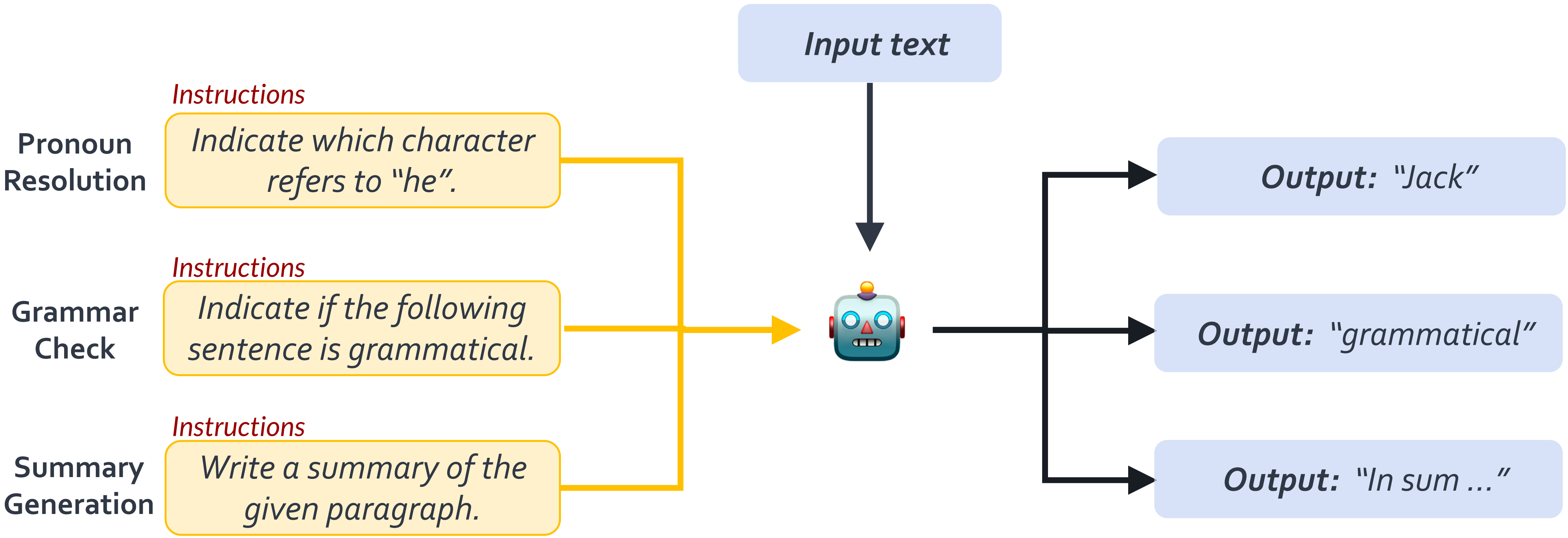
# Beyond Task-Specific Models



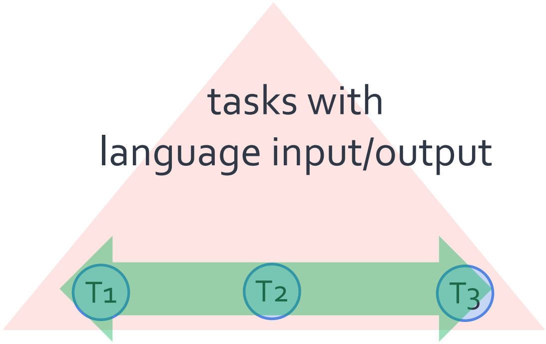



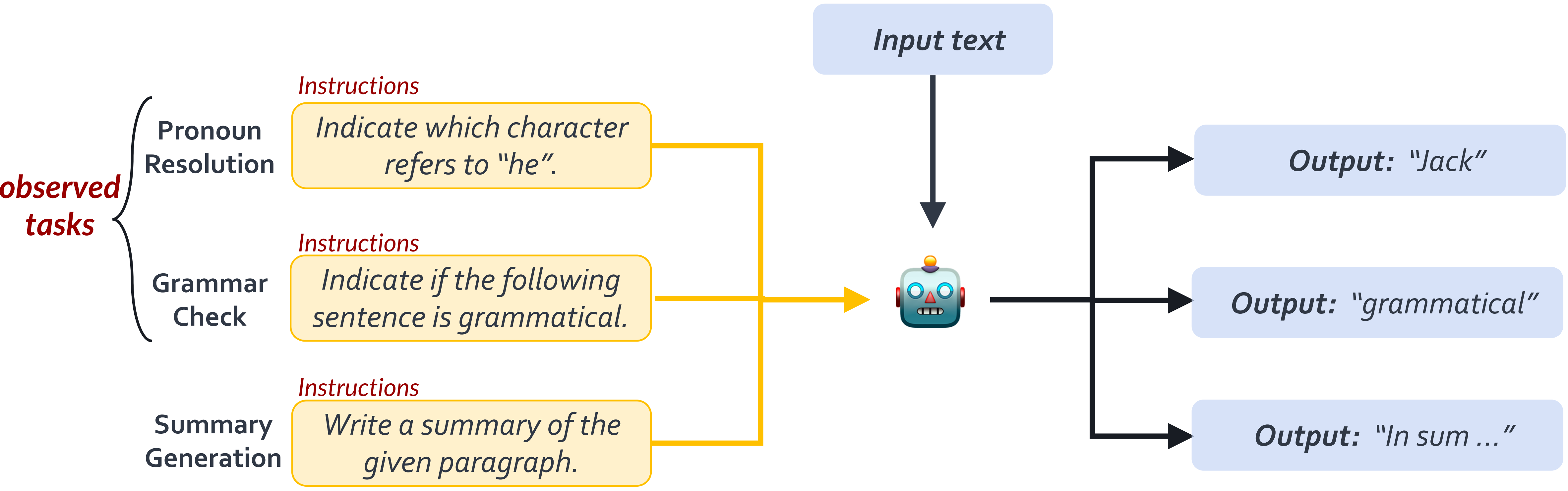
# Cross-Task Generalization



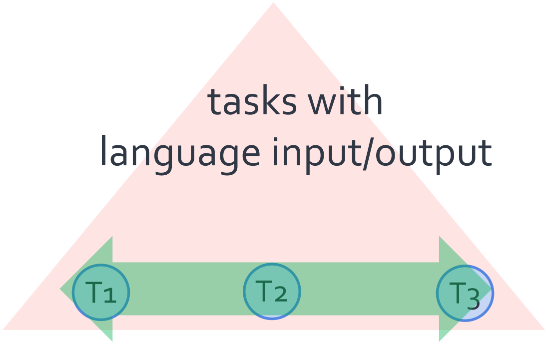



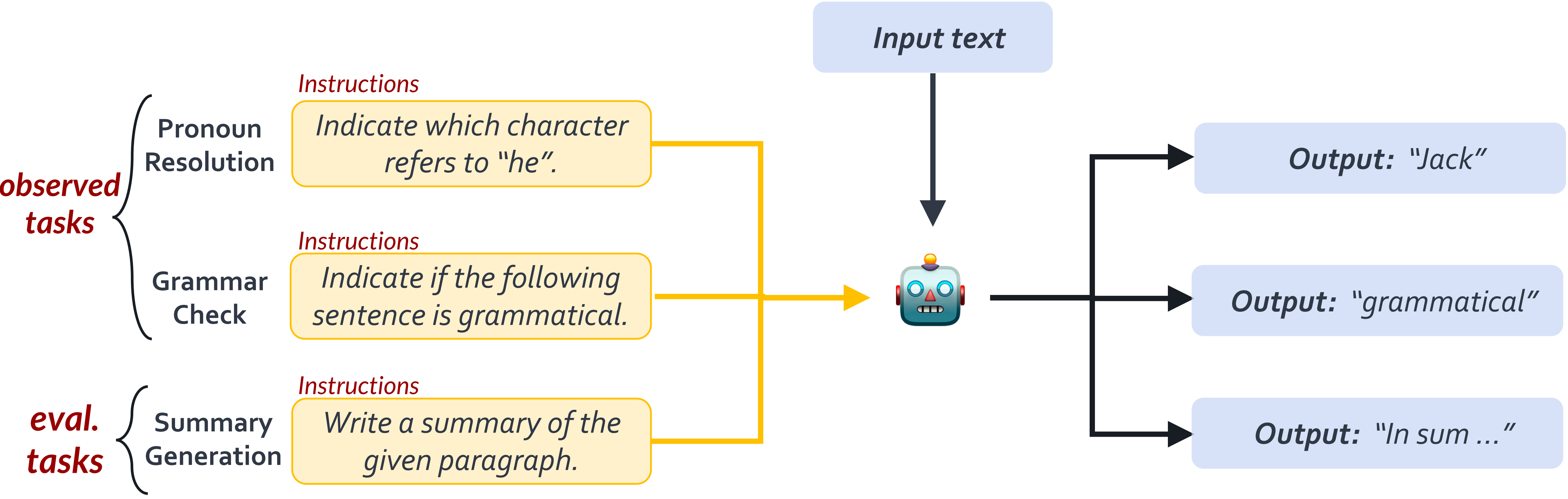
# Cross-Task Generalization



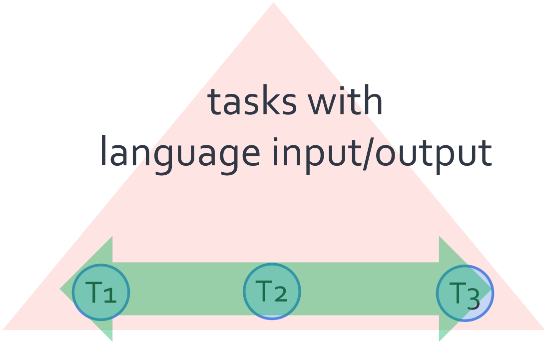



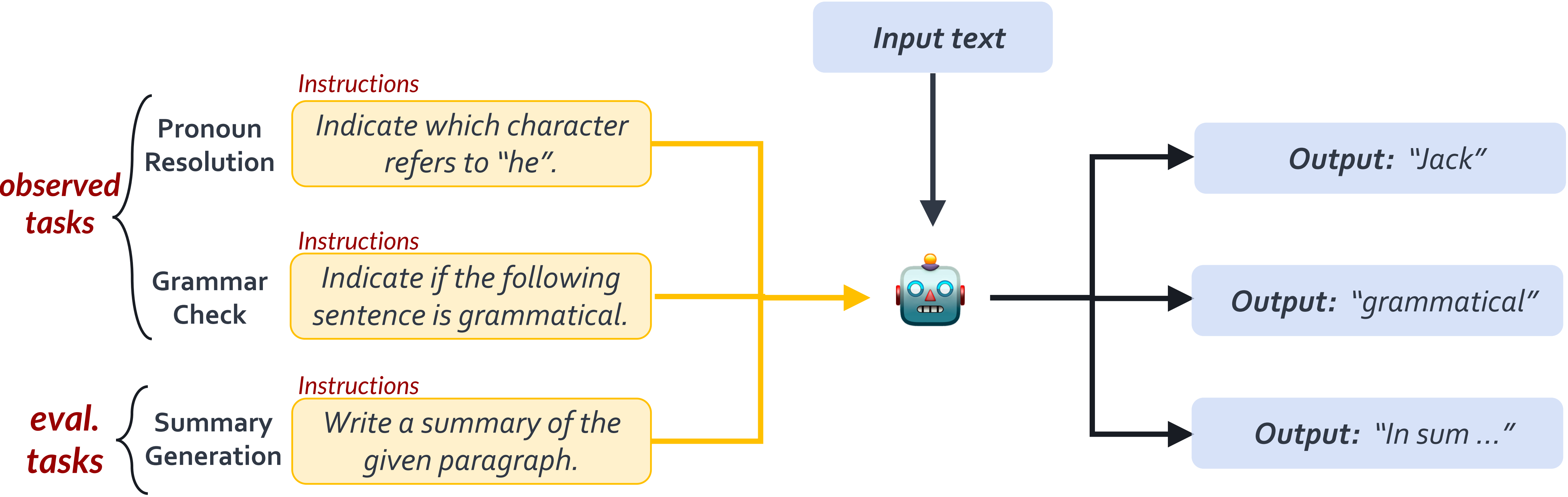
# Cross-Task Generalization



# Cross-Task Generalization

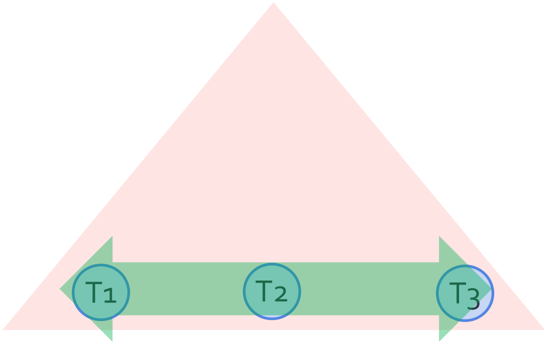


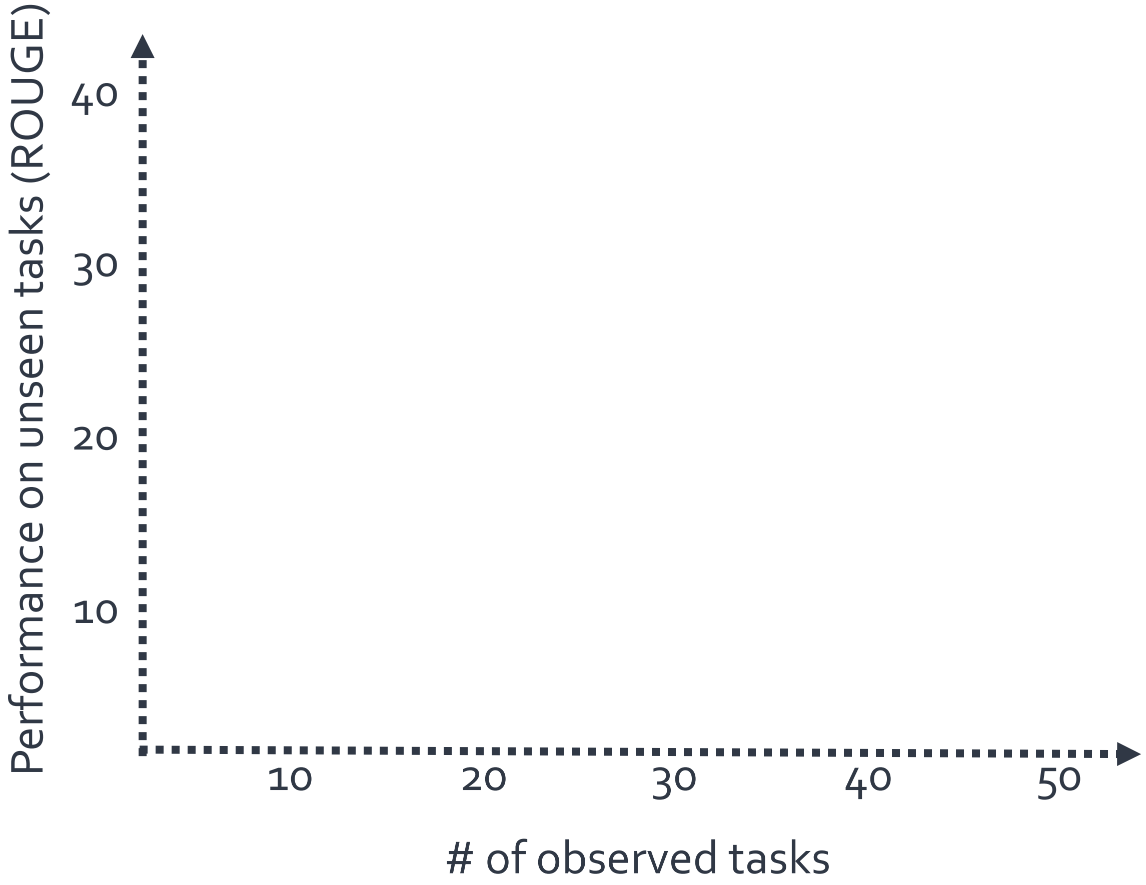




Done on "Natural Instructions" — a meta-dataset of tasks and their language instructions.

https://instructions.apps.allenai.org/

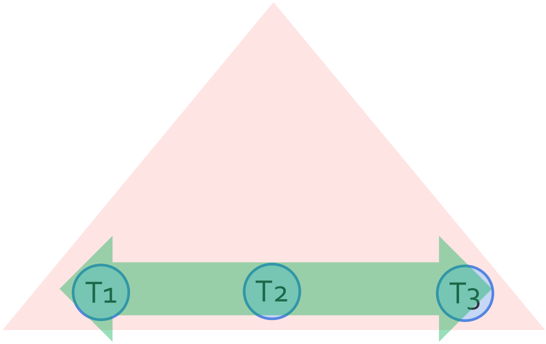
# Cross-Task Generalization

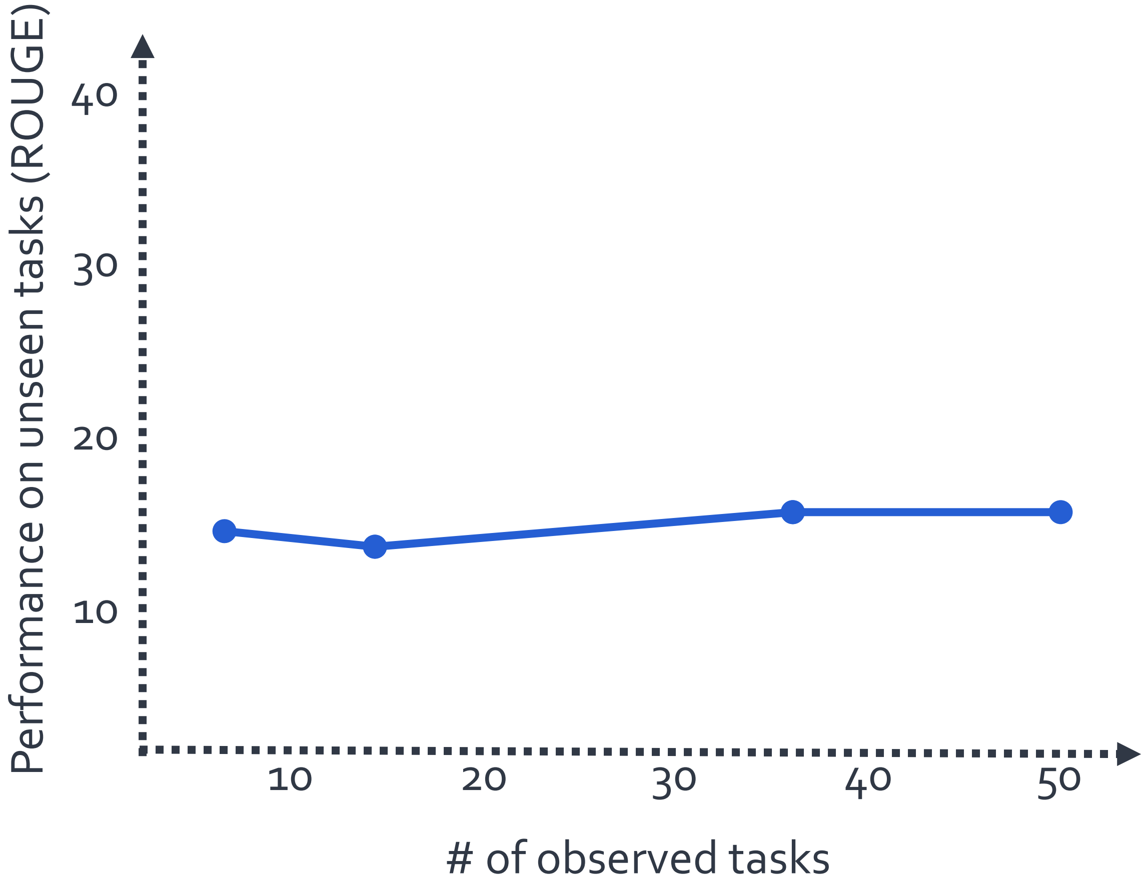



“Cross-Task Generalization via Natural Language Crowdsourcing Instructions.” Mishra, **Khashabi**, Baral, Hajishirzi, ACL’22

# Cross-Task Generalization

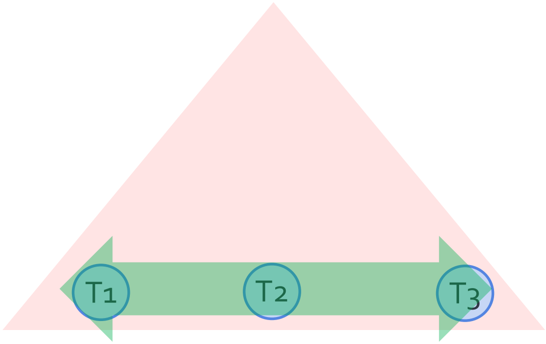


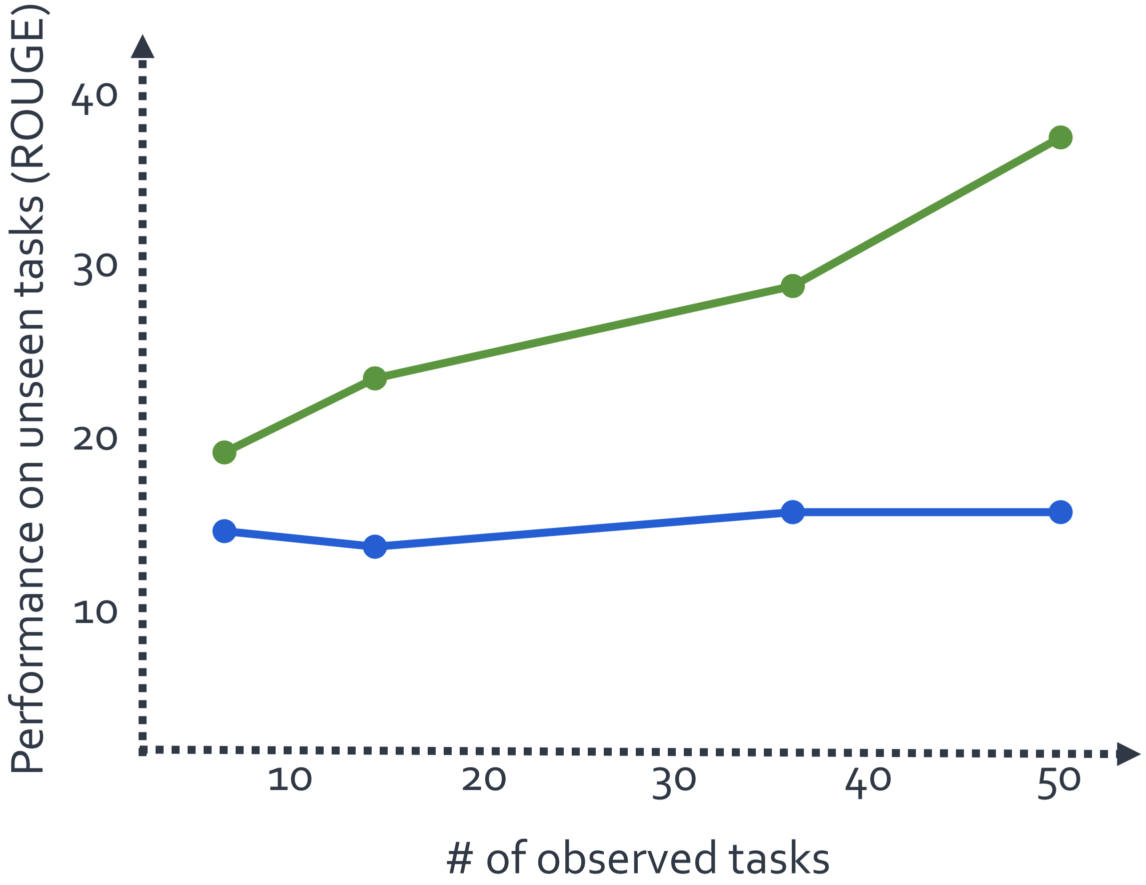

**without** Instructions

"Cross-Task Generalization via Natural Language Crowdsourcing Instructions." Mishra, **Khashabi**, Baral, Hajishirzi, ACL'22

# Cross-Task Generalization




- Performance on unseen tasks
  - **improves** with **more** observed tasks
  - when using **instructions**!

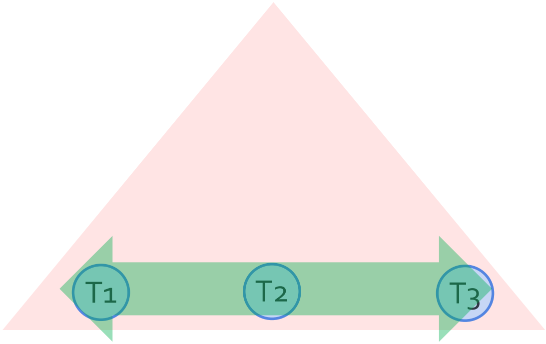



**with** Instructions

**without** Instructions

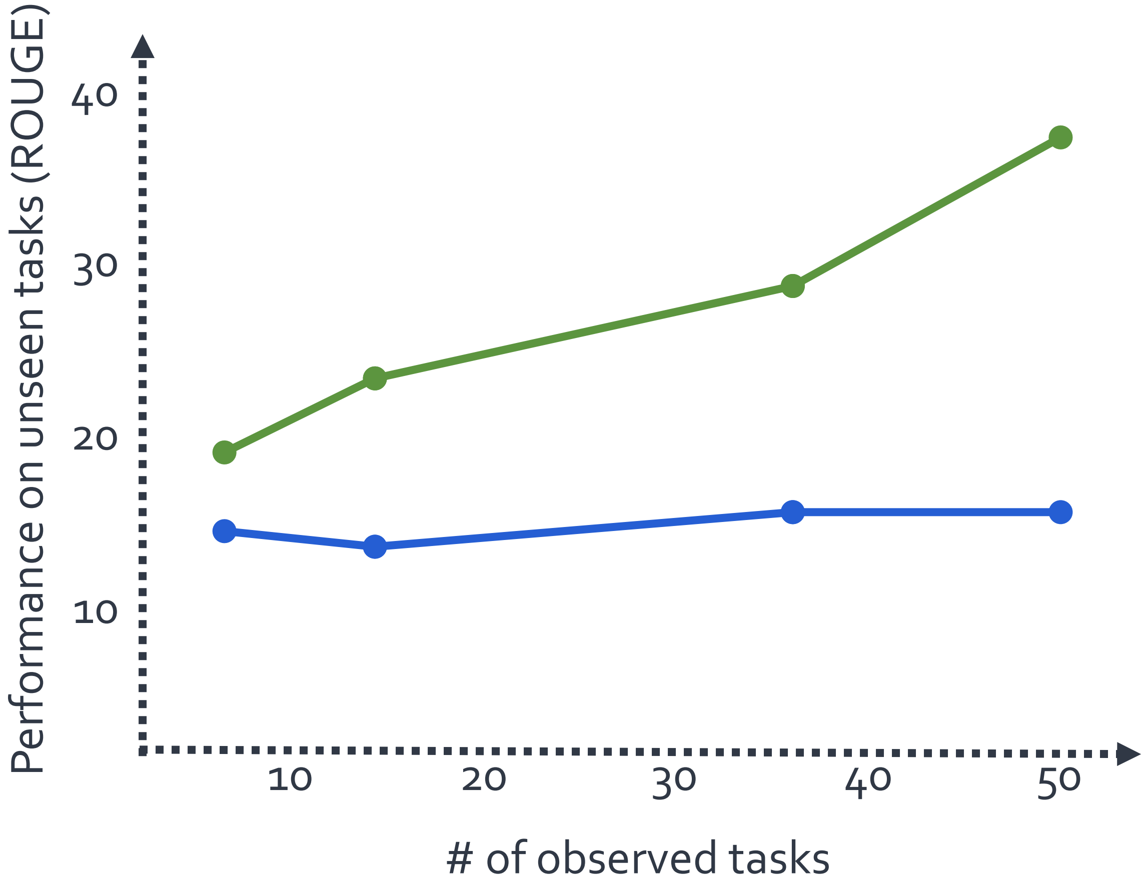
"Cross-Task Generalization via Natural Language Crowdsourcing Instructions." Mishra, **Khashabi**, Baral, Hajishirzi, ACL'22

# Cross-Task Generalization




- Performance on unseen tasks
  - **improves** with **more** observed tasks
  - when using **instructions**!

✅ **Hypothesis:** Task "instructions" are enough to induce sharedness among them.

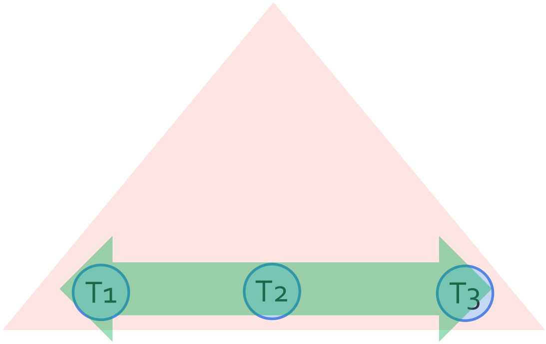



**with** Instructions

**without** Instructions

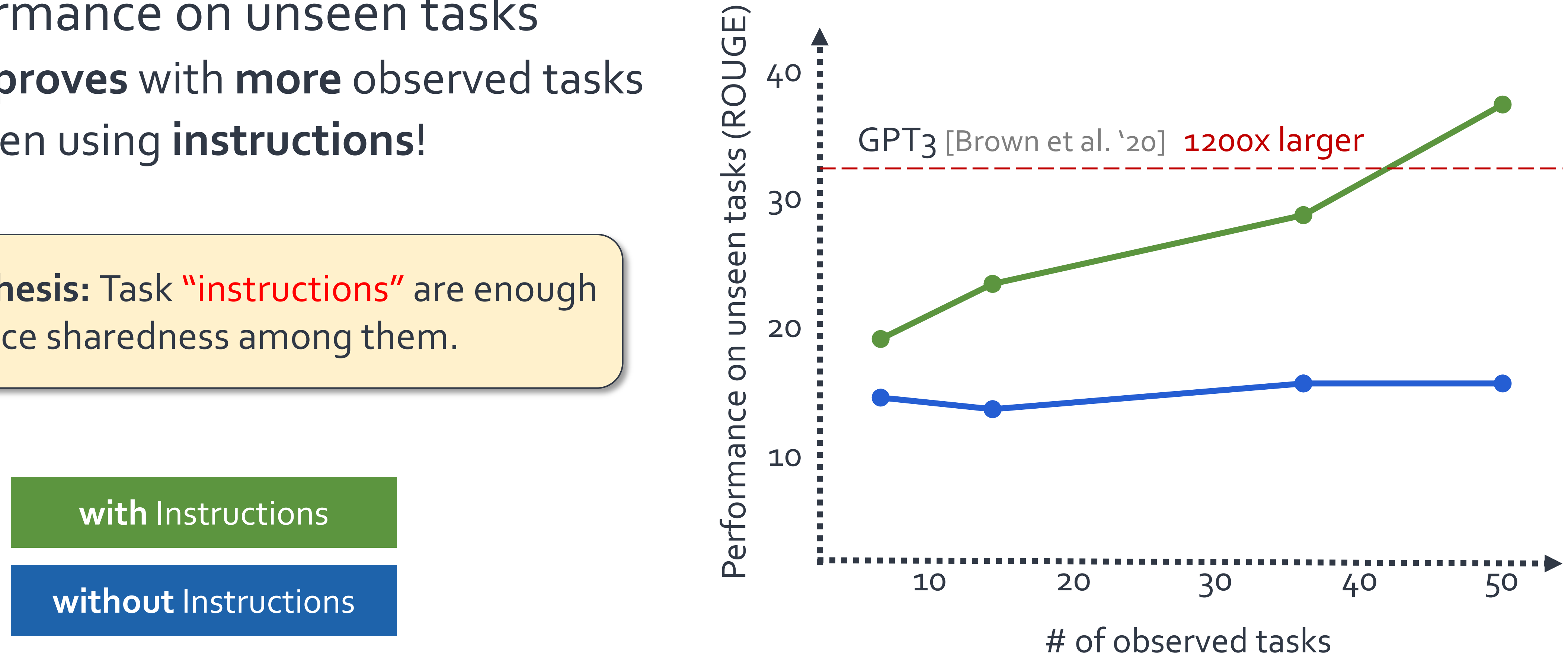
"Cross-Task Generalization via Natural Language Crowdsourcing Instructions." Mishra, **Khashabi**, Baral, Hajishirzi, ACL'22

# Cross-Task Generalization



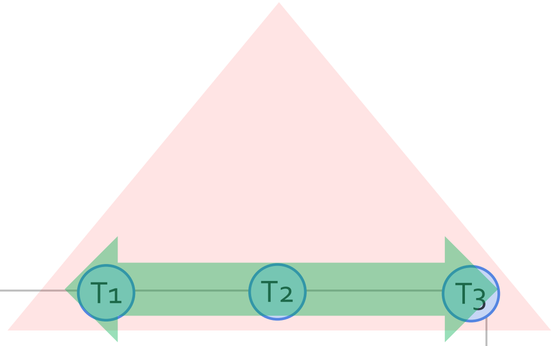

- Performance on unseen tasks
  - **improves** with **more** observed tasks
  - when using **instructions**!

**Hypothesis:** Task "instructions" are enough to induce sharedness among them.



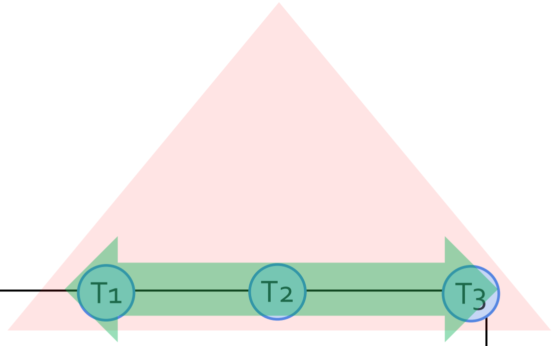

"Cross-Task Generalization via Natural Language Crowdsourcing Instructions." Mishra, **Khashabi**, Baral, Hajishirzi, ACL'22

# Summary



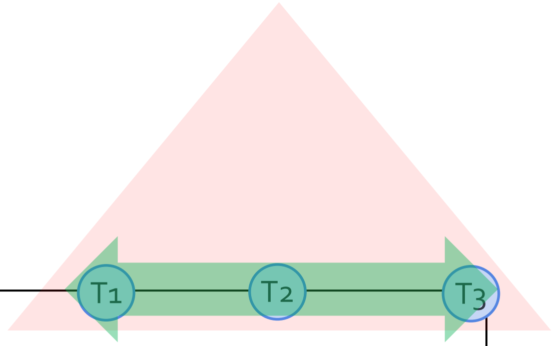

- **Motivating Question:** Can we build a single model that generalizes to unseen tasks?
- Generalization to unseen tasks improves when utilizing instructions.
- Toward systems w/ better "alignment" with human asks. [Christian '20]
- **Open questions:**
  - When does this generalization work? When does it not?

# Summary



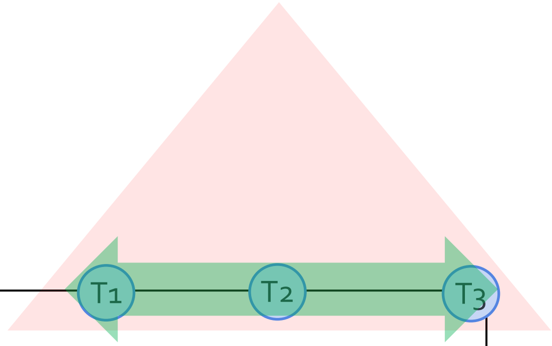

- **Motivating Question:** Can we build a single model that generalizes to unseen tasks?
- Generalization to unseen tasks improves when utilizing instructions.
- Toward systems w/ better "alignment" with human asks. [Christian '20]
- **Open questions:**
  - When does this generalization work? When does it not?

# Summary



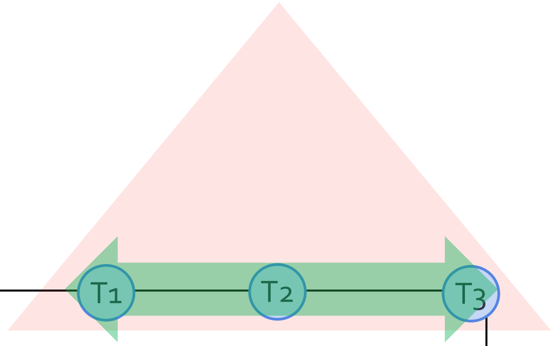

- **Motivating Question:** Can we build a single model that generalizes to unseen tasks?
- Generalization to unseen tasks improves when utilizing instructions.
- Toward systems w/ better "alignment" with human asks. [Christian '20]
- **Open questions:**
  - When does this generalization work? When does it not?

# Summary



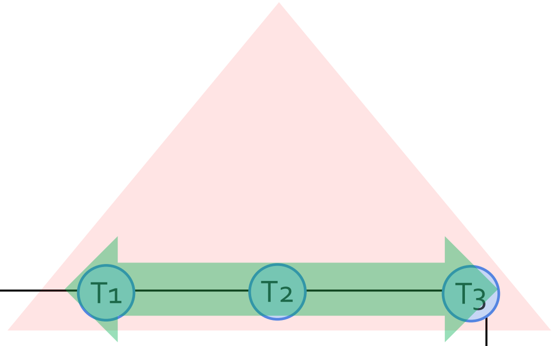

- **Motivating Question:** Can we build a single model that generalizes to unseen tasks?
- Generalization to unseen tasks improves when utilizing instructions.
- Toward systems w/ better "alignment" with human asks. [Christian '20]
- **Open questions:**
  - When does this generalization work? When does it not?

# Summary




- **Motivating Question:** Can we build a single model that generalizes to unseen tasks?
- Generalization to unseen tasks improves when utilizing instructions.
- Toward systems w/ better "alignment" with human asks. [Christian '20]
- **Open questions:**
  - When does this generalization work? When does it not?

# Summary




- **Motivating Question:** Can we build a single model that generalizes to unseen tasks?

- Generalization to unseen tasks improves when utilizing instructions.

- Toward systems w/ better "alignment" with human asks. [Christian '20]

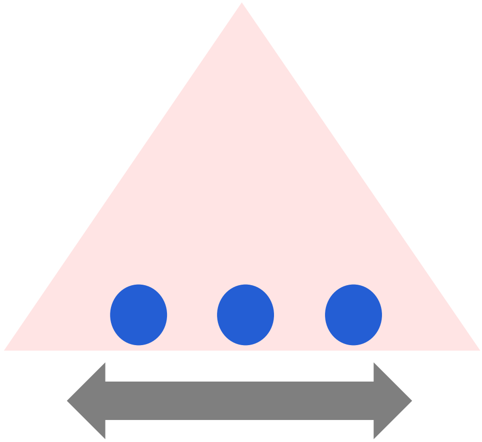
- **Open questions:**
  - When does this generalization work? When does it not?

# Talk Outline

Generality in “breadth” — tackling a *variety* of tasks

Generality in “depth” — tackling more *complex* tasks

*Future work:* Toward broad, interactive reasoning








UnifiedQA
*EMNLP Findings ’20*

Natural Instructions
*ACL ‘22*

ModularQA
*NAACL ’21*

# Talk Outline

Generality in "breadth" — tackling a variety of tasks

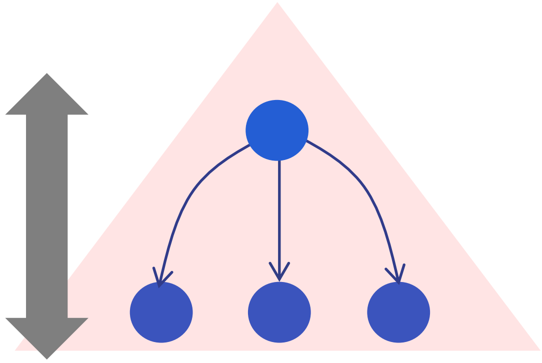
Generality in "depth" — tackling more complex tasks

*Future work:* Toward broad, interactive reasoning

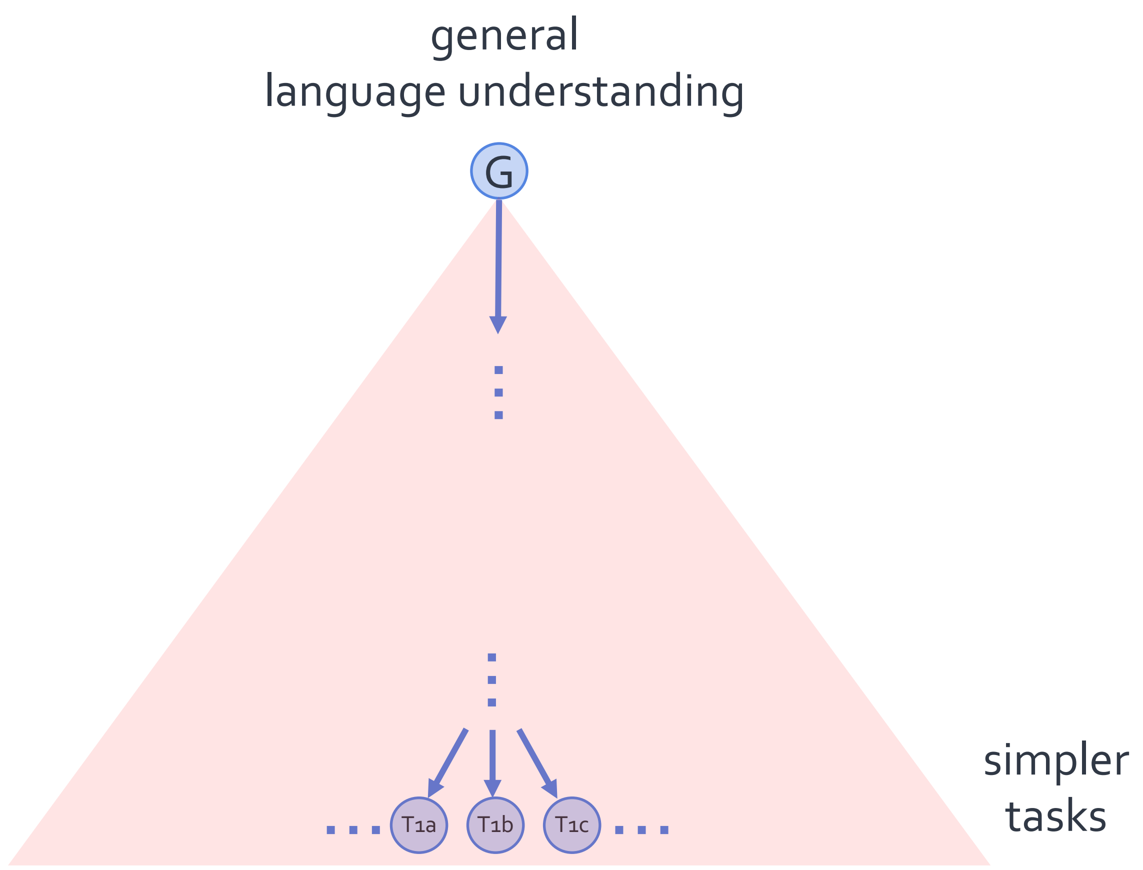


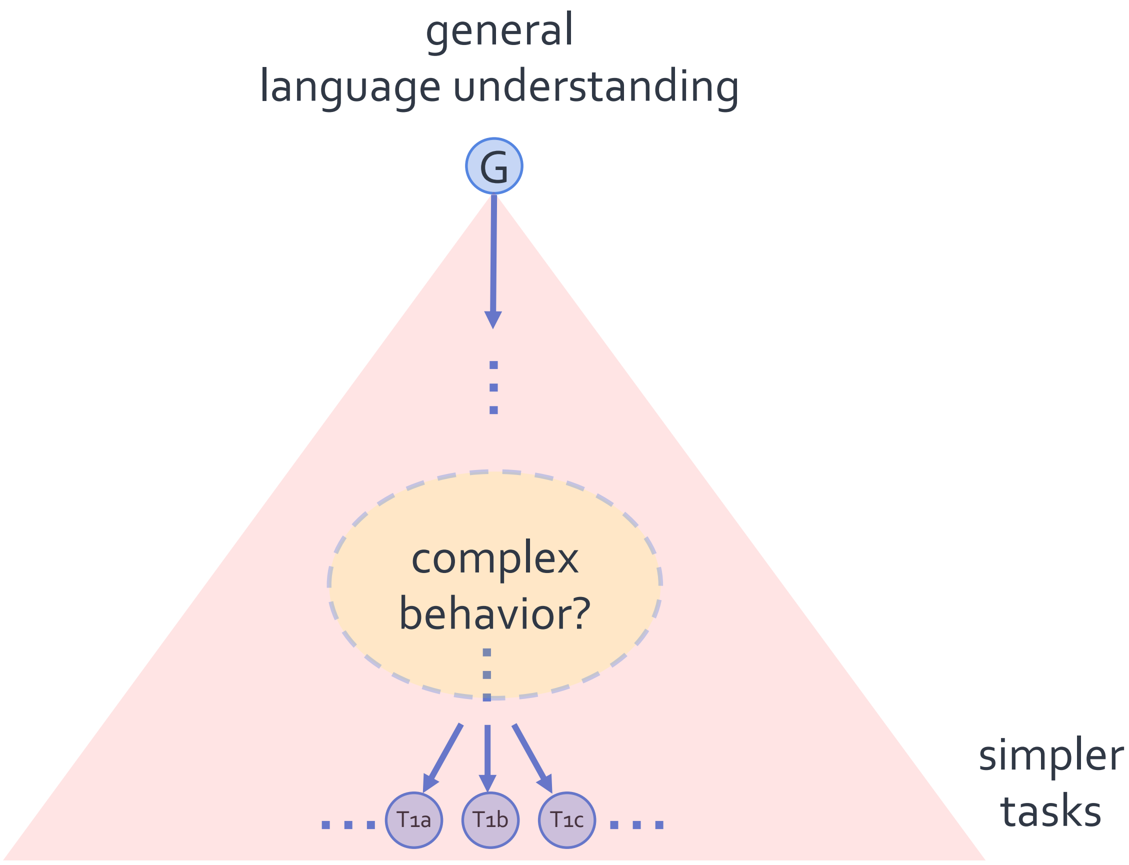
UnifiedQA *EMNLP Findings '20*

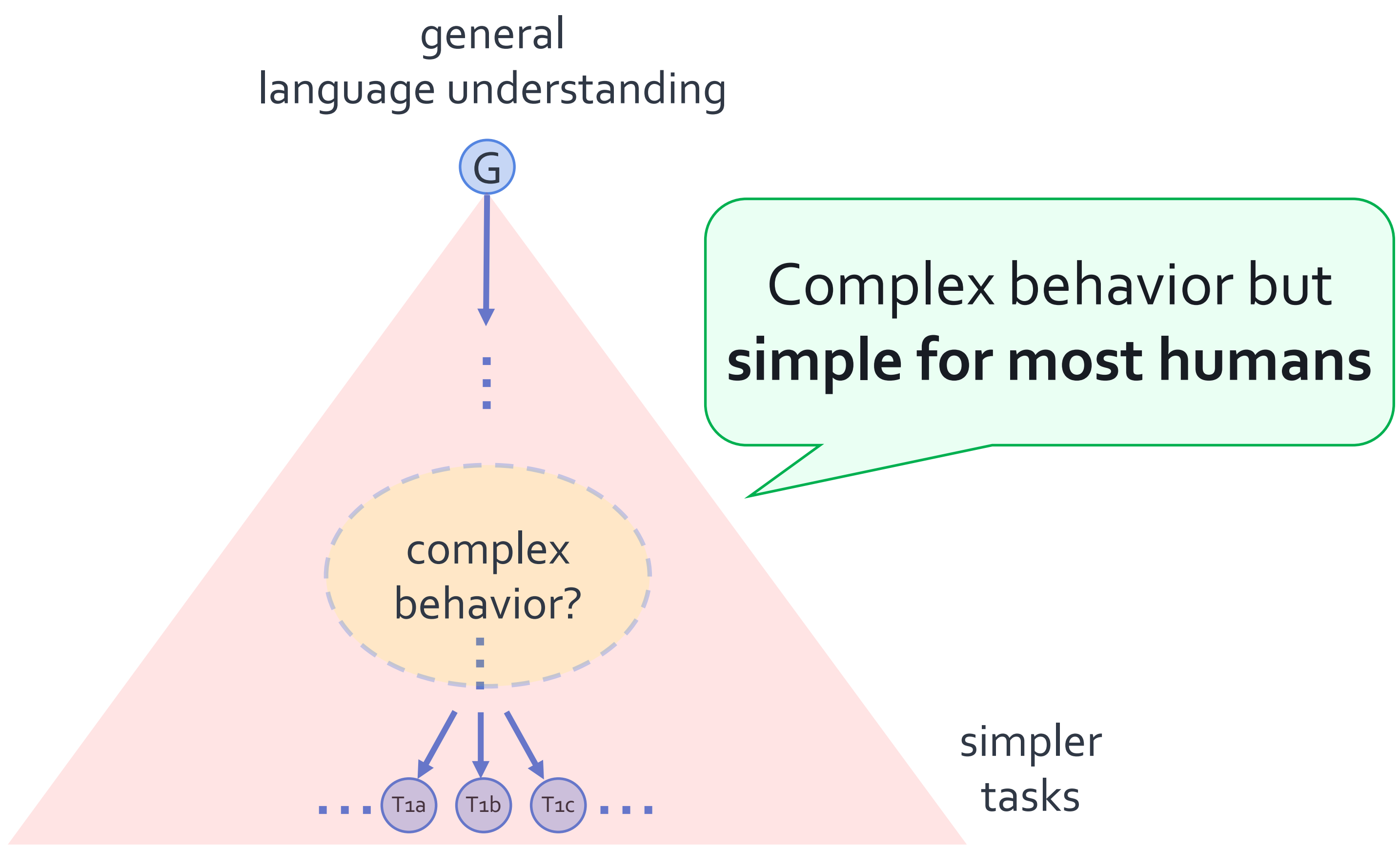
Natural Instructions *ACL '22*



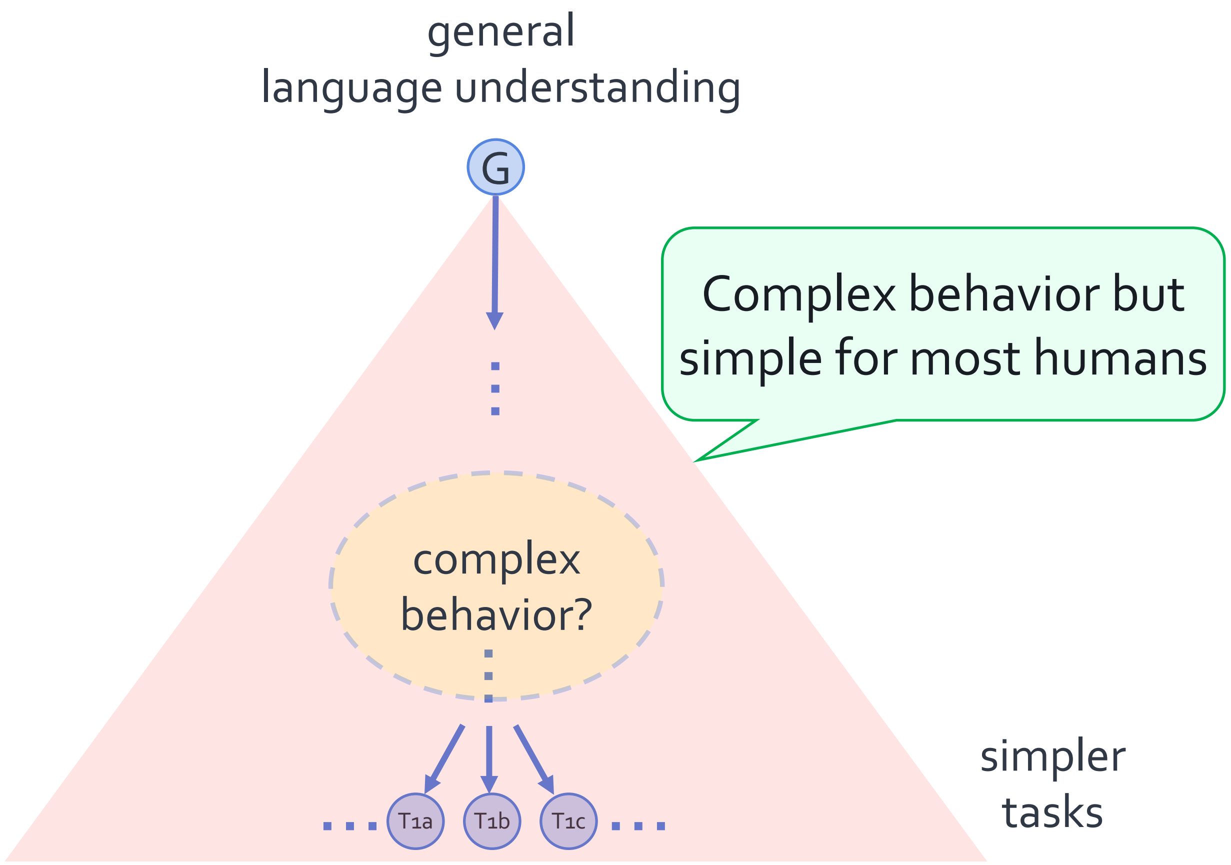
ModularQA *NAACL '21*


general
language understanding
G
T1a
T1b
T1c
simpler
tasks

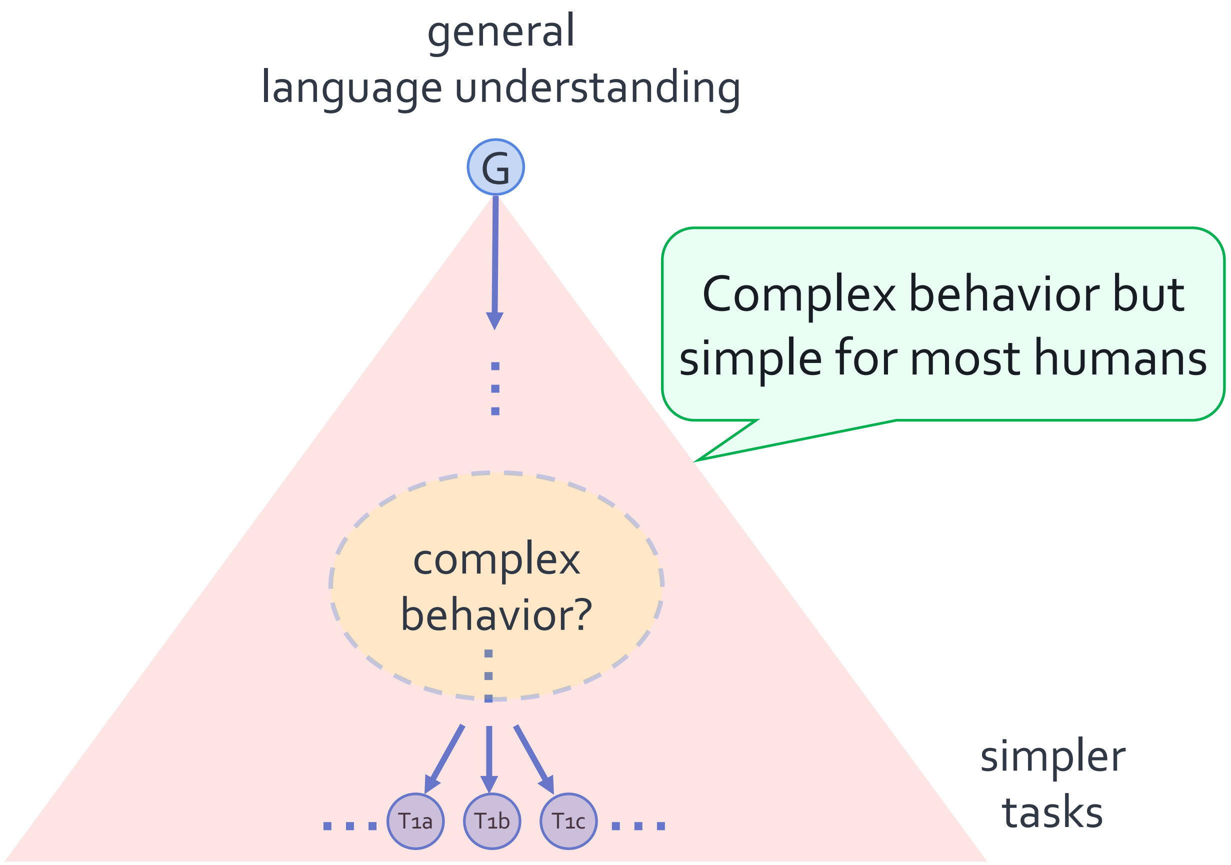

general
language understanding
G
complex
behavior?
T1a
T1b
T1c
simpler
tasks


general
language understanding
G
complex
behavior?
T1a
T1b
T1c
simpler
tasks
Complex behavior but
**simple for most humans**

- **Interactivity** — can lead to complex phenomena, through simple steps.



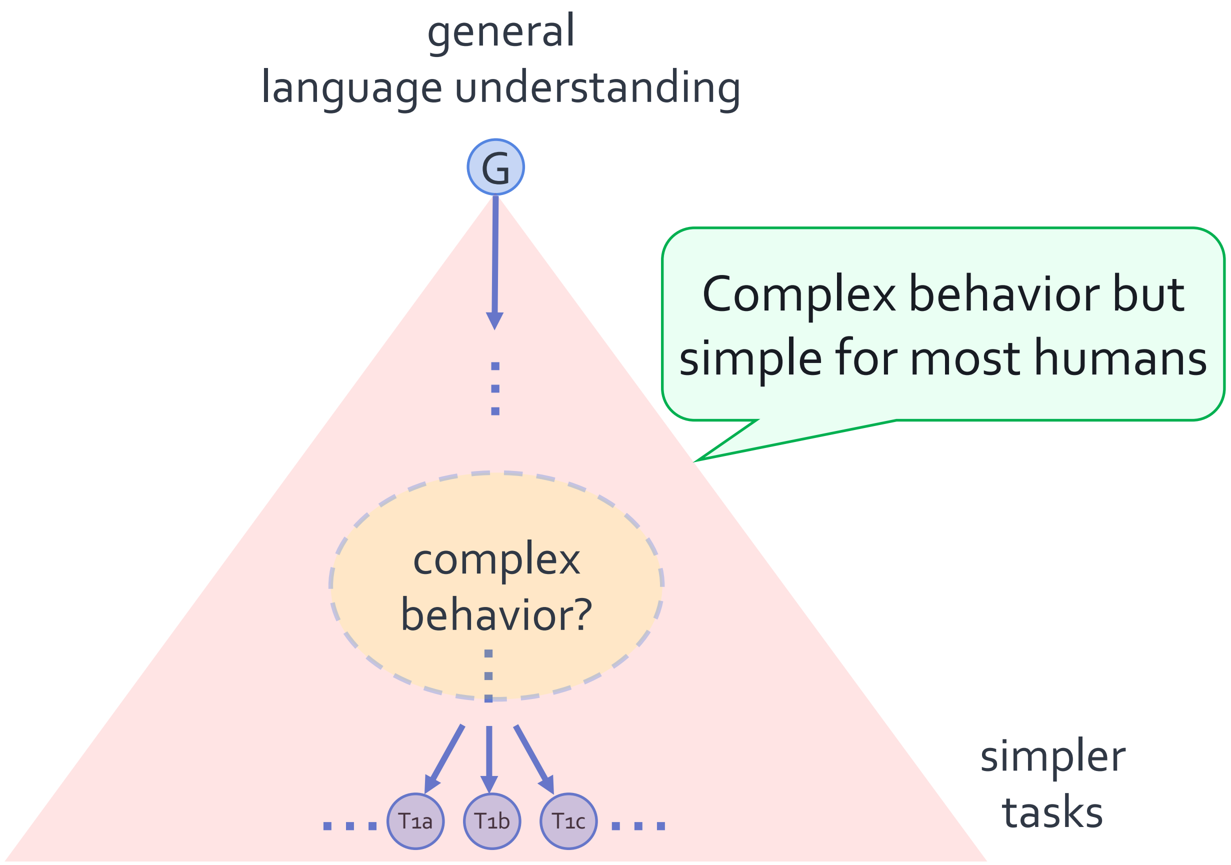
- **Interactivity —** can lead to complex phenomena, through simple steps.

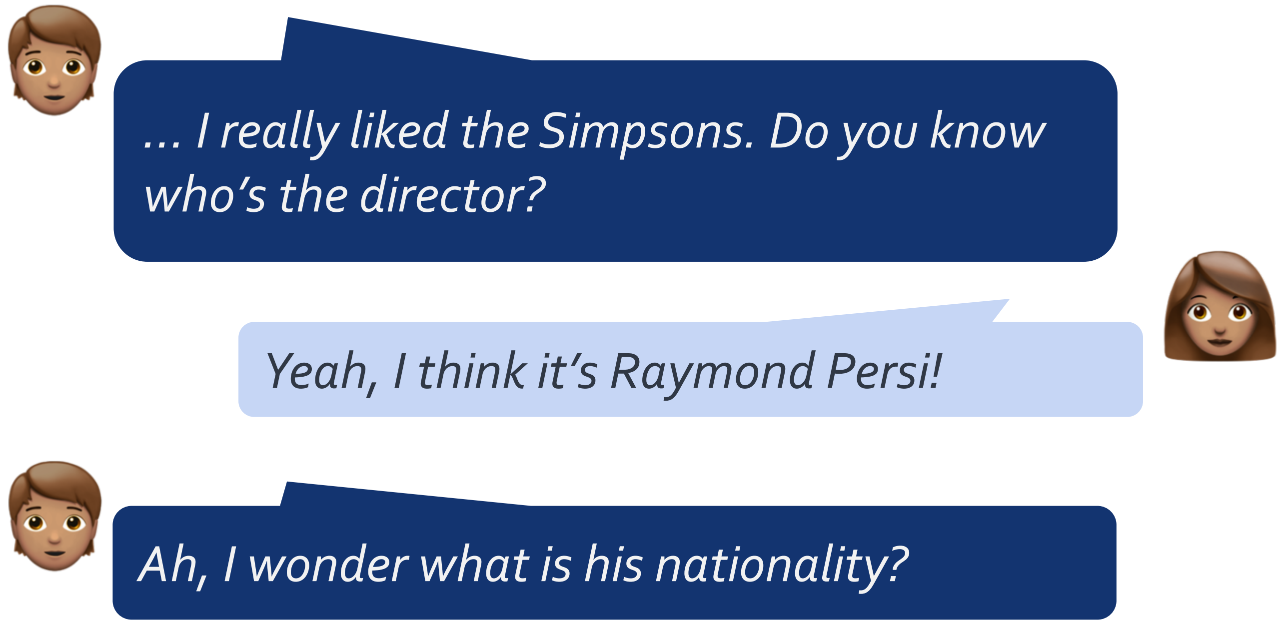
- **Interactivity** — can lead to complex phenomena, through simple steps.

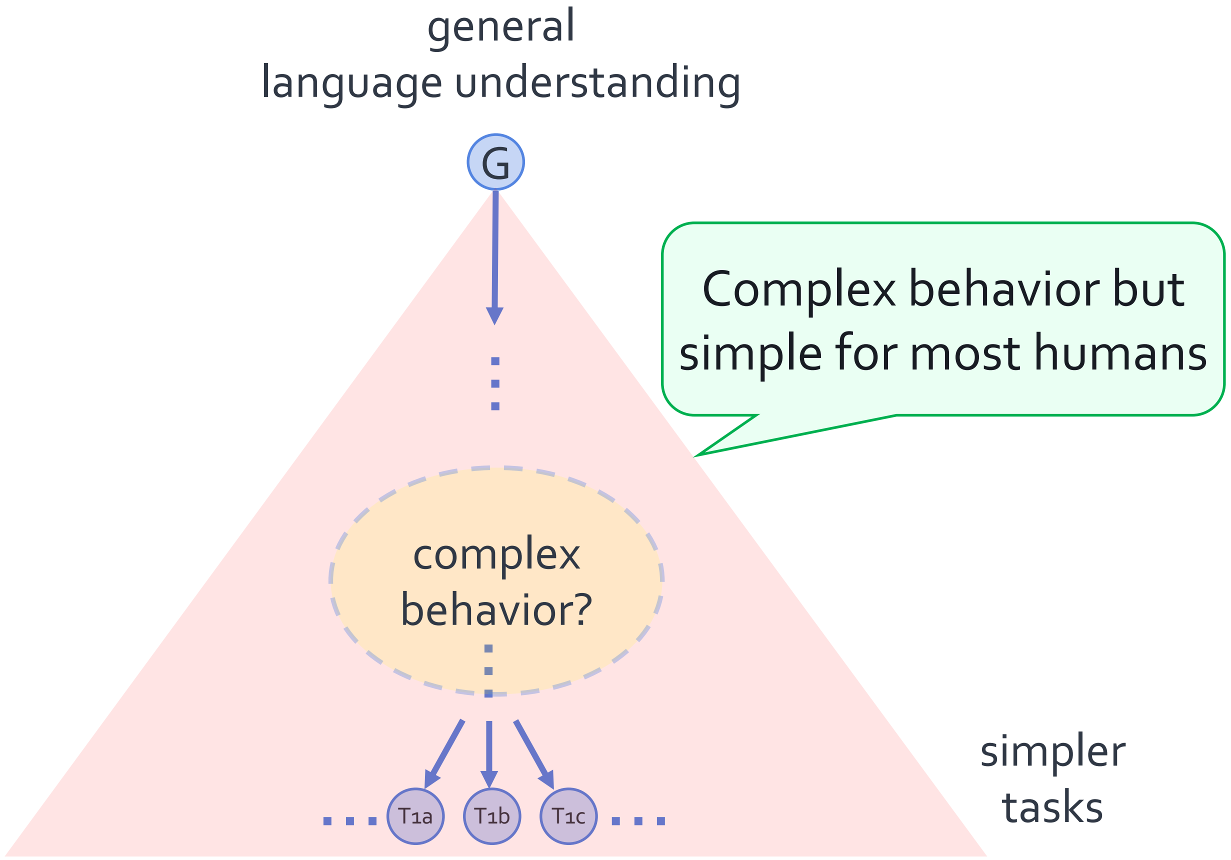
*… I really liked the Simpsons. Do you know who's the director?*

*Yeah, I think it's Raymond Persi!*



- **Interactivity** — can lead to complex phenomena, through simple steps.



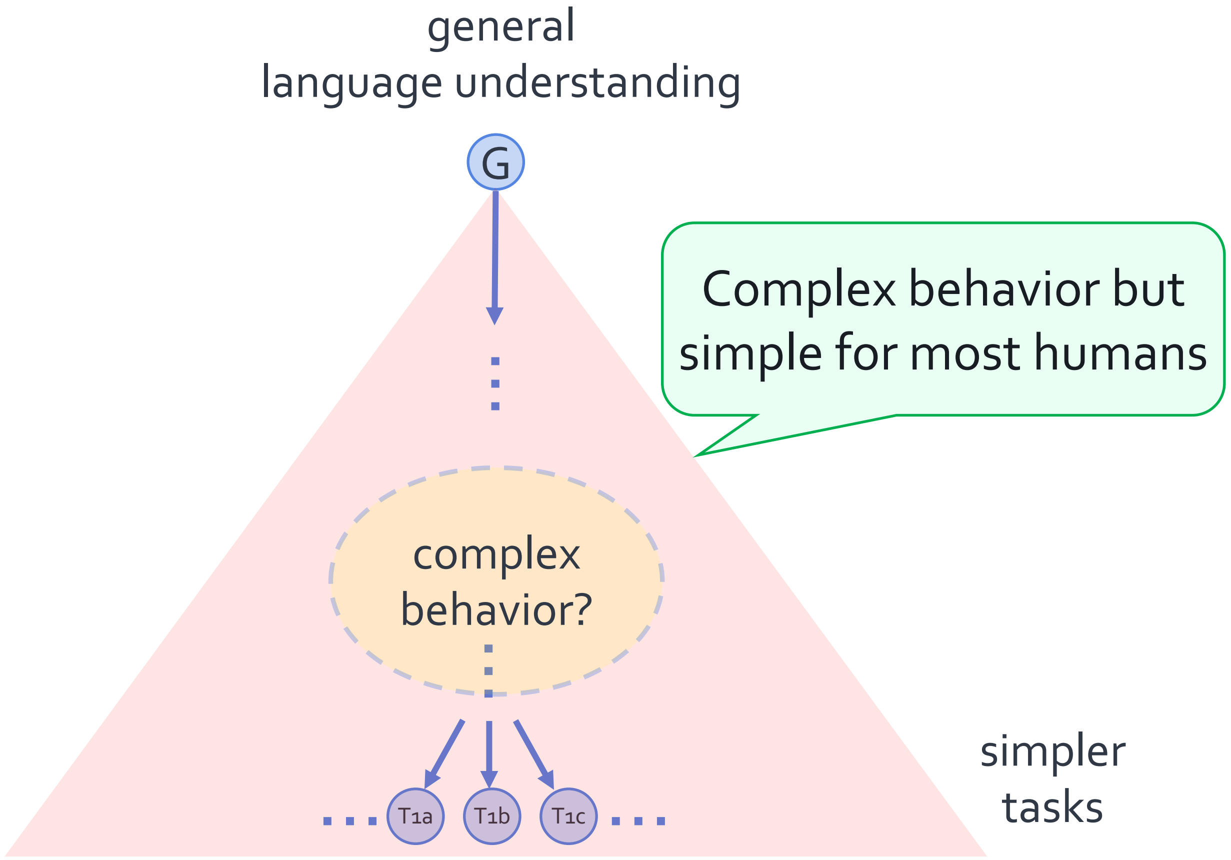
- **Interactivity** — can lead to complex phenomena, through simple steps.

*... I really liked the Simpsons. Do you know who's the director?*

*Yeah, I think it's Raymond Persi!*

*Ah, I wonder what is his nationality?*

*According to Wikipedia he's American.*



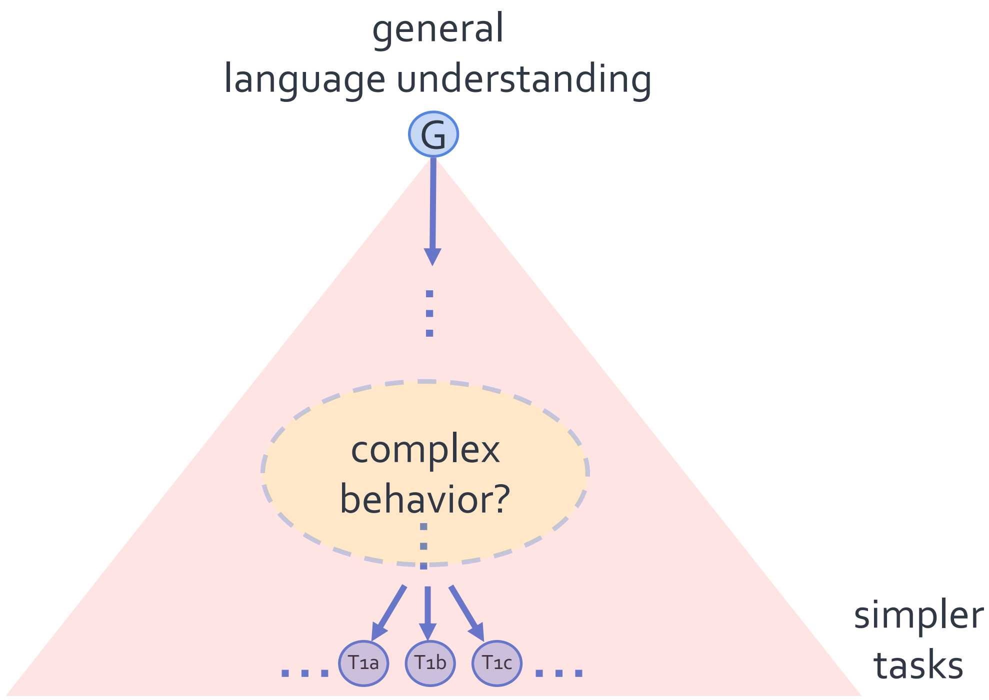
- **Interactivity** — can lead to complex phenomena, through simple steps.

*… I really liked the Simpsons. Do you know who's the director?*

*Yeah, I thinks it's Raymond Persi!*

*Ah, I wonder what is his nationality?*

*According to Wikipedia he's American.*



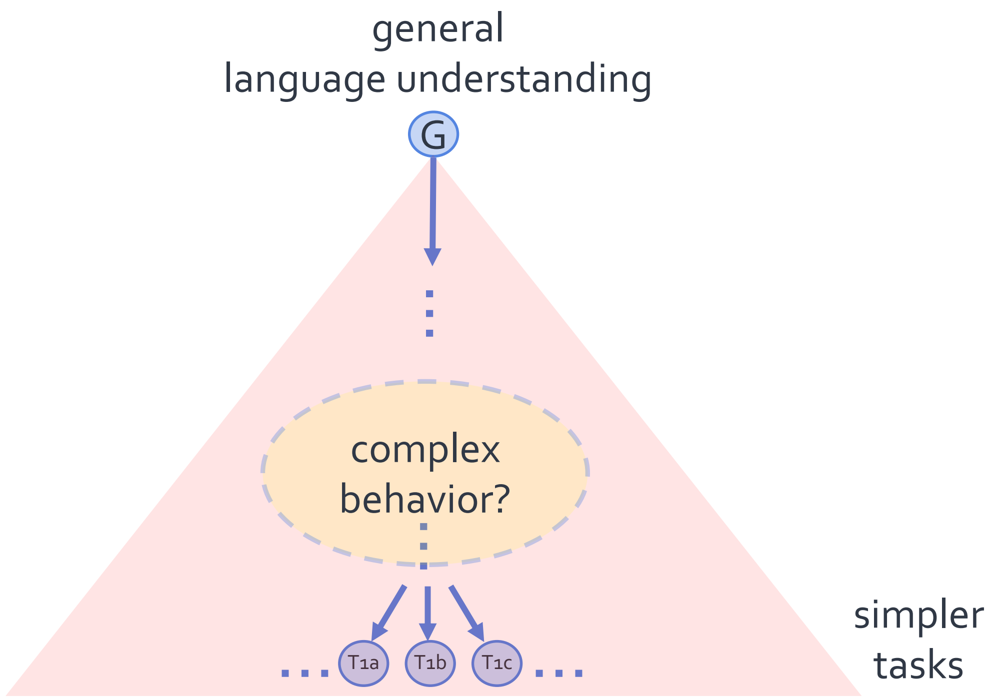
- **Interactivity** — can lead to complex phenomena, through simple steps.

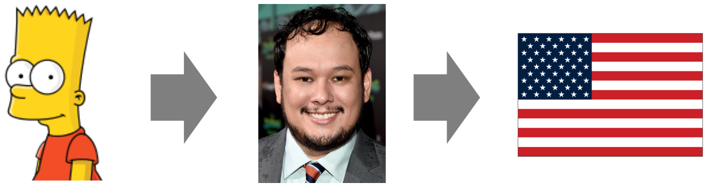
*… I really liked the Simpsons. Do you know who’s the director?*

*Yeah, I thinks it’s Raymond Persi!*

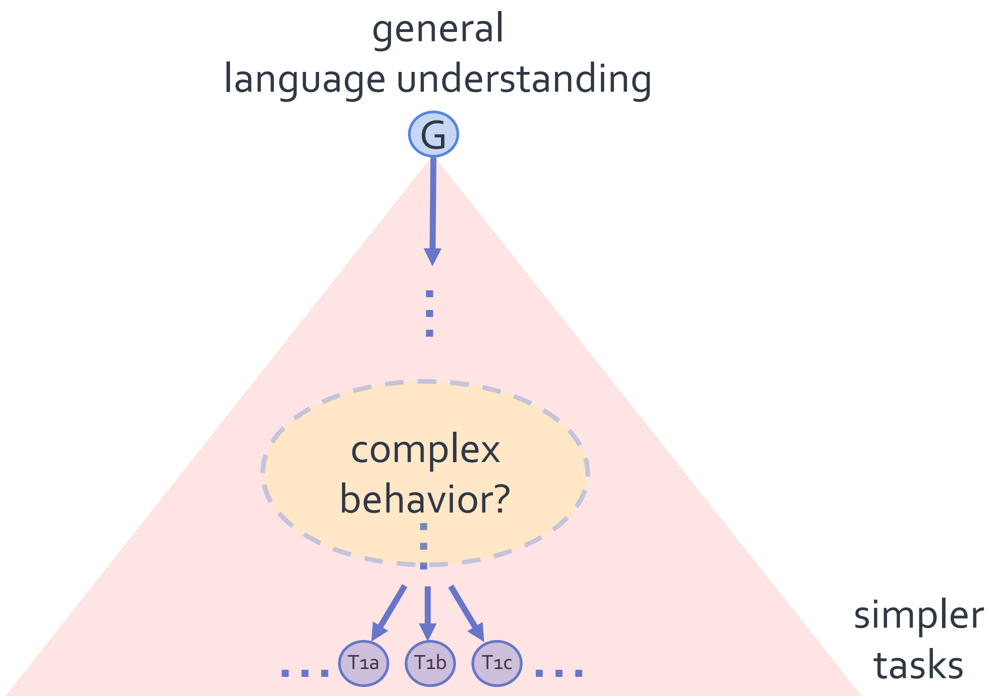
*Ah, I wonder what is his nationality?*

*According to Wikipedia he’s American.*



- **Interactivity** — can lead to complex phenomena, through simple steps.

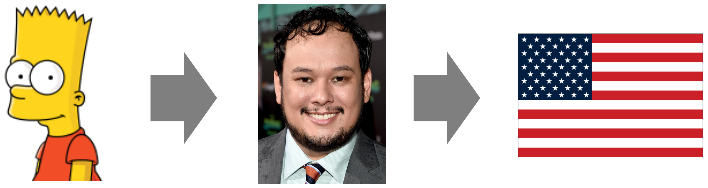



*... I really liked the Simpsons. Do you know who's the director?*

*Yeah, I thinks it's Raymond Persi!*

*Ah, I wonder what is his nationality?*

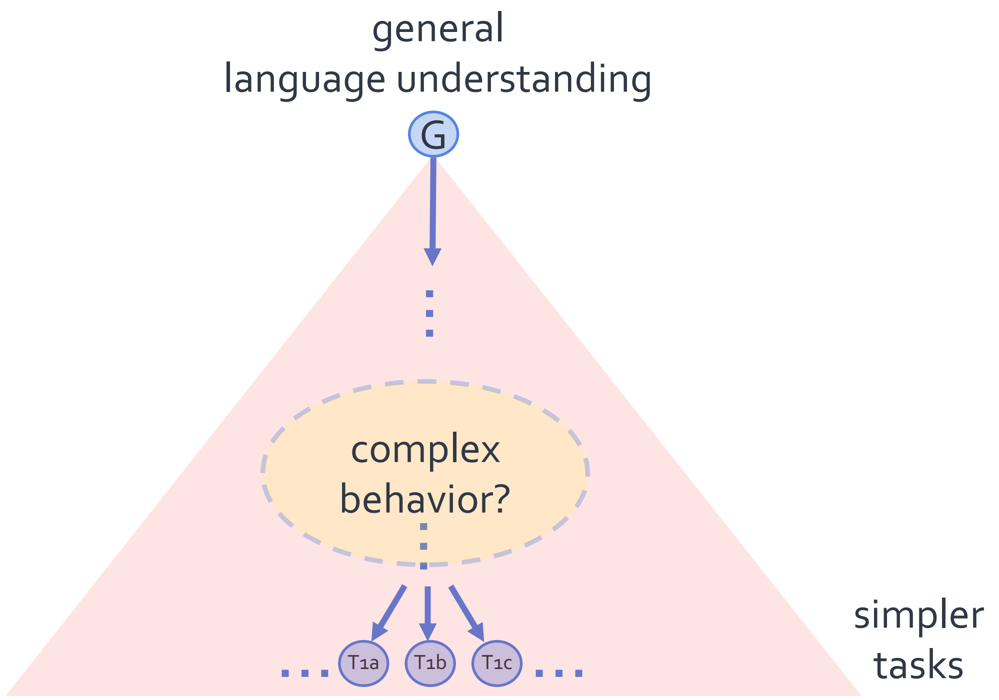
*According to Wikipedia he's American.*

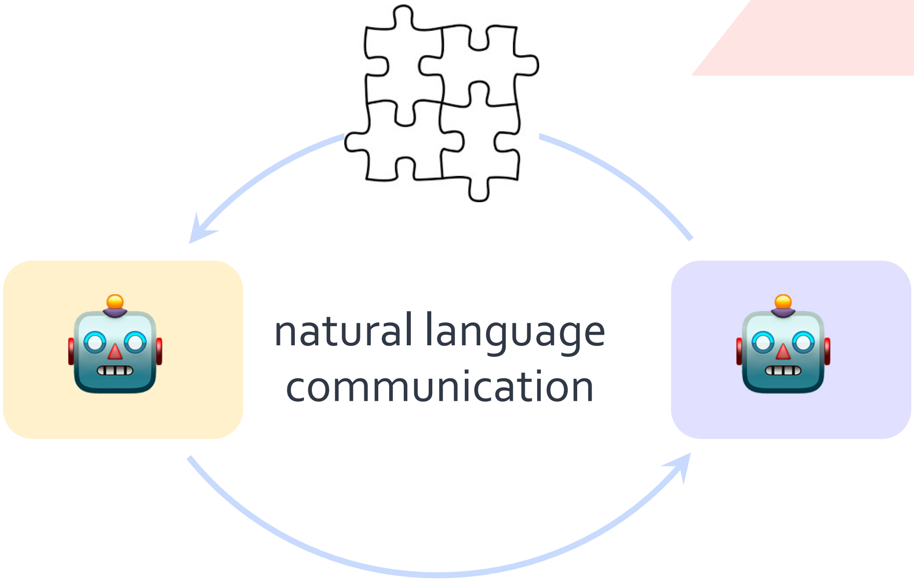


*complex question*

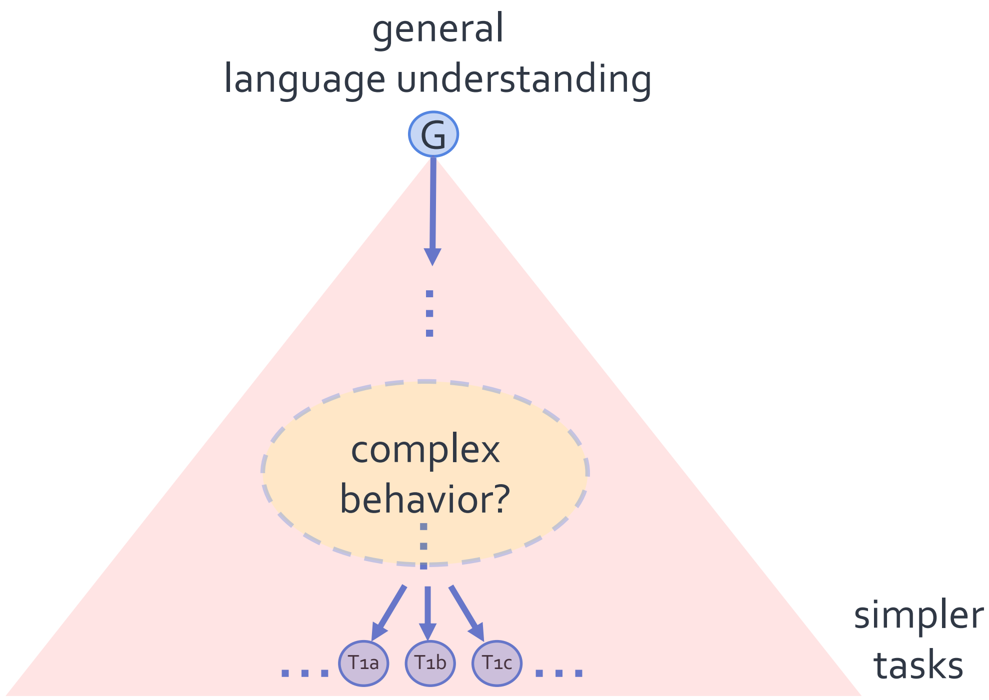
*"What is the nationality of the Simpsons director?"*

- **Interactivity** — can lead to complex phenomena, through simple steps.
- **Setup:**
    - Communications between models
    - Goal oriented

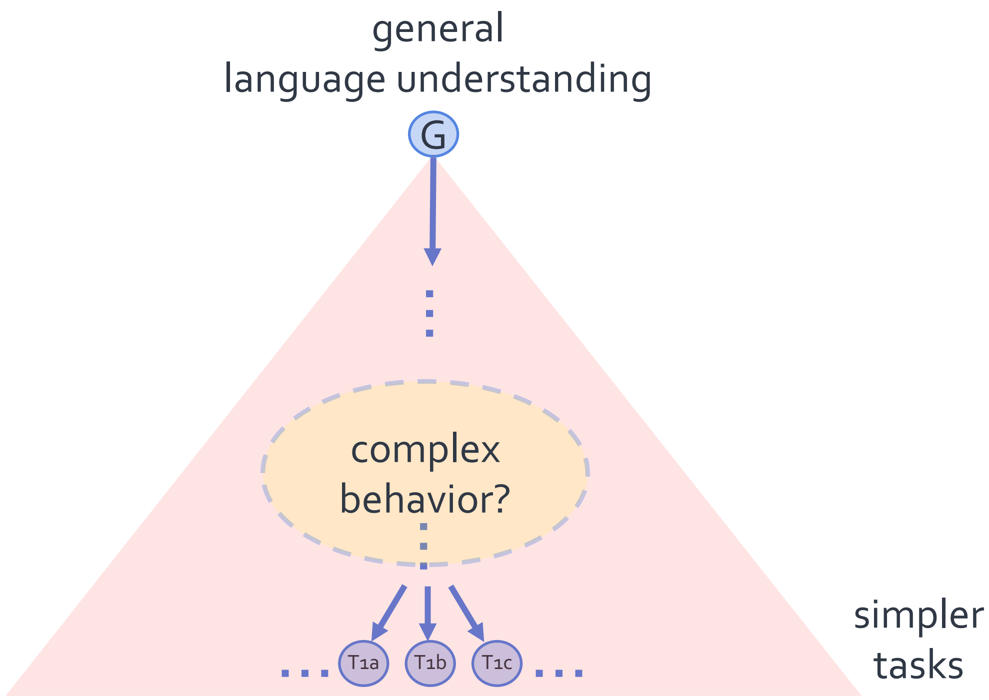


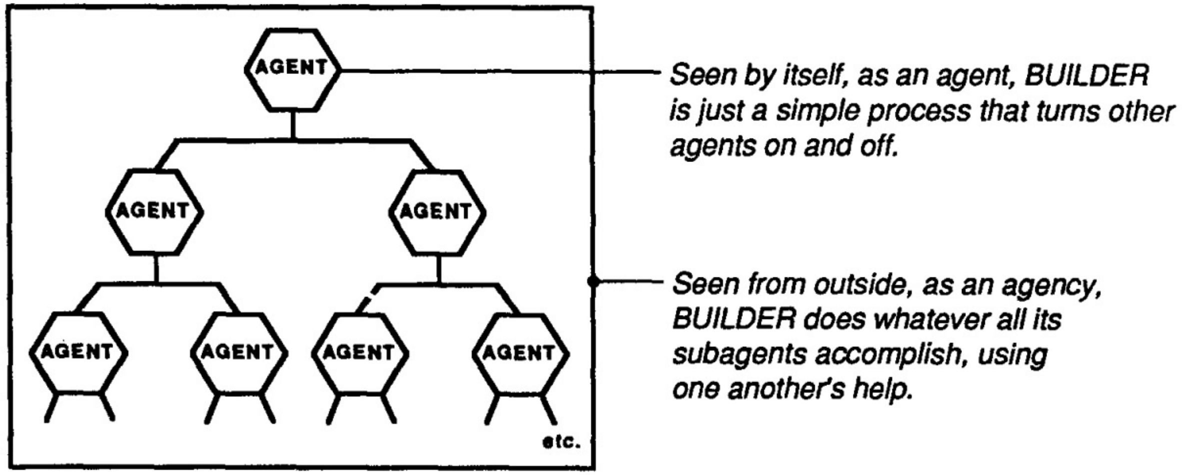



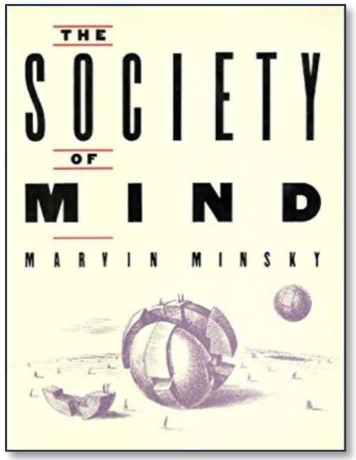
- **Interactivity** — can lead to complex phenomena, through simple steps.
- **Setup:**
  - Communications between models
  - Goal oriented



- **Interactivity —** can lead to complex phenomena, through simple steps.
- **Setup:**
  - Communications between models
  - Goal oriented




*"human intelligence … built up from the interactions of simple parts called agents"*

[Minsky, '70s]

# Text Modular Networks

Interactive communication
for solving complex questions

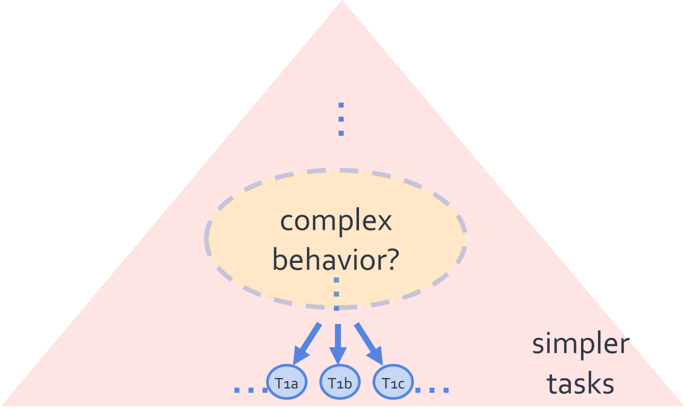
Tushar Khot , **Daniel Khashabi**, Kyle Richardson
Peter Clark and Ashish Sabharwal

*NAACL 2021*

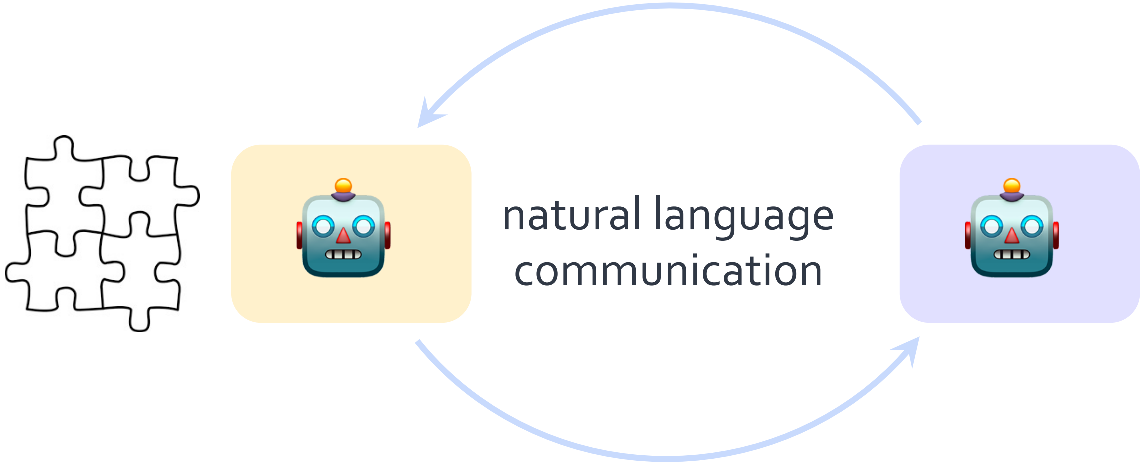




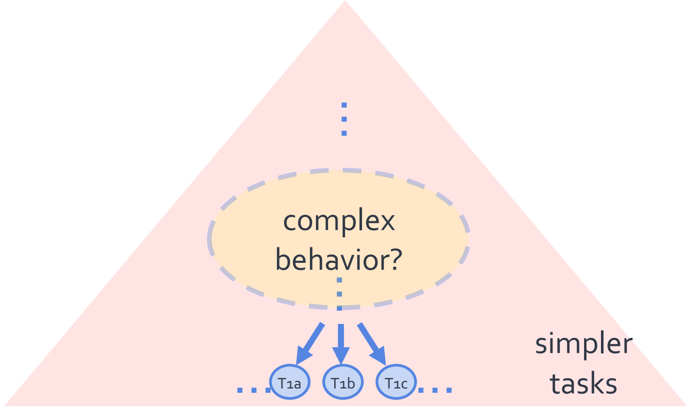
# Complex Problem Solving As Communication



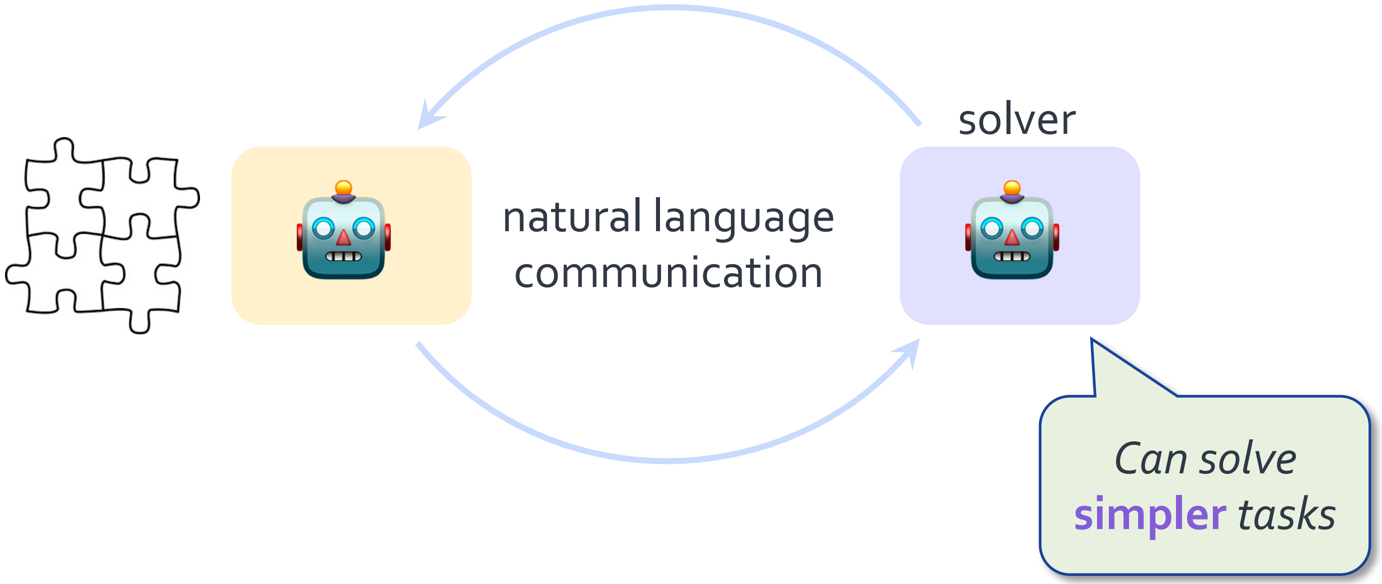

- **Setup:**
  - Communications between models
  - Goal oriented



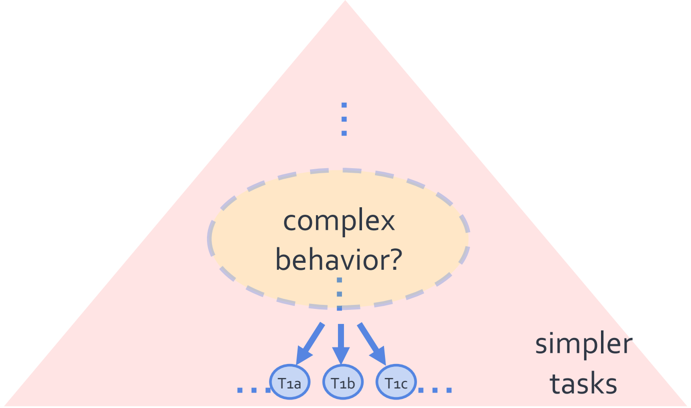
# Complex Problem Solving As Communication



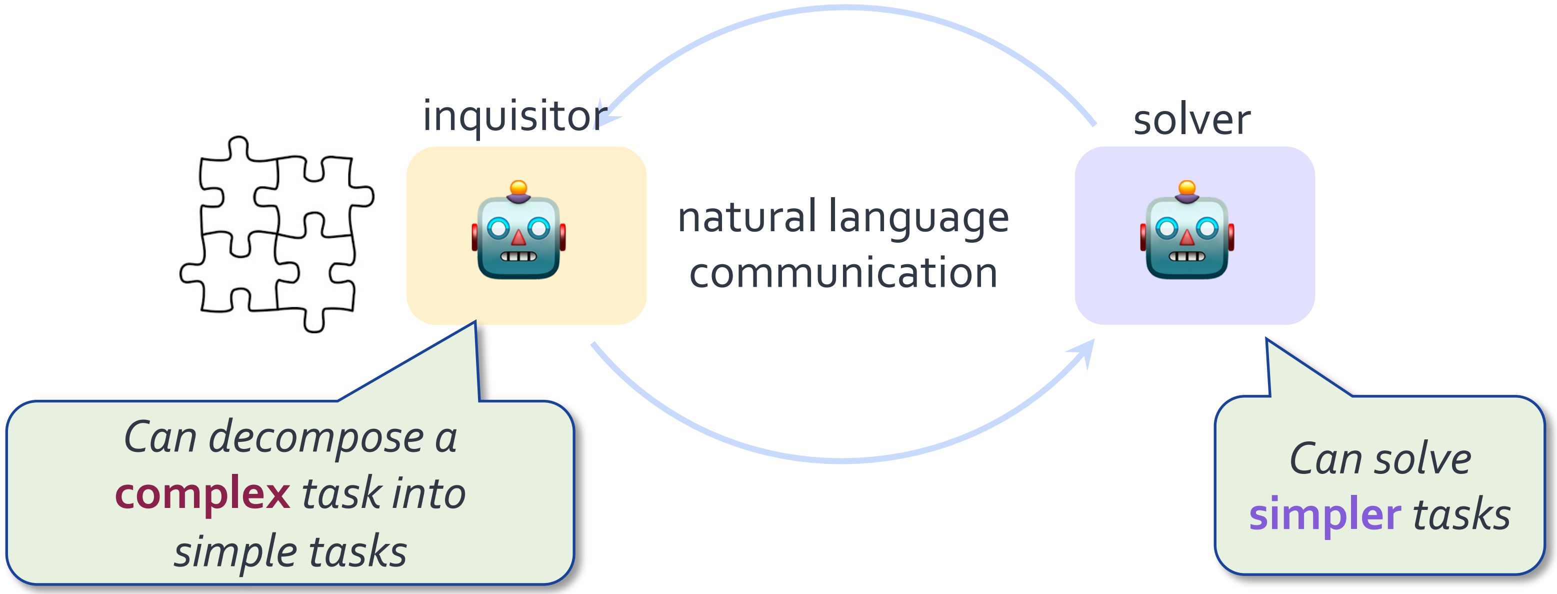

- **Setup:**
  - Communications between models
  - Goal oriented
  - Roles: solver and …



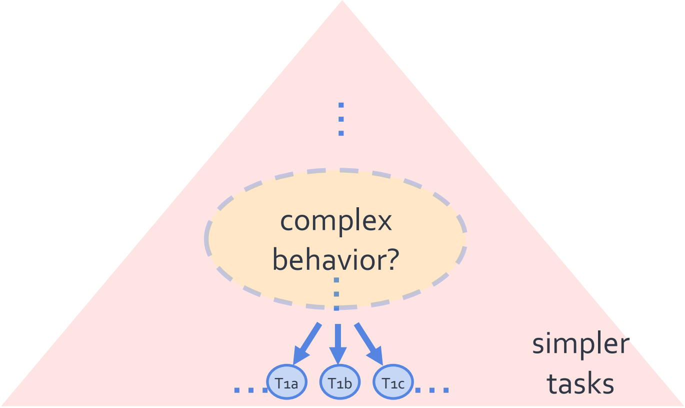
# Complex Problem Solving As Communication



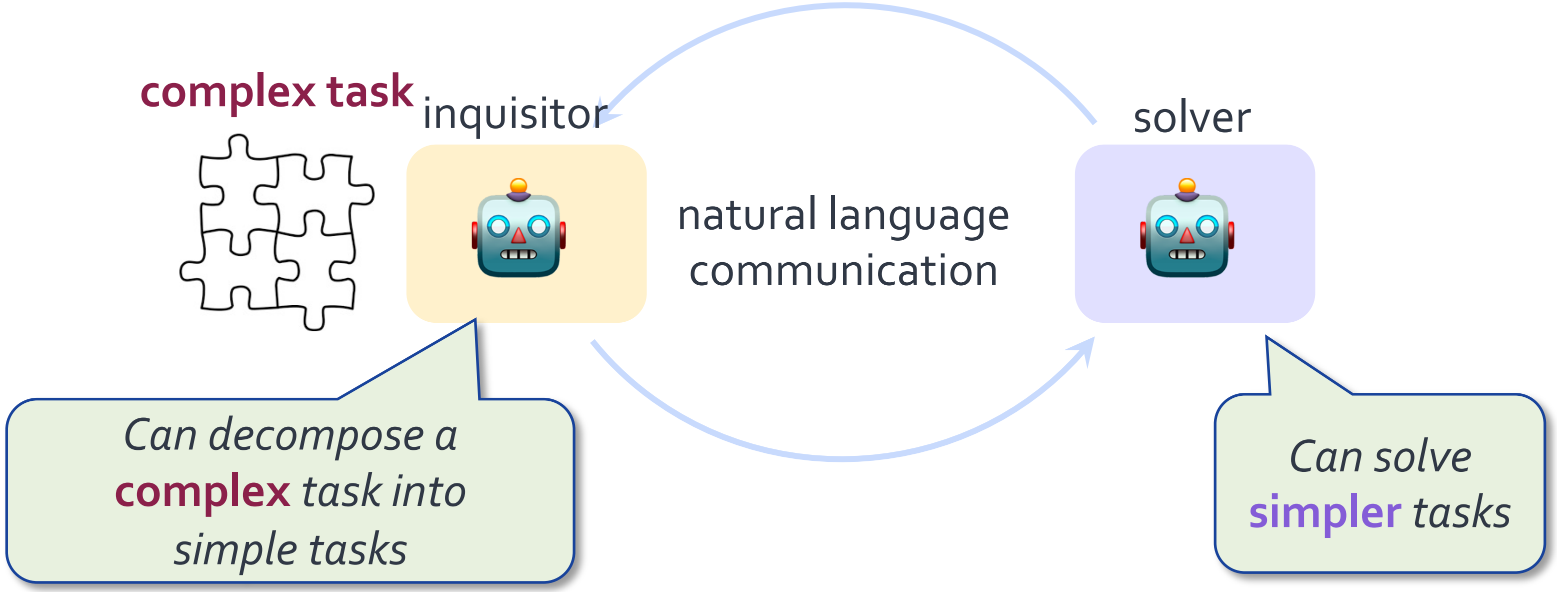

- **Setup:**
  - Communications between models
  - Goal oriented
  - Roles: solver and inquisitor



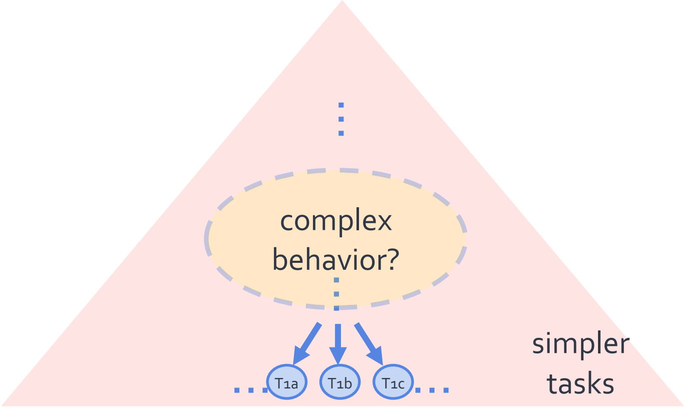
# Complex Problem Solving As Communication



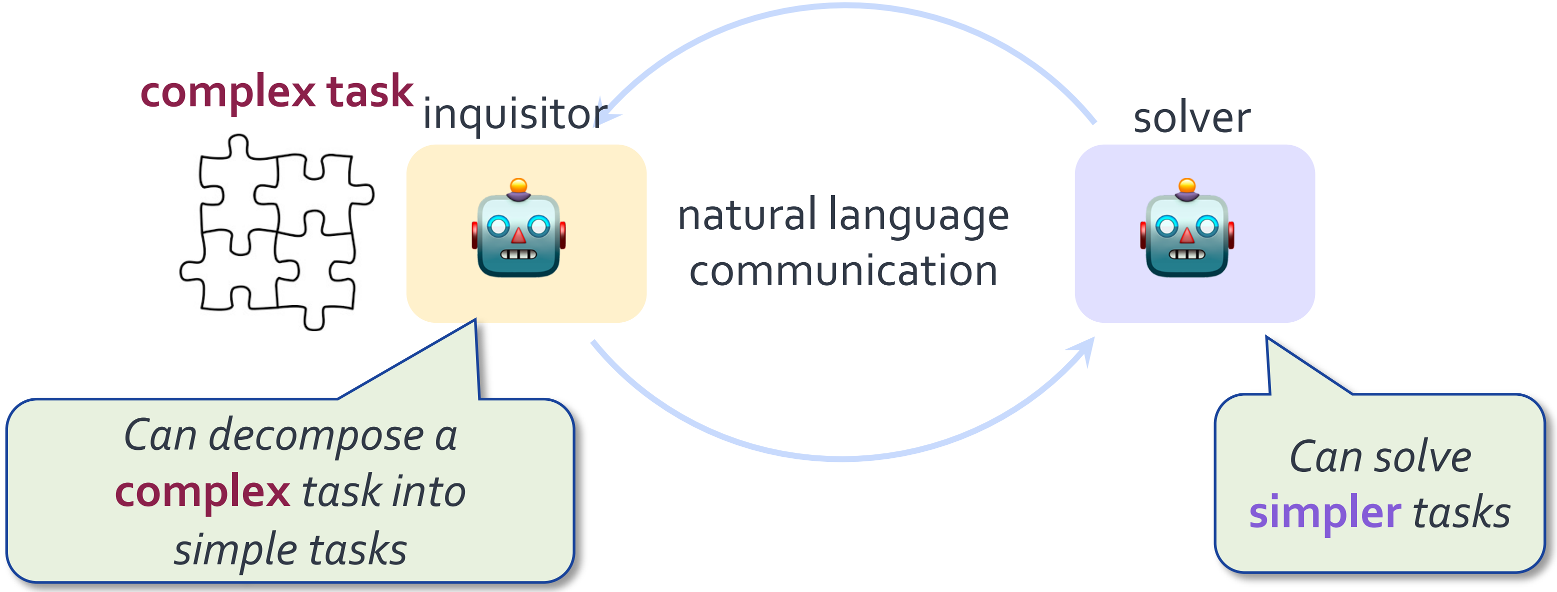

- **Setup:**
  - Communications between models
  - Goal oriented
  - Roles: solver and inquisitor



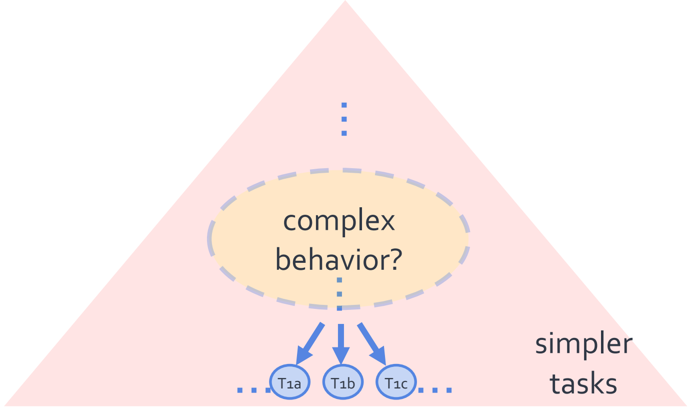
# Complex Problem Solving As Communication



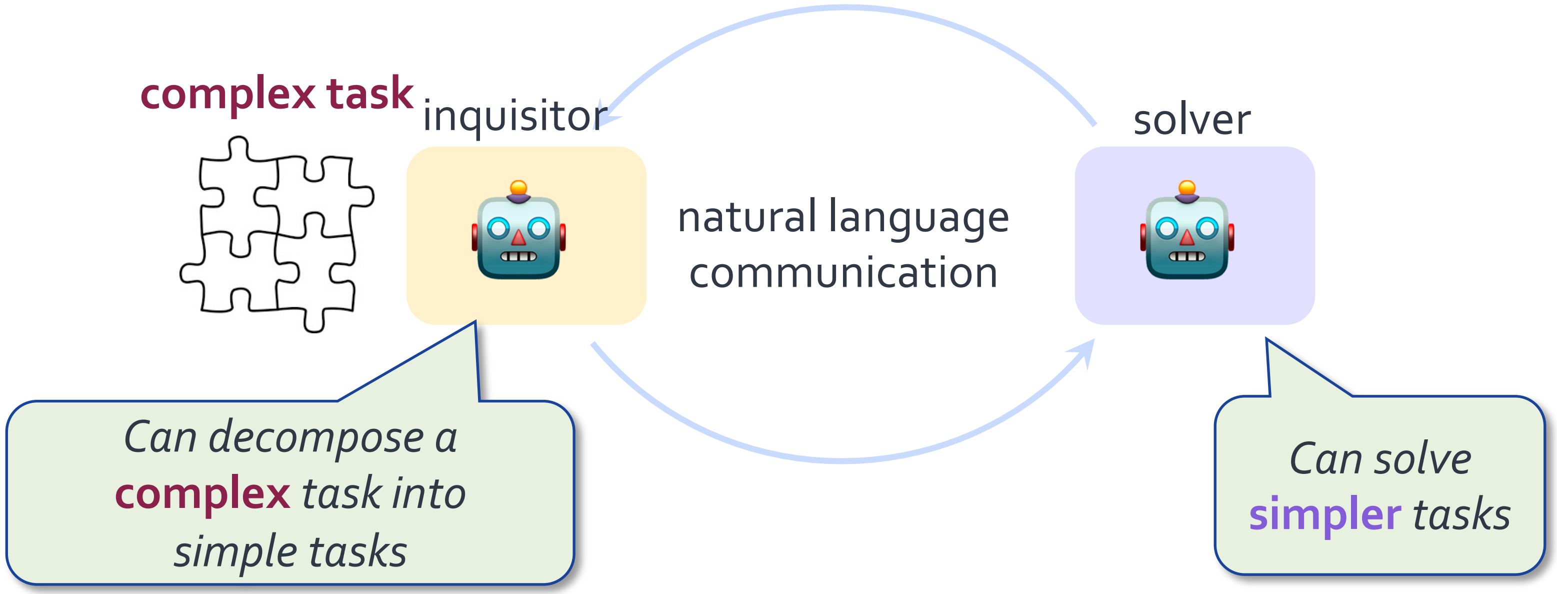

- **Setup:**
  - Communications between models
  - Goal oriented
  - Roles: solver and inquisitor



# Complex Problem Solving As Communication

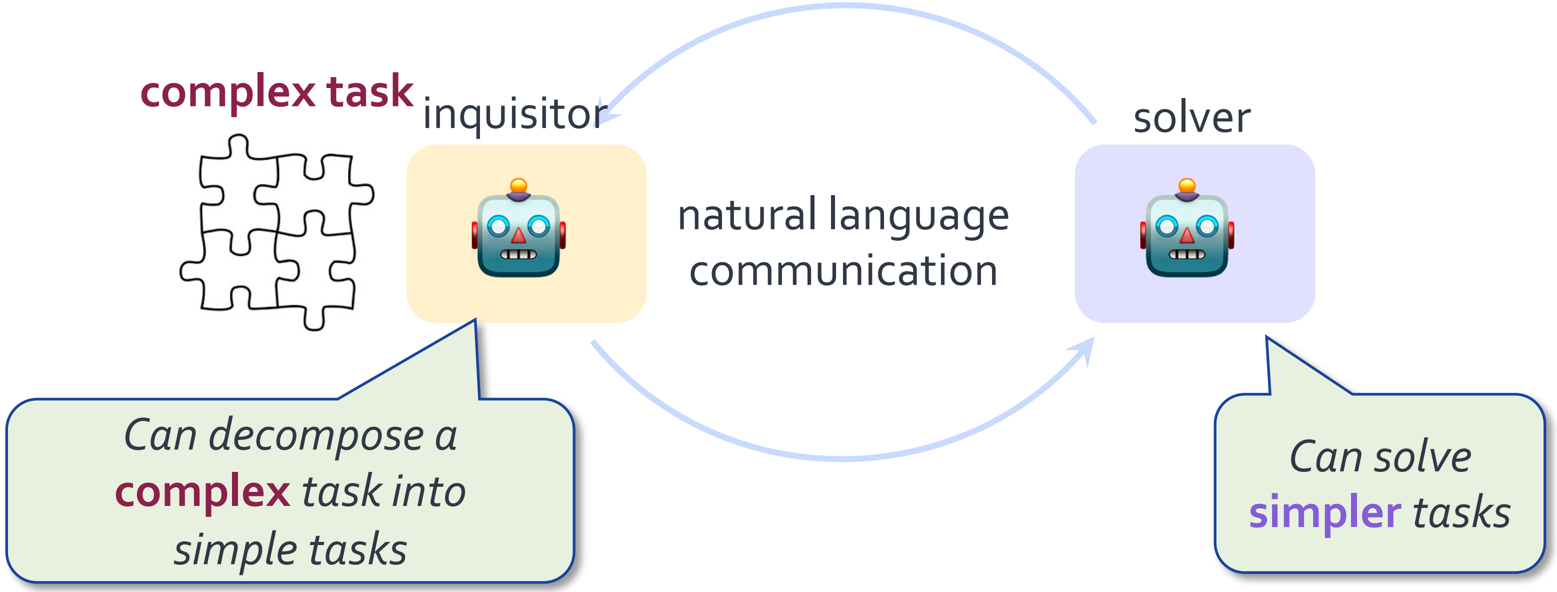



- **Setup:**
  - Communications between models
  - Goal oriented
  - Roles: solver and inquisitor

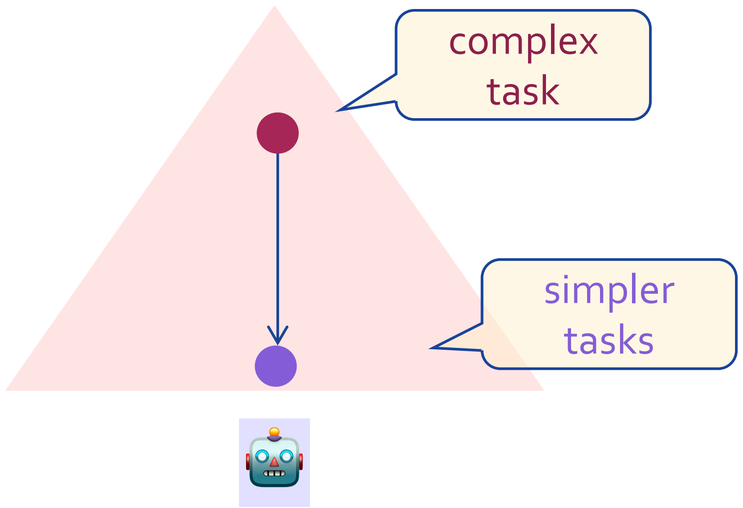


# Complex Problem Solving As Communication

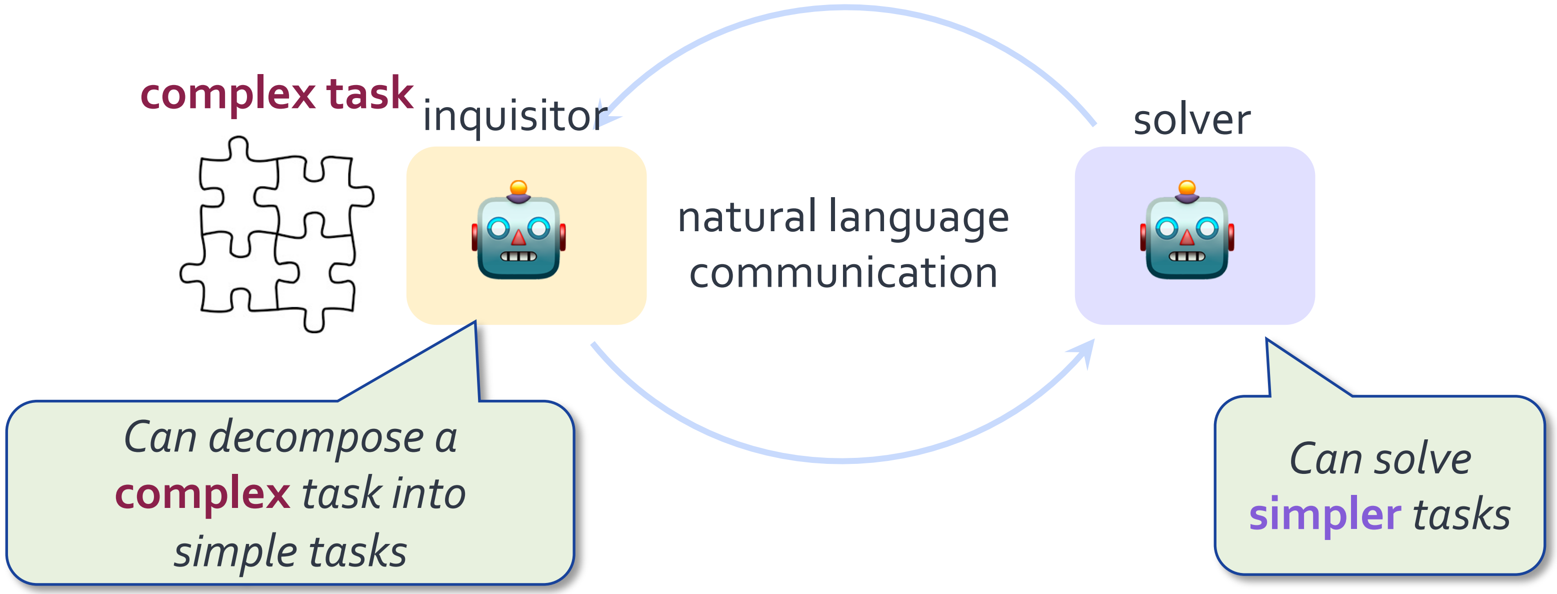
- **Setup:**
  - Communications between models
  - Goal oriented
  - Roles: solver and inquisitor



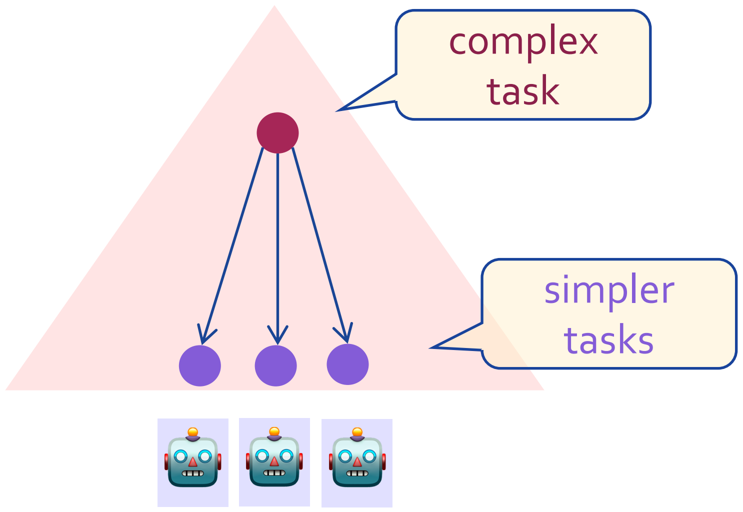
# Complex Problem Solving As Communication



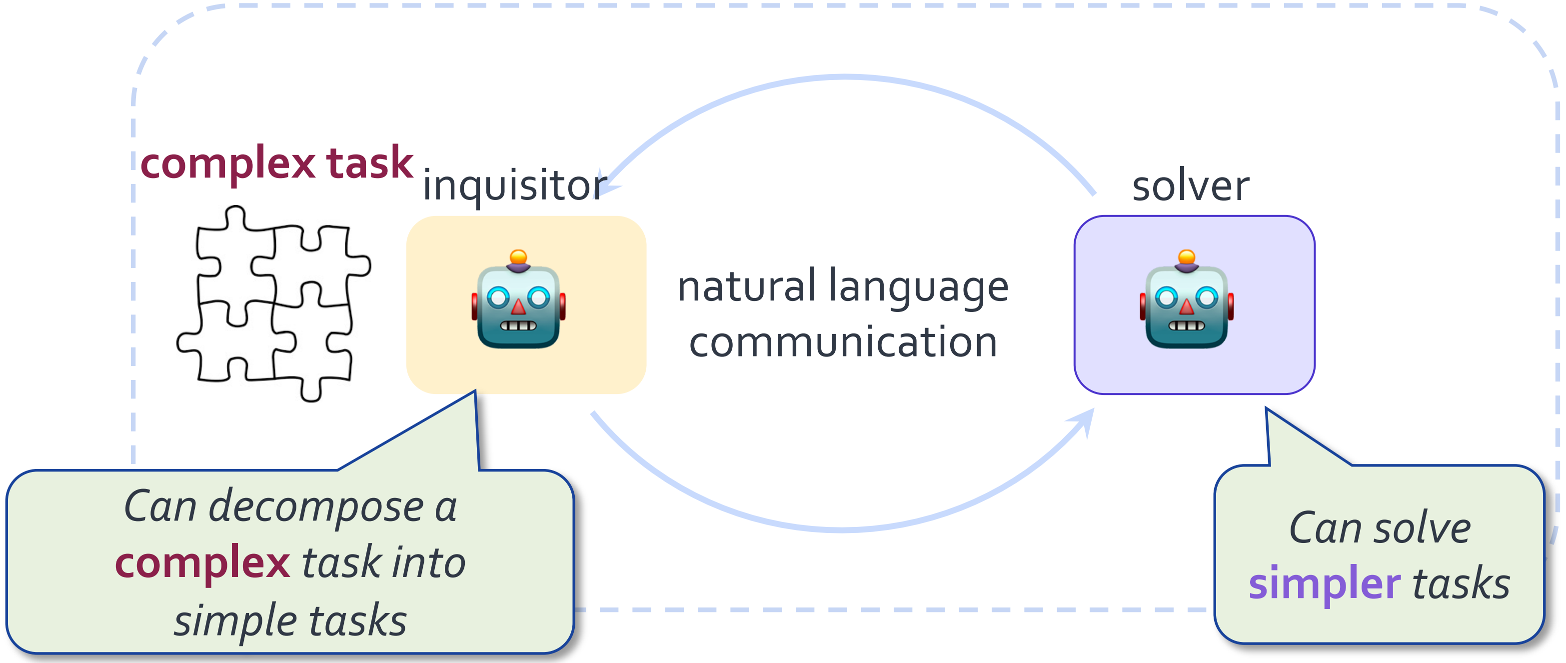

- **Setup:**
  - Communications between models
  - Goal oriented
  - Roles: solver and inquisitor



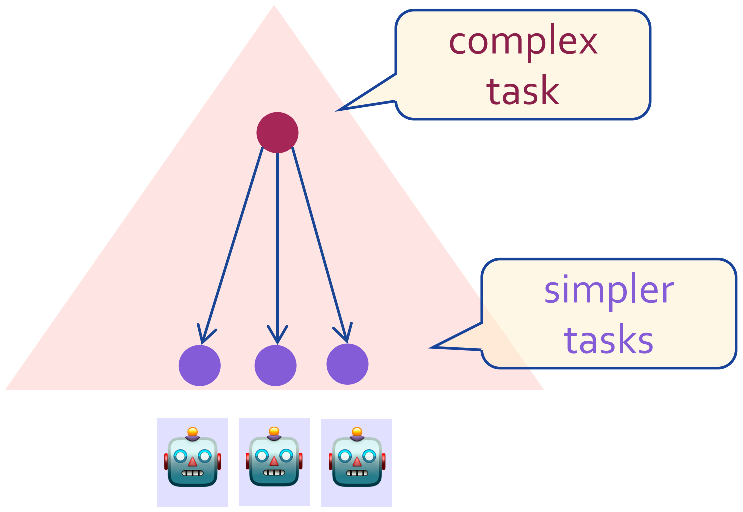
# Text Modular Networks (TMN)



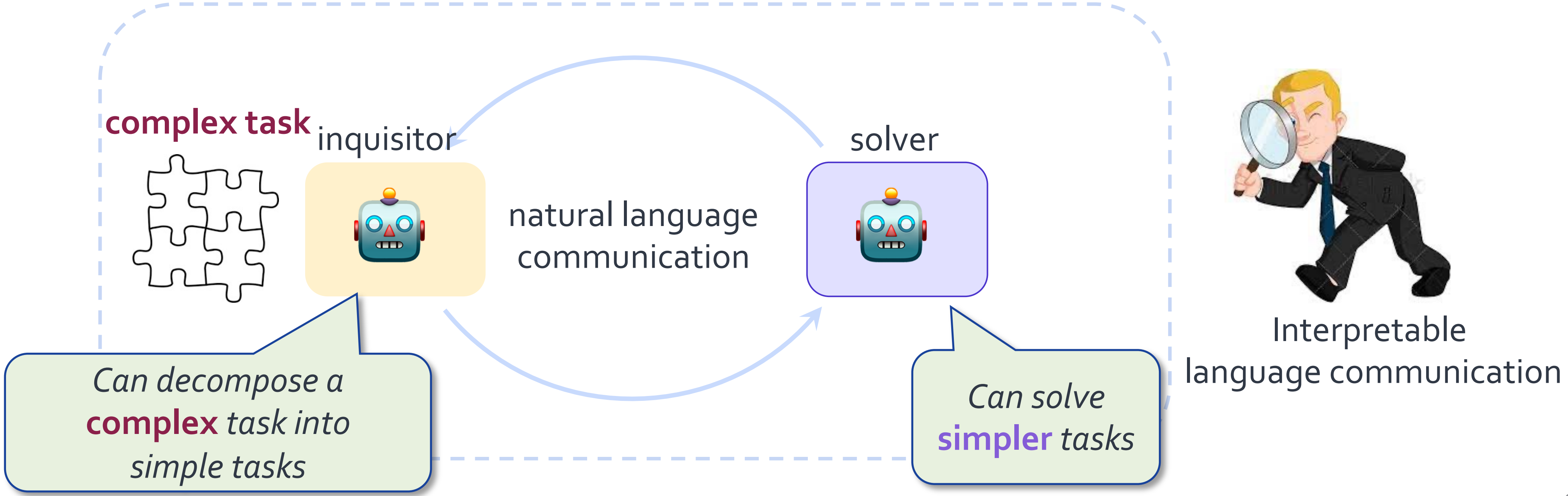

A general framework that leverages existing simpler models –neural or symbolic– through interactive communication.



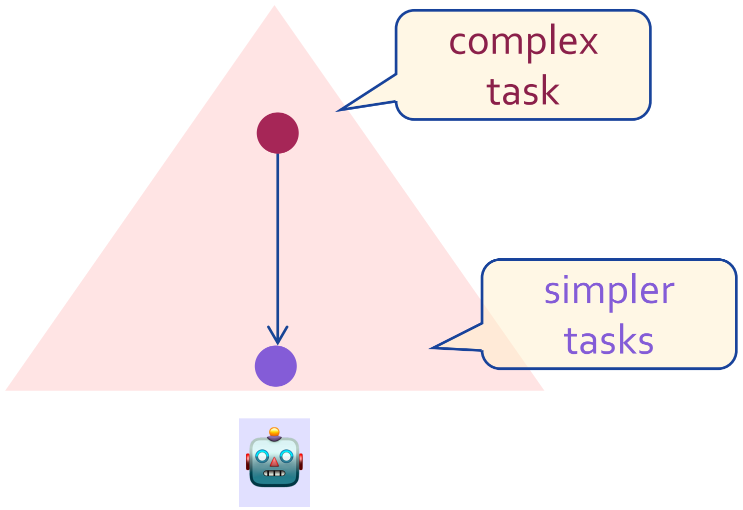
# Text Modular Networks (TMN)



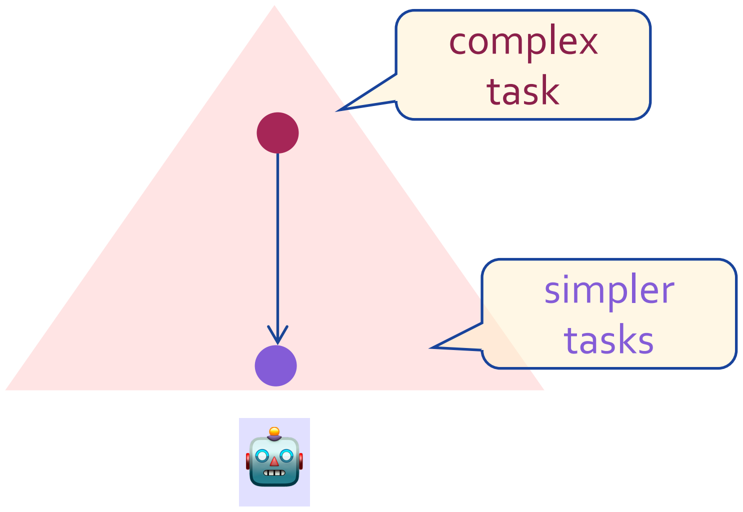

A general framework that leverages existing simpler models –neural or symbolic– through interactive communication.



# Text Modular Networks (TMN)



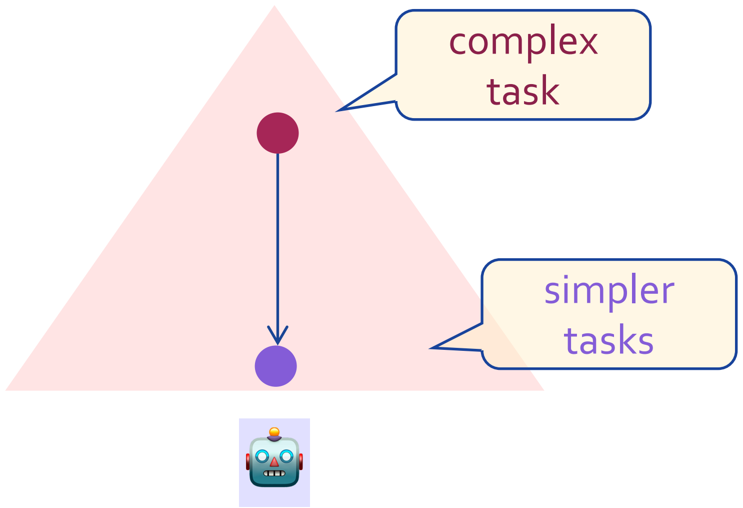
# Text Modular Networks (TMN)




*complex question*

*"What is the nationality of the Simpsons director?"*

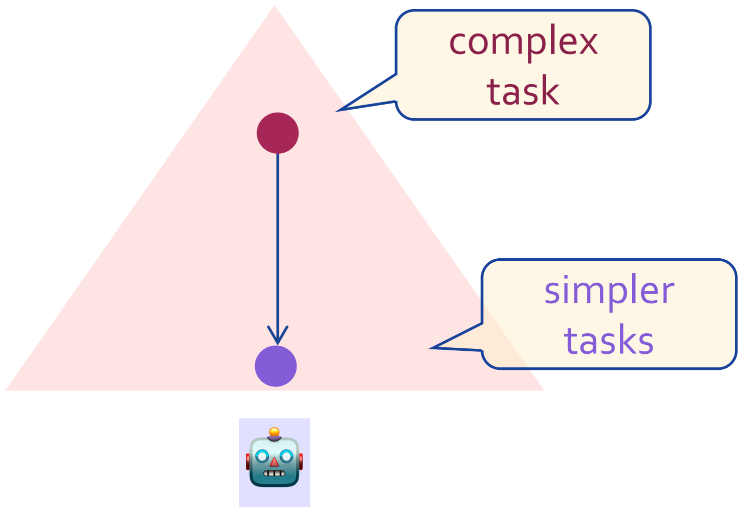
# Text Modular Networks (TMN)




*complex question*

*"What is the nationality of the Simpsons director?"*

# Text Modular Networks (TMN)




*complex question*

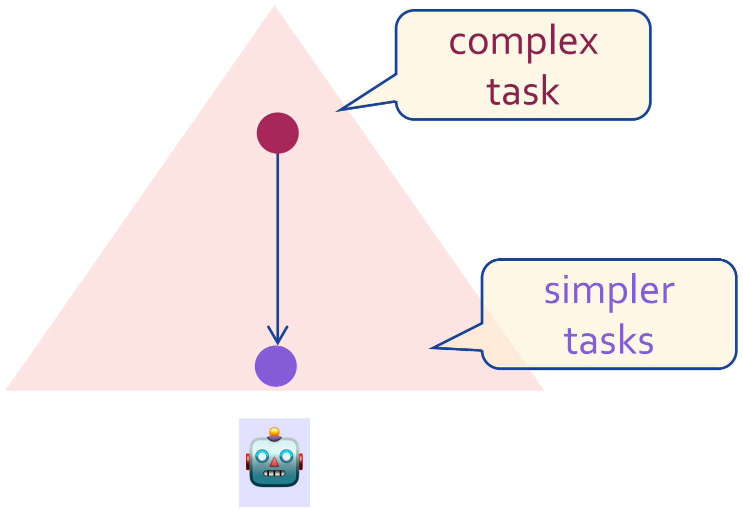
*"What is the nationality of the Simpsons director?"*

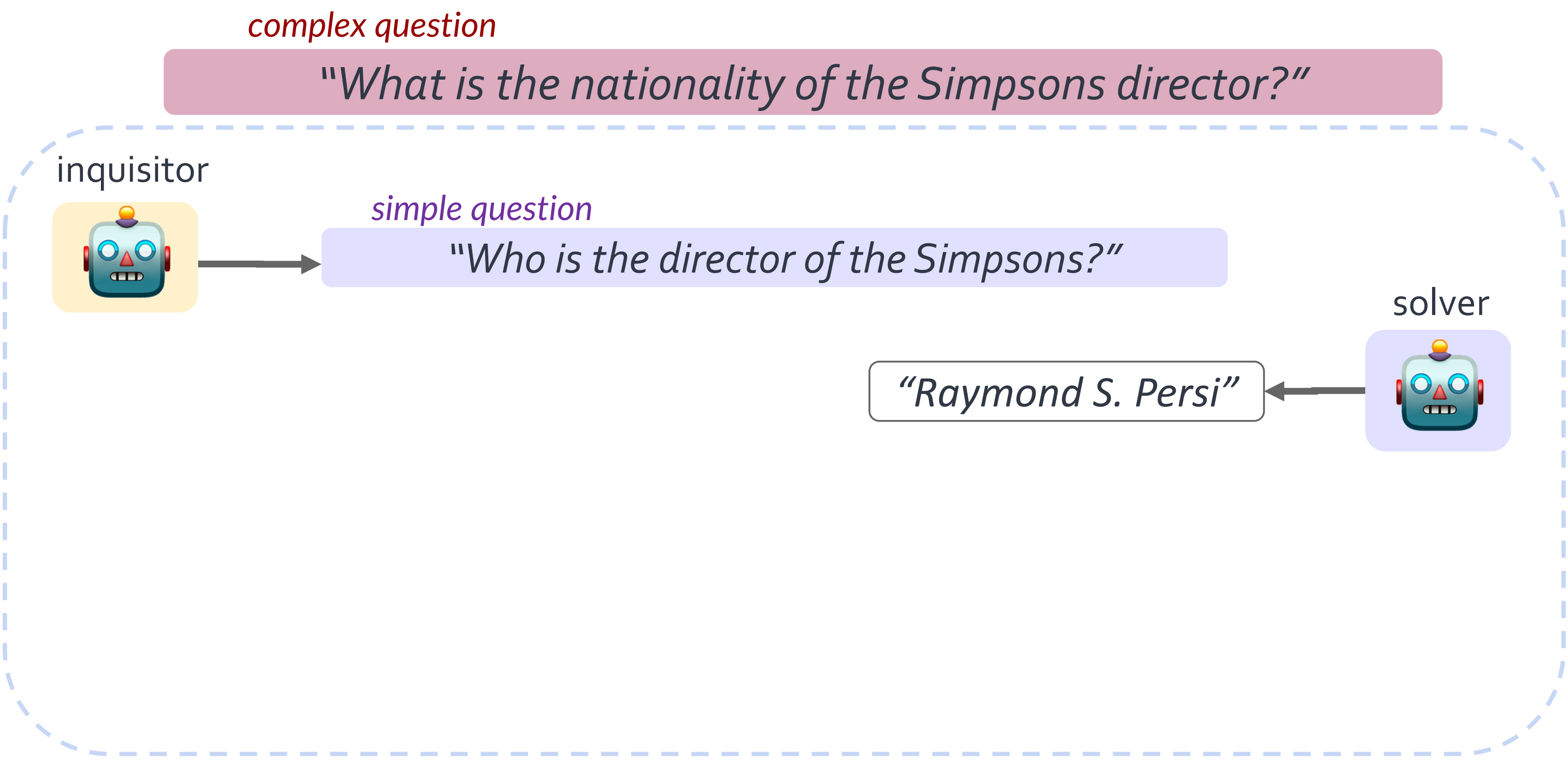
inquisitor

*simple question*

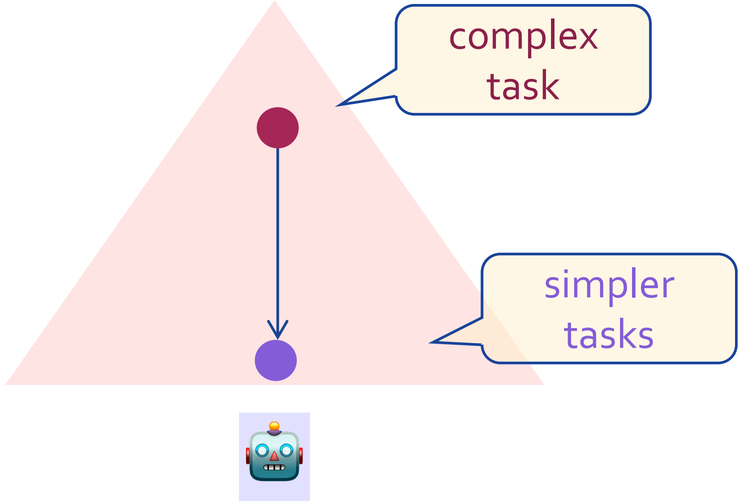
*"Who is the director of the Simpsons?"*

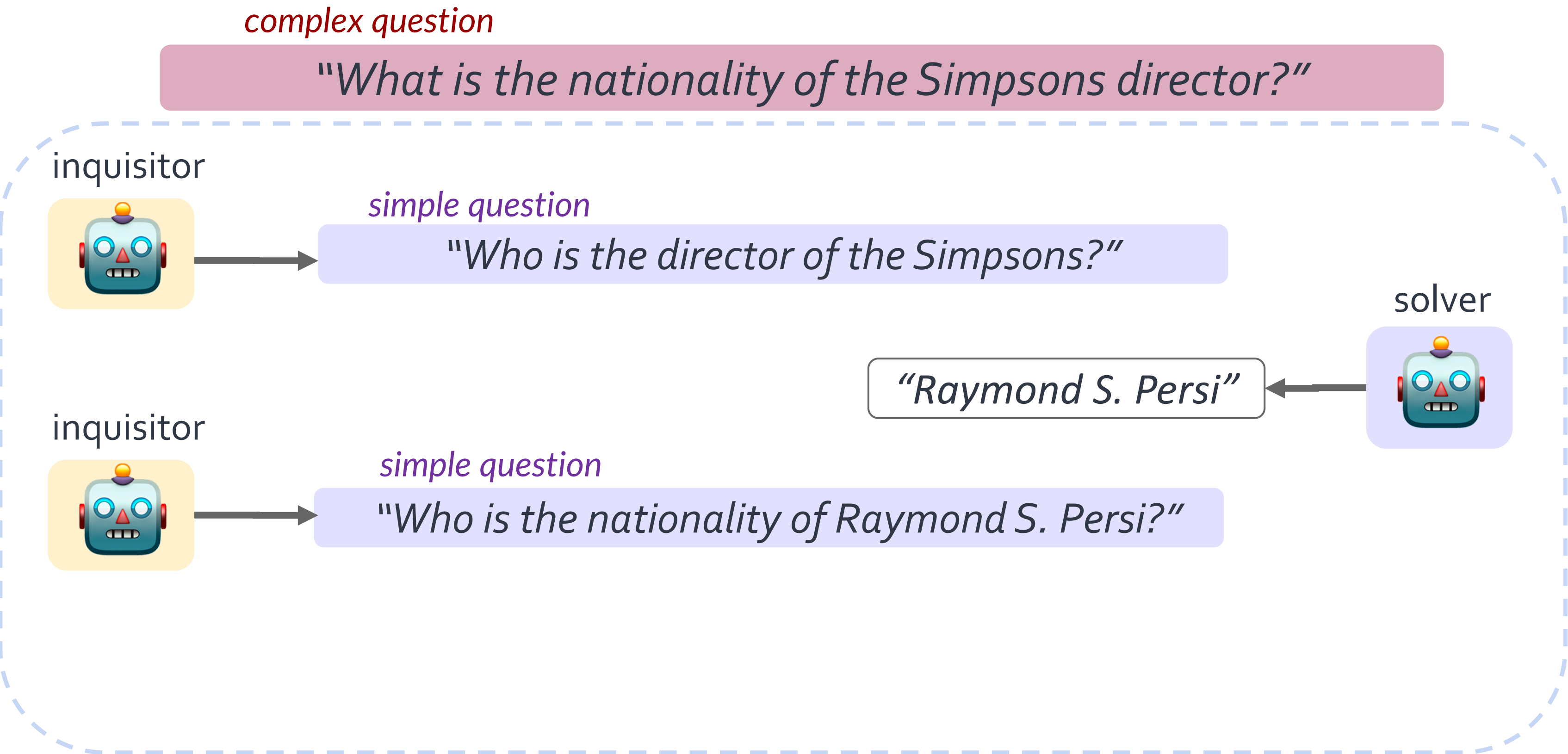
# Text Modular Networks (TMN)



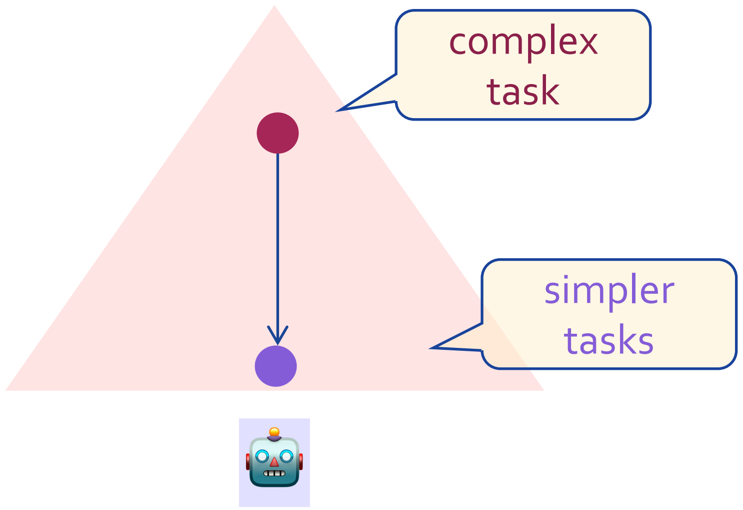



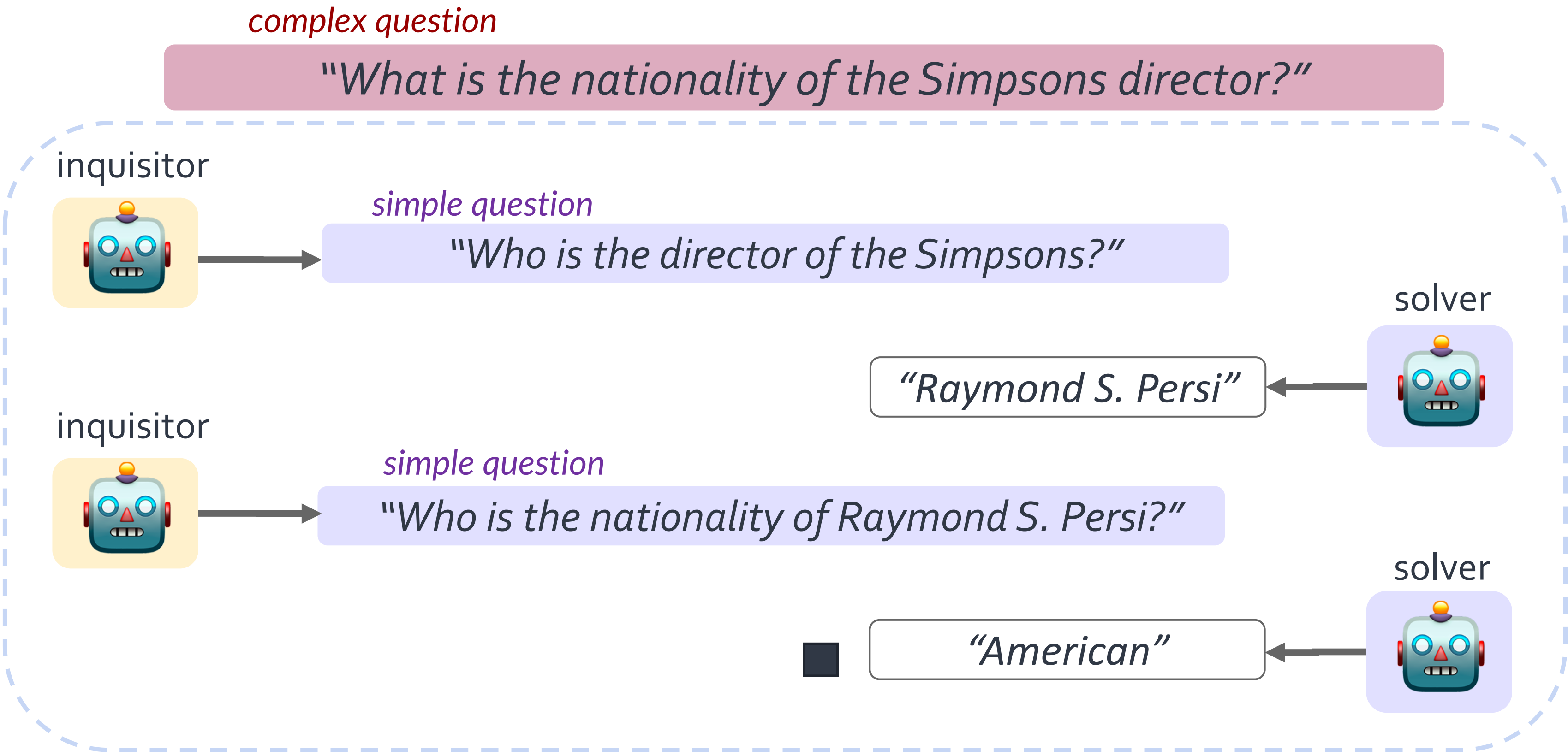
# Text Modular Networks (TMN)



# Text Modular Networks (TMN)



# Demo

https://modularqa-demo.apps.allenai.org/

Selected Reasoning [Ans: **American** ] 0.0044

**Question: What is the nationality of Simpson's "Little Big Girl" director?**

Who was the director of "Little Big Girl"? Curr. Penalty: 0.0000

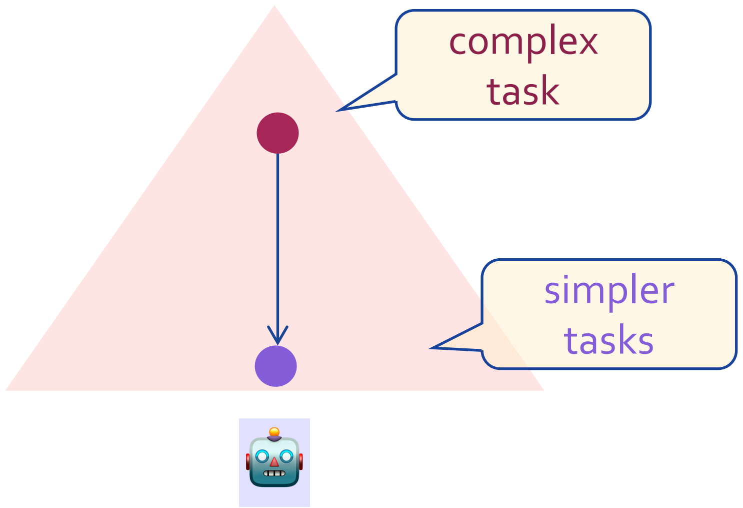
Raymond S. Persi via module: SQUAD QA

What is Raymond S. Persi's nationality? Curr. Penalty: 0.0000

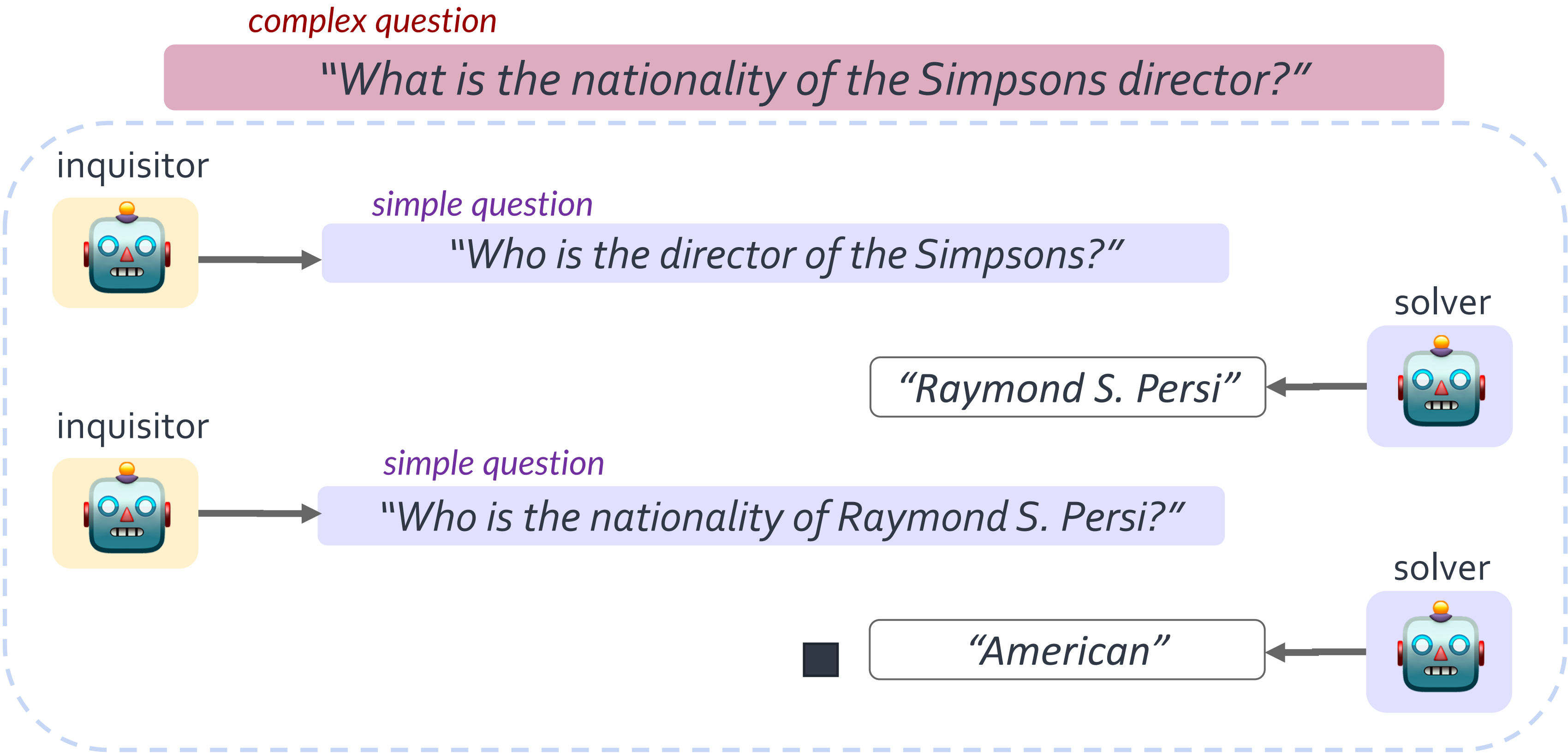
American via module: SQUAD QA

**Answer: American** Final Penalty: 0.0044

# Text Modular Networks (TMN)

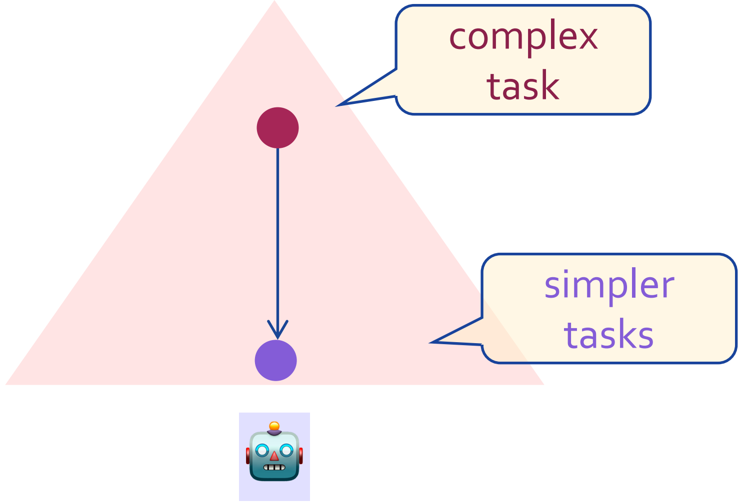


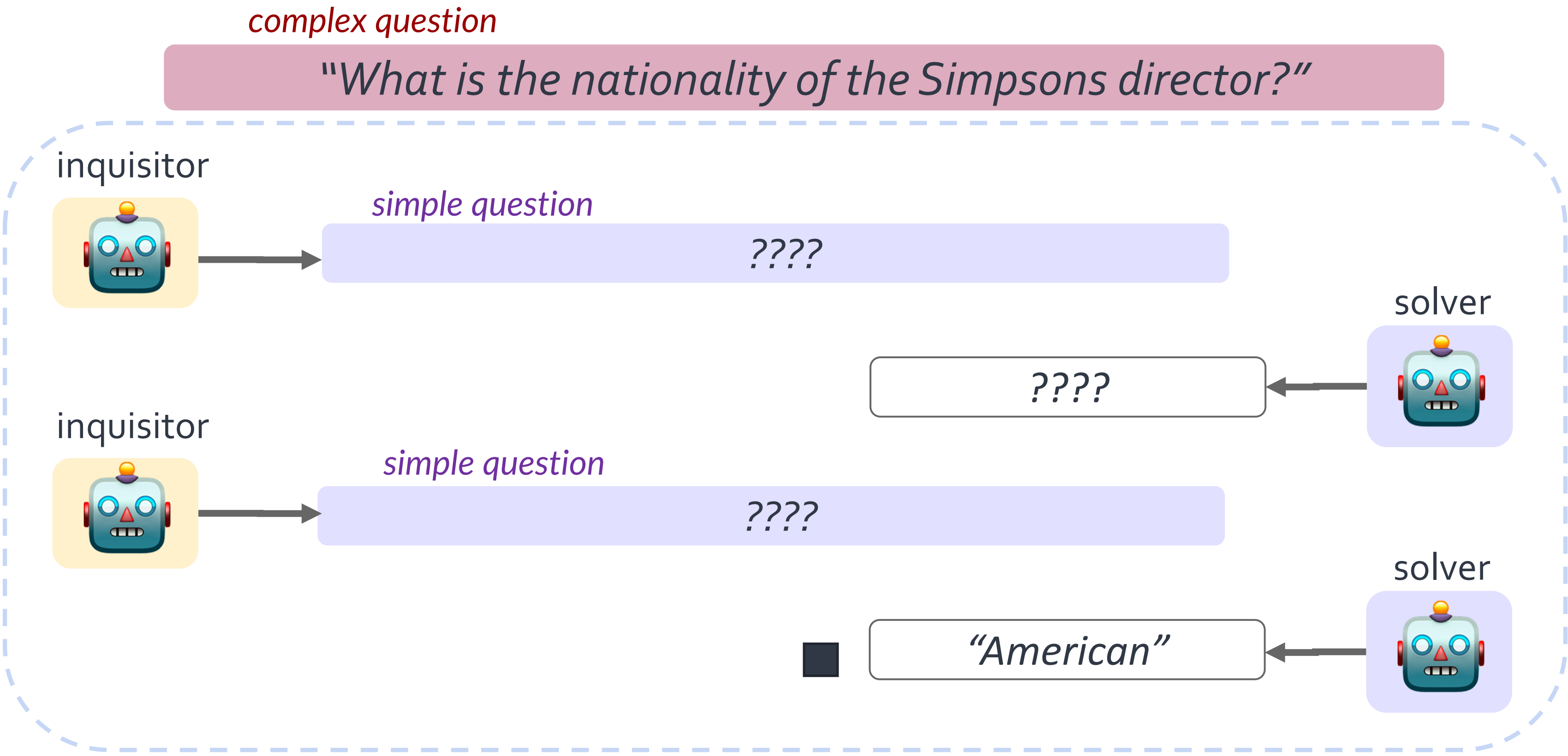

**Research question:** Can we learn to solve complex questions via language interactions with existing, simpler models?



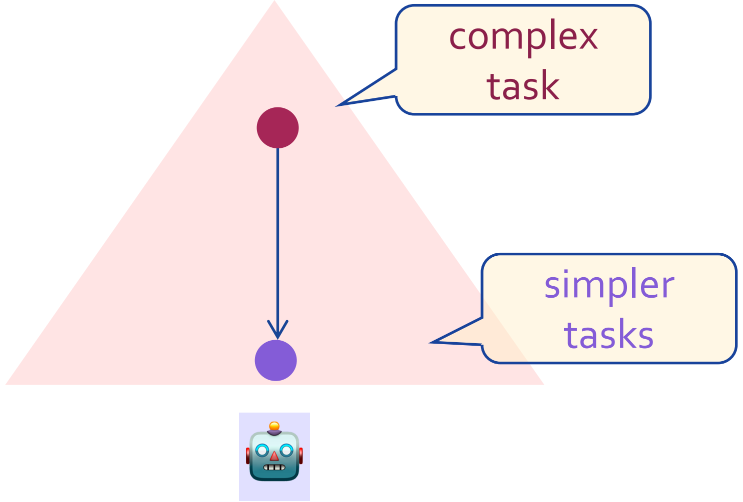
# Text Modular Networks (TMN)

**Research question:** Can we learn to solve complex questions via language interactions with existing, simpler models?

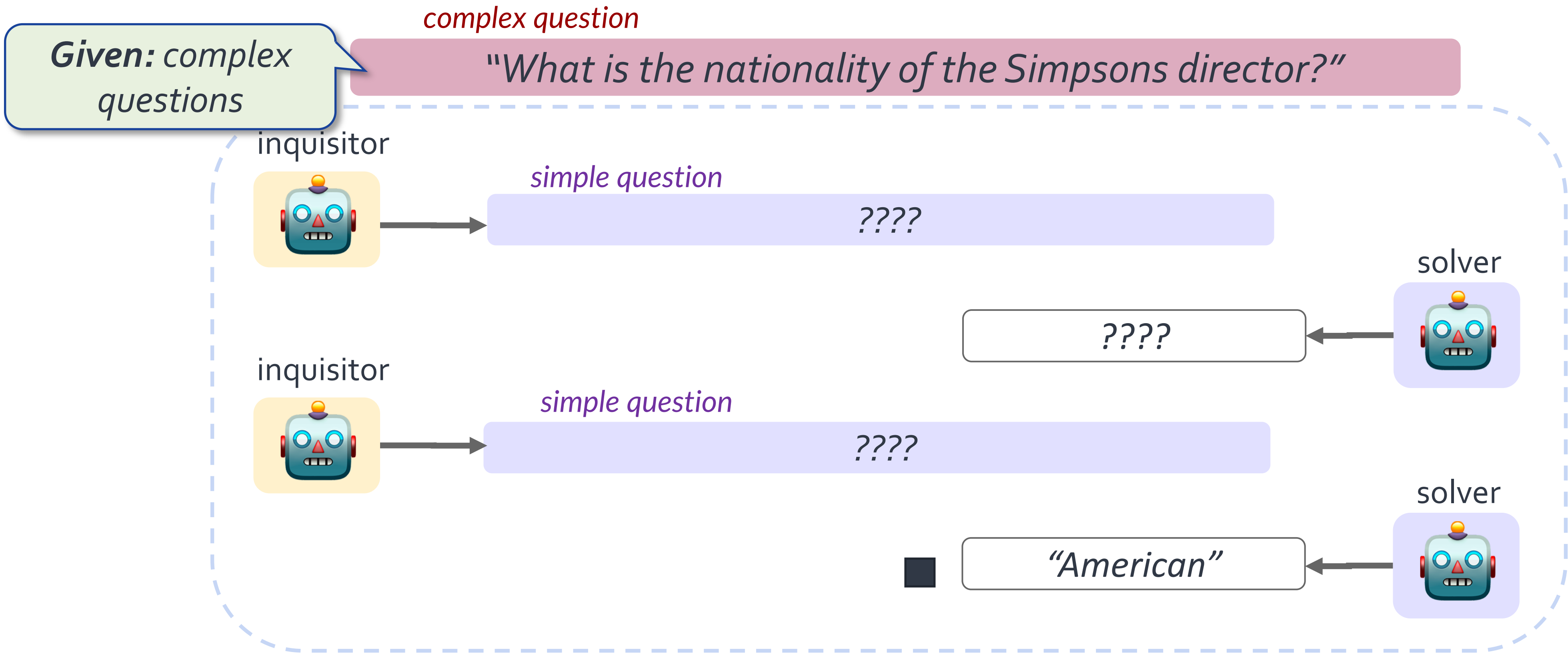


# Text Modular Networks (TMN)

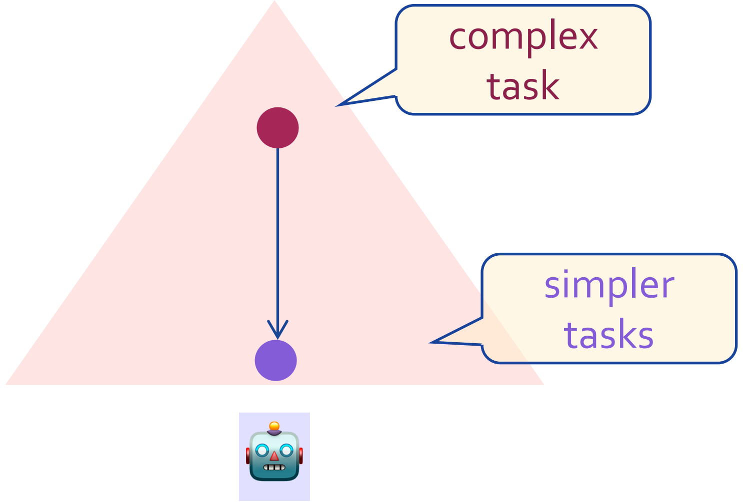



**Research question:** Can we learn to solve complex questions via language interactions with existing, simpler models?

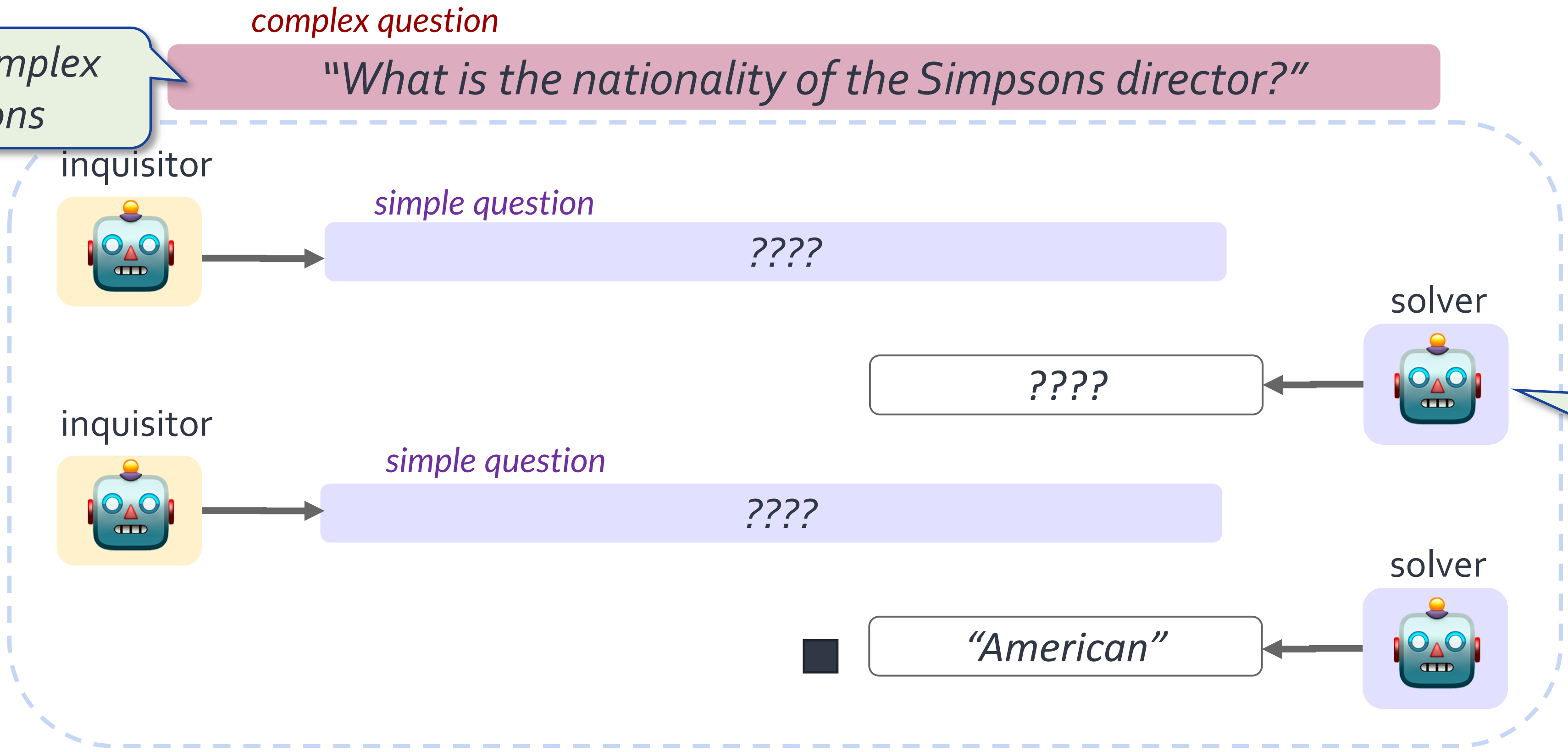


# Text Modular Networks (TMN)

**Research question:** Can we learn to solve complex questions via language interactions with existing, simpler models?

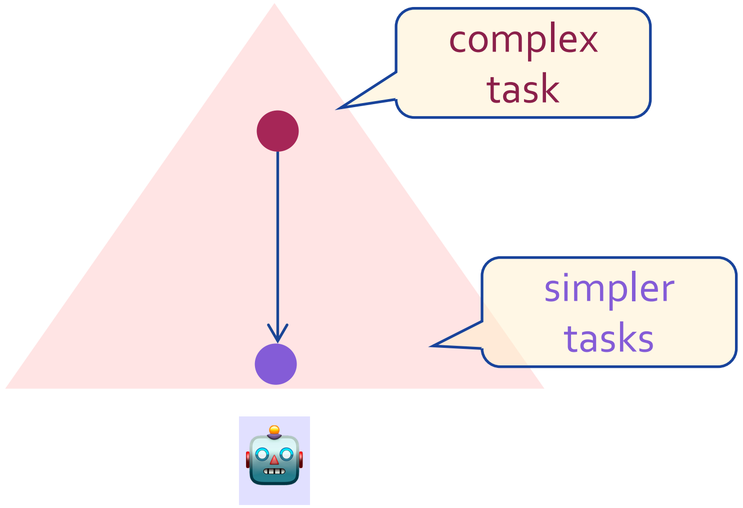



***Given:*** *complex questions*

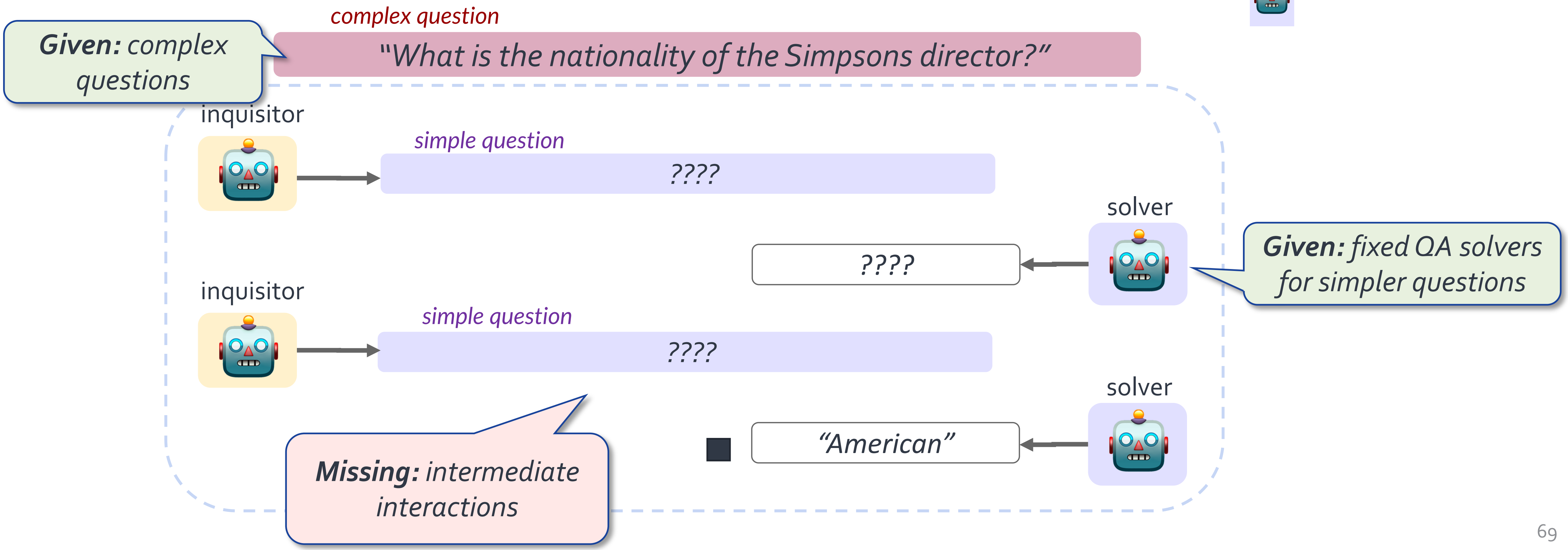



***Given:*** *fixed QA solvers for simpler questions*

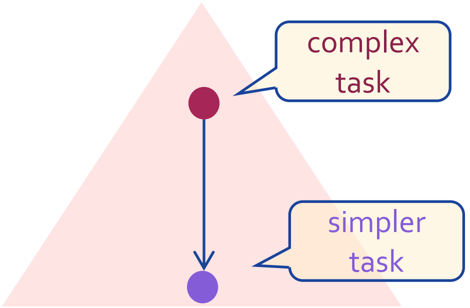
# Text Modular Networks (TMN)



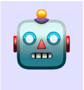

**Research question:** Can we learn to solve complex questions via language interactions with existing, simpler models?



# Approach: Overview

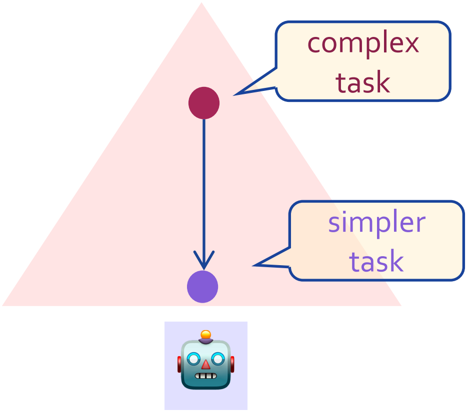



complex question "*What is the nationality of the Simpsons director?*"



answer "*American*"

# Approach: Overview




complex question

*"What is the nationality of the Simpsons director?"*

candidate decompositions
(simple questions)

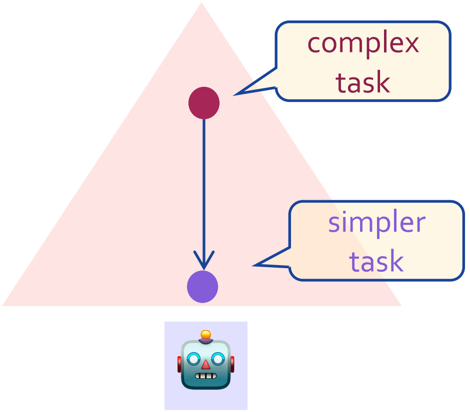
*Who…?* *When…?* *Where…?* *What…?* …

Need to be understandable to the simple models.

answer

*"American"*

# Approach: Overview




complex question

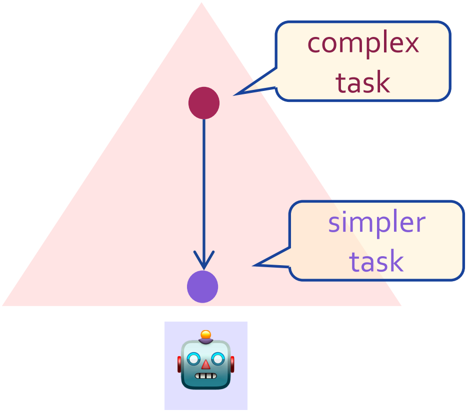
*"What is the nationality of the Simpsons director?"*

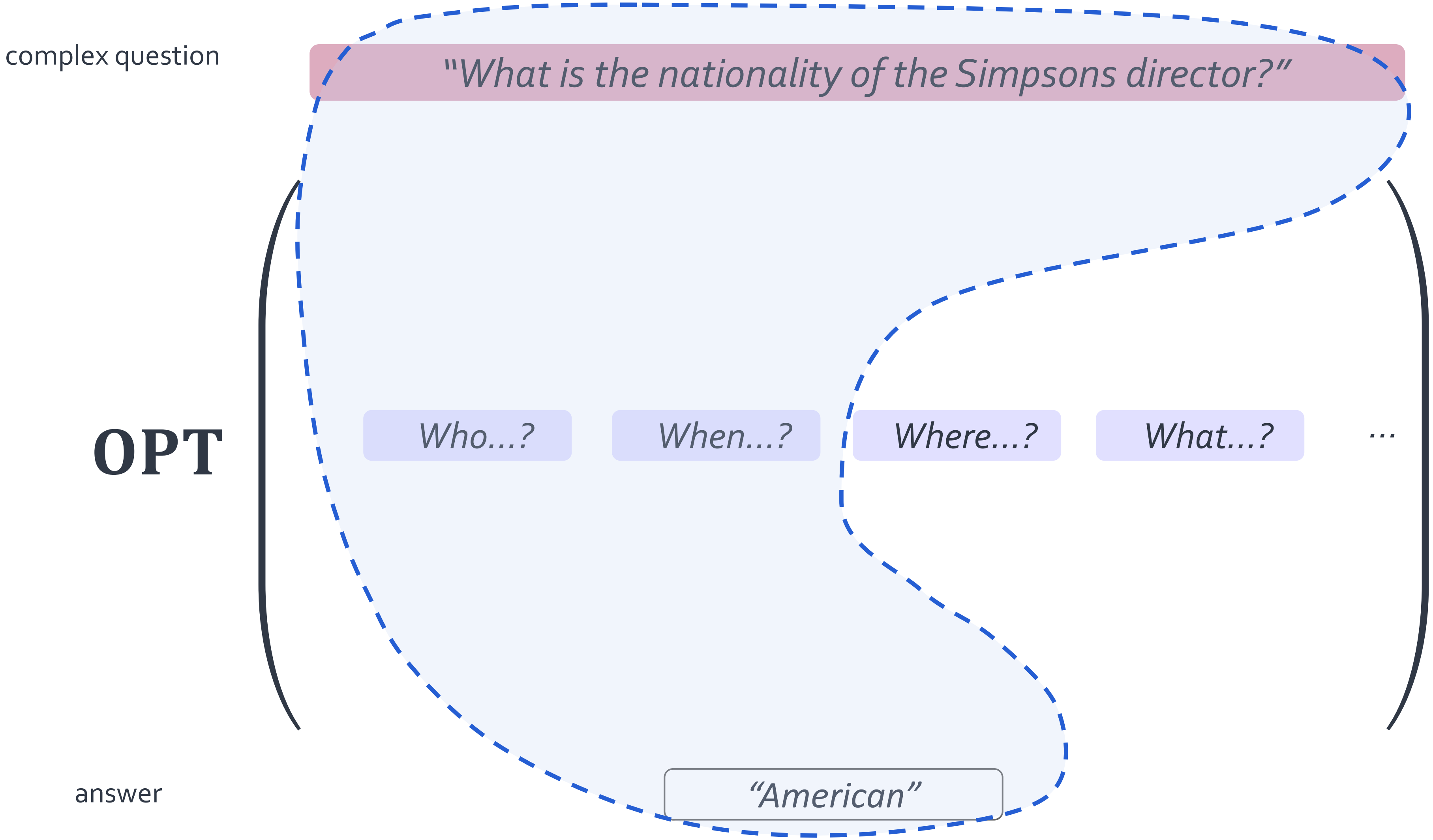
*Who...?* *When...?* *Where...?* *What...?* ...

answer

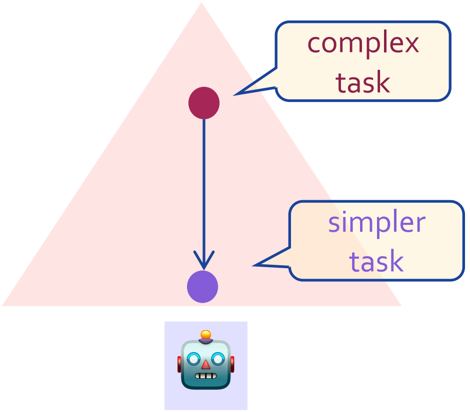
*"American"*

# Approach: Overview



# Step 1: Relevant Documents



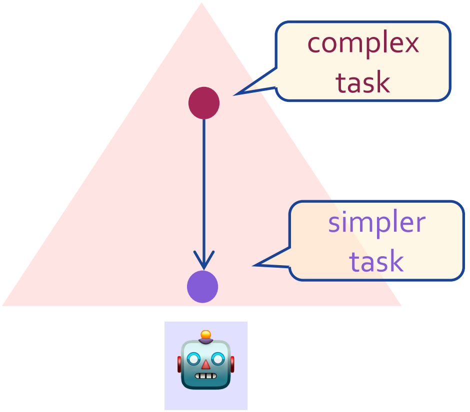

complex question

*"What is the nationality of the Simpsons director?"*

answer

*"American"*

Step 1: # Relevant Documents




complex question

*"What is the nationality of the Simpsons director?"*

*"Little Big Girl" is the twelfth episode of "The Simpsons"'s eighteenth season. It originally aired on the Fox network in the United States on February 11, 2007. It was written by Don Payne, and directed by Raymond S. Persi. Natalie Portman guest starred as a new character, Darcy. The title is a play on the Dustin Hoffman movie "Little Big Man". The last time the title was parodied was in season 11's "Little Big Mom."*

*album with his own rendition of "Iisa Lang Tayo". ....*

*Never-Ending Story". Persi went on to work as a sequence director ...*

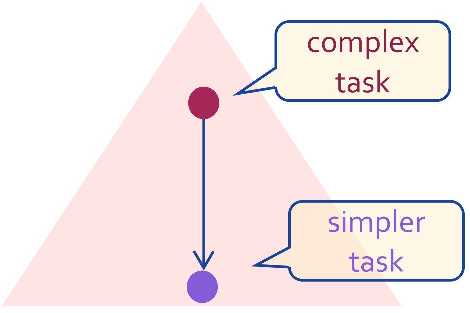
*and the last episode directed by Wes Archer. ...*

*meaning "ineffectual or weak, someone failing to show ....*

answer

*"American"*

# Step 2: Language of Simple QA Models




complex question

*"What is the nationality of the Simpsons director?"*

*"Little Big Girl" is the twelfth episode of "The Simpsons"'s eighteenth season. It originally aired on the Fox network in the United States on February 11, 2007. It was written by Don Payne, and directed by Raymond S. Persi. Natalie Portman guest starred as a new character, Darcy. The title is a play on the Dustin Hoffman movie "Little Big Man". The last time the title was parodied was in season 11's "Little Big Mom."*

*album with his own rendition of "Iisa Lang Tayo". ....*

*on to work as a sequence director ...*

*Archer. ...*

*meaning "ineffectual or weak, so*

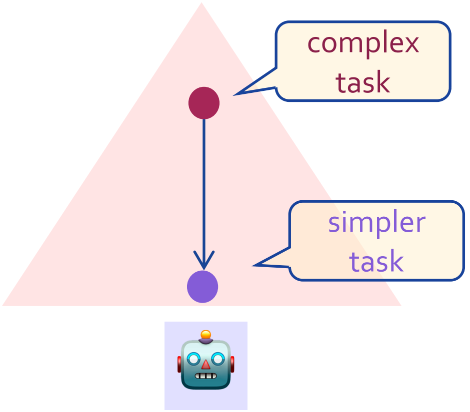
*"Who is the director of Simpson's 'Little Big Girl'?"*

Understandable to the simple models.

answer

*"American"*

# Step 2: Language of Simple QA Models




complex question

*"What is the nationality of the Simpsons director?"*

| | |
|---|---|
| *"Little Big Girl" is in which season of "the Simpsons"'s?* → | *eighteenth* |
| *"Who is the director of Simpson's 'Little Big Girl'?"* → | *Raymond Persi* |
| *"Little Big Girl" is which episode of "the Simpsons"'s?* → | *twelfth* |
| *When was 'Little Big Girl' aired in USA?* → | *February 11, 2007* |
| *Who is the writer of 'Little Big Girl' episode?* → | *Don Payne* |
| ⋮ | ⋮ |

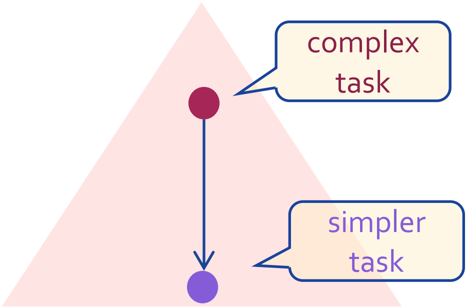
question-answers as an expressive knowledge representation medium.

[He et al., '15, FitzGerald et al. '18]

answer

*"American"*

# Step 3: Subset Selection via Optimization




complex question

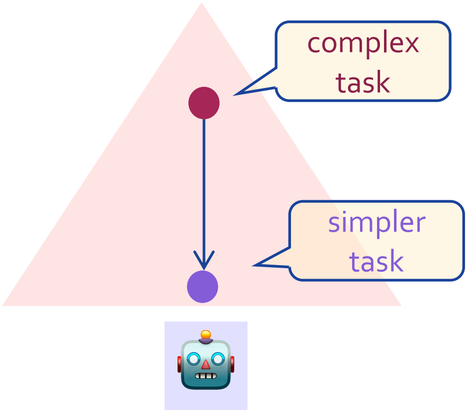
*"What is the nationality of the Simpsons director?"*

| | |
|---|---|
| *"Little Big Girl" is in which season of "the Simpsons"'s?* → | *eighteenth* |
| *"Who is the director of Simpson's 'Little Big Girl'?"* → | *Raymond Persi* |
| *"Little Big Girl" is which episode of "the Simpsons"'s?* → | *twelfth* |
| *When was 'Little Big Girl' aired in USA?* → | *February 11, 2007* |
| *Who is the writer of 'Little Big Girl' episode?* → | *Don Payne* |
| ⋮ | ⋮ |

answer

*"American"*

# Step 3: Subset Selection via Optimization




complex question

*"What is the nationality of the Simpsons director?"*

| | |
|---|---|
| *"Little Big Girl" is in which season of "the Simpsons"'s?* → | *eighteenth* |
| *"Who is the director of Simpson's 'Little Big Girl'?"* → | *Raymond Persi* |
| *"Little Big Girl" is which episode of "the Simpsons"'s?* → | *twelfth* |
| *When was 'Little Big Girl' aired in USA?* → | *February 11, 2007* |
| *Who is the writer of 'Little Big Girl' episode?* → | *Don Payne* |
| ⋮ | ⋮ |

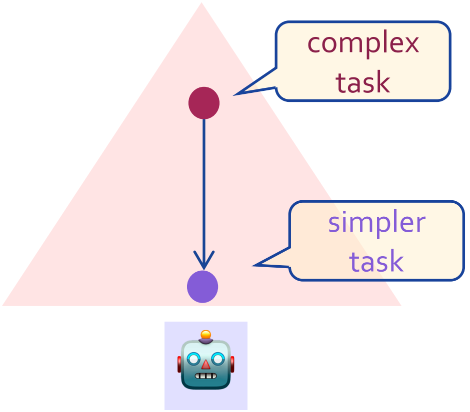
$$\begin{cases} \text{maximize} & \boldsymbol{c}^T\boldsymbol{x} \\ \text{subject to} & \boldsymbol{A}\boldsymbol{x} \le \boldsymbol{b} \\ & \boldsymbol{x} \ge \boldsymbol{0} \\ & \boldsymbol{x} \in \mathbb{Z}^n \end{cases}$$

discrete constrained search

answer

*"American"*

# Step 3: Subset Selection via Optimization




complex question

*"What is the nationality of the Simpsons director?"*

| | |
|---|---|
| *"Little Big Girl" is in which season of "the Simpsons"'s?* | *eighteenth* |
| *"Who is the director of Simpson's 'Little Big Girl'?"* | *Raymond Persi* |
| *"Little Big Girl" is which episode of "the Simpsons"'s?* | *twelfth* |
| *When was 'Little Big Girl' aired in USA?* | *February 11, 2007* |
| *Who is the writer of 'Little Big Girl' episode?* | *Don Payne* |

⋮

$$\begin{cases} \text{maximize} & \boldsymbol{c}^T\boldsymbol{x} \\ \text{subject to} & \boldsymbol{A}\boldsymbol{x} \le \boldsymbol{b} \\ & \boldsymbol{x} \ge \boldsymbol{0} \\ & \boldsymbol{x} \in \mathbb{Z}^n \end{cases}$$

discrete constrained search

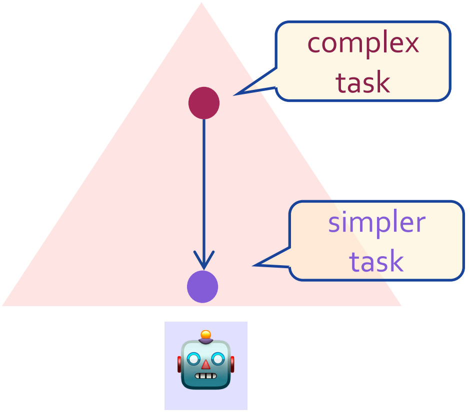
Find a subset of the questions, such that:

1. form a "desirable reasoning structure".

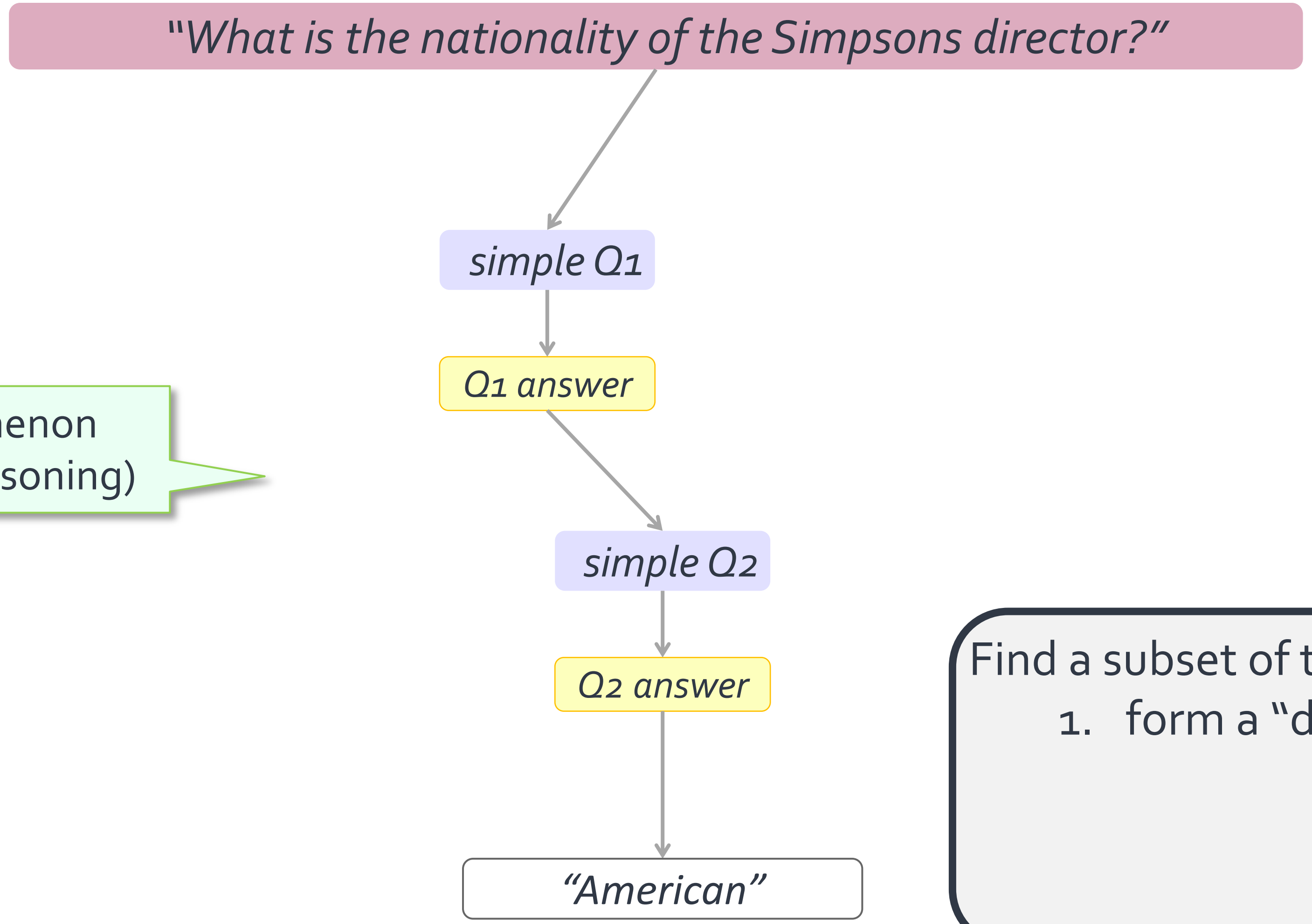
answer

*"American"*

# Step 3: Decomposition via Optimization




complex question




answer

Bridging phenomenon (e.g., deductive reasoning)

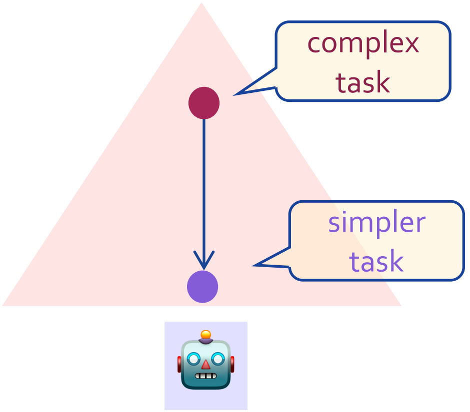
$$
\begin{cases}
\text{maximize} & \boldsymbol{c}^T\boldsymbol{x} \\
\text{subject to} & \boldsymbol{A}\boldsymbol{x} \le \boldsymbol{b} \\
 & \boldsymbol{x} \ge \boldsymbol{0} \\
 & \boldsymbol{x} \in \mathbb{Z}^n
\end{cases}
$$

discrete constrained search

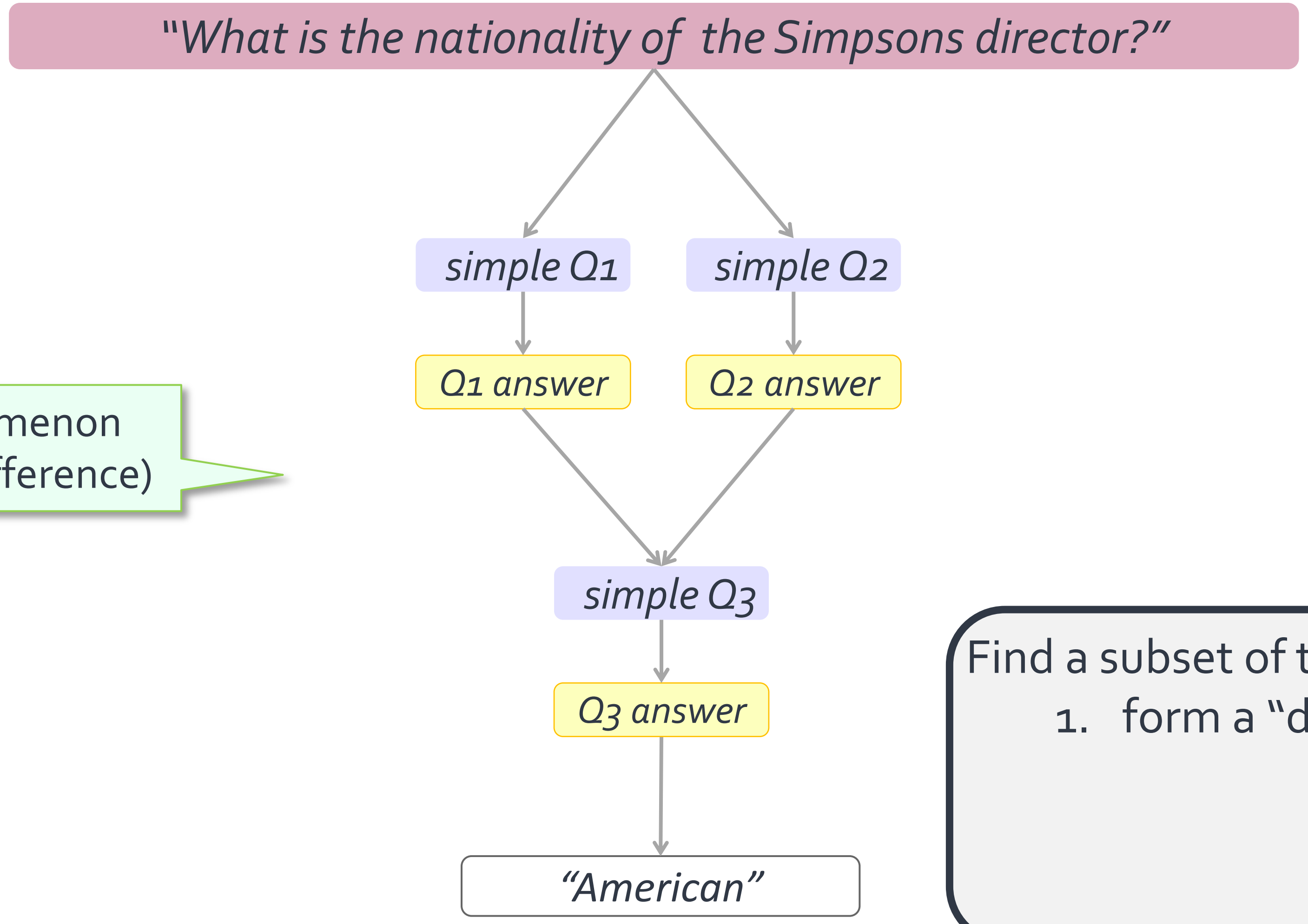
Find a subset of the questions, such that:

1. form a "desirable reasoning structure".

# Step 3: Decomposition via Optimization




complex question




answer

Comparison phenomenon
(e.g., conjunction, difference)

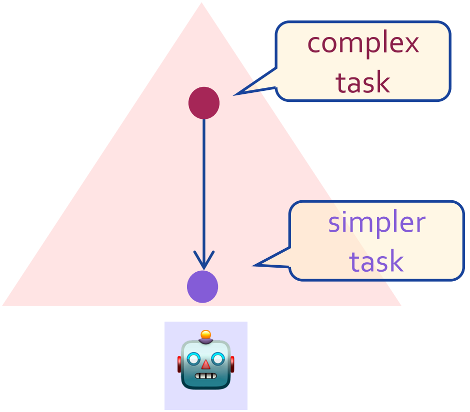
$$\begin{cases} \text{maximize} & \boldsymbol{c}^T \boldsymbol{x} \\ \text{subject to} & A\boldsymbol{x} \leq \boldsymbol{b} \\ & \boldsymbol{x} \geq \boldsymbol{0} \\ & \boldsymbol{x} \in \mathbb{Z}^n \end{cases}$$

discrete constrained search

Find a subset of the questions, such that:

1. form a "desirable reasoning structure".

# Step 3: Decomposition via Optimization




complex question

*"What is the nationality of the Simpsons director?"*

*"Who is the director of the Simpsons?"*

*"Raymond Persi"*

*"What is the nationality of Raymond S. Persi?"*

*"American"*

answer

*"American"*

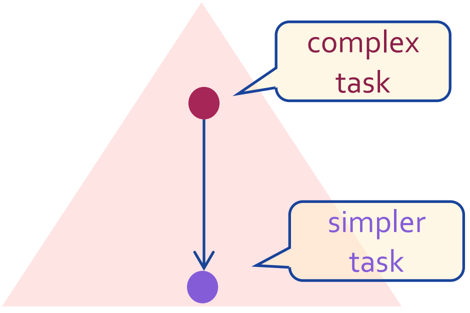
$$\begin{cases} \text{maximize} & \boldsymbol{c}^T\boldsymbol{x} \\ \text{subject to} & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \\ & \boldsymbol{x} \geq \boldsymbol{0} \\ & \boldsymbol{x} \in \mathbb{Z}^{\boldsymbol{n}} \end{cases}$$

discrete constrained search

Find a subset of the questions, such that:

1. form a "desirable reasoning structure".

# Step 3: Decomposition via Optimization




complex question

*"What is the nationality of the Simpsons director?"*

*"Who is the director of the Simpsons?"*

*"Raymond Persi"*

*"What is the nationality of Raymond S. Persi?"*

*"American"*

answer

*"American"*

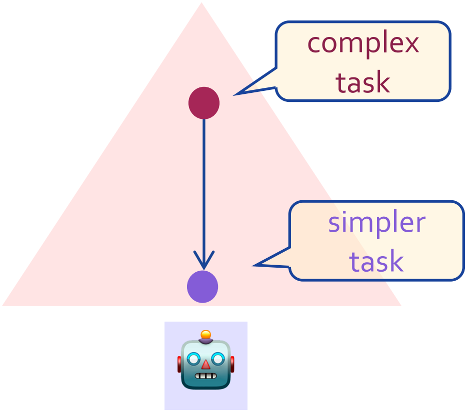
$$\begin{cases} \text{maximize} & \boldsymbol{c}^T\boldsymbol{x} \\ \text{subject to} & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \\ & \boldsymbol{x} \geq \boldsymbol{0} \\ & \boldsymbol{x} \in \mathbb{Z}^n \end{cases}$$

discrete constrained search

Find a subset of the questions, such that:

1. form a "desirable reasoning structure".
2. satisfy sparsity/regularization factors:
   - Small pairwise overlap.
   - Cover the complex question.

Step 4: **Learn to Decompose**




complex question

*"What is the nationality of the Simpsons director?"*

↓

*"Who is the director of the Simpsons?"*

↓

*"Raymond Persi"*

↓

*"What is the nationality of Raymond S. Persi?"*

↓

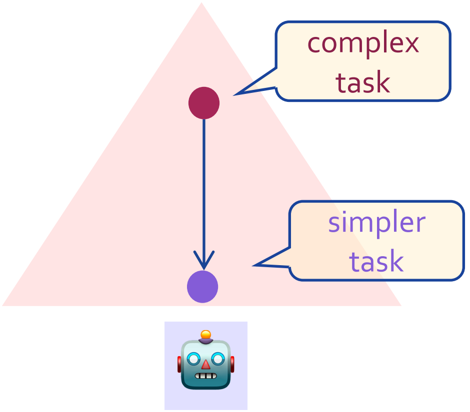
*"American"*

↓

answer

*"American"*

# Step 4: Learn to Decompose




complex question

"What is the nationality of the Simpsons director?"

"Who is the director of the Simpsons?"

"Raymond Persi"

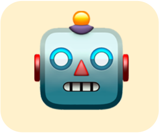
"What is the nationality of Raymond S. Persi?"

"American"

answer

"American"

inquisitor



Trained on [noisy] decompositions

**No decomposition annotation needed!**

# Summary of Empirical Observations

| | DROP (F1) [Ran et al. 19] | HotPotQA (F1) [Ran et al. 19] |
|---|---|---|
| NumNet [Ran et al. 19] | 92 | 🤔 |
| Quark [Groeneveld et al. 20] | 🤔 | 76 |
| TMN [this work] | 88 | 62 |

- **Competitive** with dataset-specific.

“Text Modular Networks: Learning to Decompose Tasks in the Language of Existing Models.” Khot, **K**, Richardson, Sabharwal NAACL‘21

# Summary of Empirical Observations

- **Competitive** with dataset-specific.

| | DROP (F1) [Ran et al. 19] | HotPotQA (F1) [Ran et al. 19] |
|---|---|---|
| **NumNet** [Ran et al. 19] | 92 | 🤔 |
| **Quark** [Groeneveld et al. 20] | 🤔 | 76 |
| **TMN** [this work] | 88 | 62 |

"Text Modular Networks: Learning to Decompose Tasks in the Language of Existing Models." Khot, K, Richardson, Sabharwal NAACL'21

# Summary of Empirical Observations

- **Competitive** with dataset-specific.

| | DROP (F1) [Ran et al. 19] | HotPotQA (F1) [Ran et al. 19] |
|---|---|---|
| **NumNet** [Ran et al. 19] | 92 | 🤔 |
| **Quark** [Groeneveld et al. 20] | 🤔 | 76 |
| **TMN** [this work] | 88 | 62 |

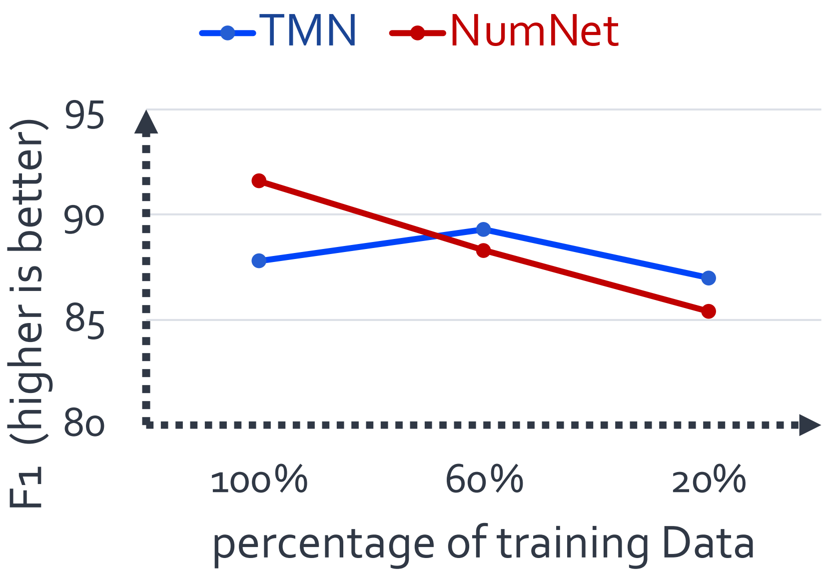
"Text Modular Networks: Learning to Decompose Tasks in the Language of Existing Models." Khot, K, Richardson, Sabharwal NAACL'21

# Summary of Empirical Observations

- **Competitive** with dataset-specific.

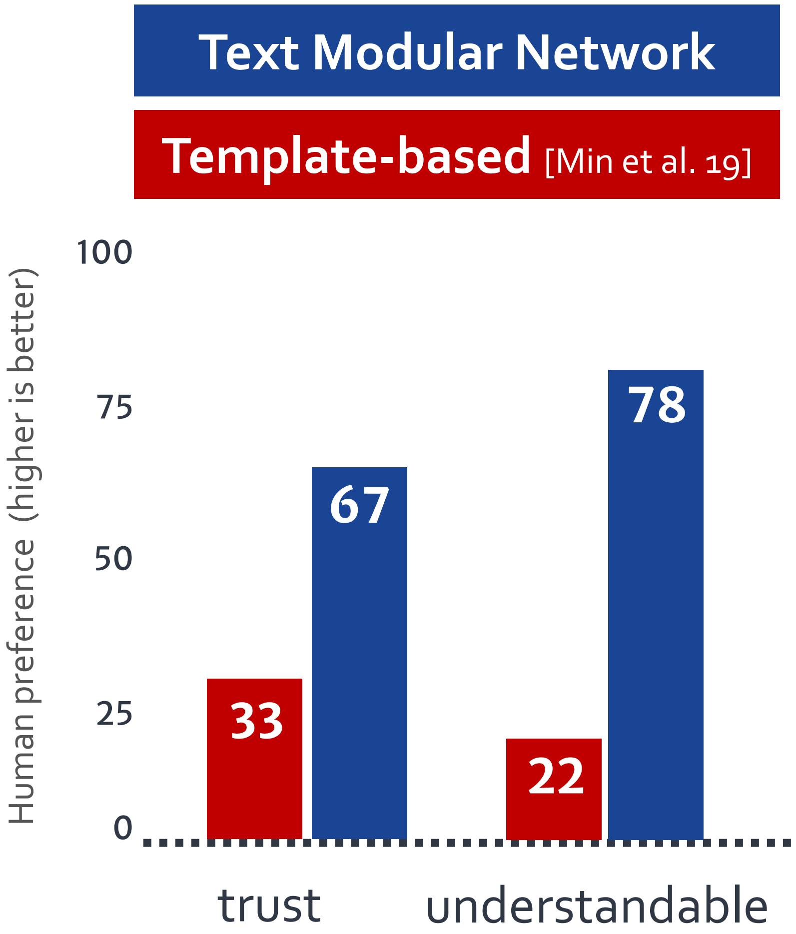
| | DROP (F1) [Ran et al. 19] | HotPotQA (F1) [Ran et al. 19] |
|---|---|---|
| **NumNet** [Ran et al. 19] | 92 | 🤔 |
| **Quark** [Groeneveld et al. 20] | 🤔 | 76 |
| **TMN** [this work] | 88 | 62 |

# Summary of Empirical Observations




- **Competitive** with dataset-specific.
- **Sample efficient** — requires fewer examples to reach a certain accuracy.

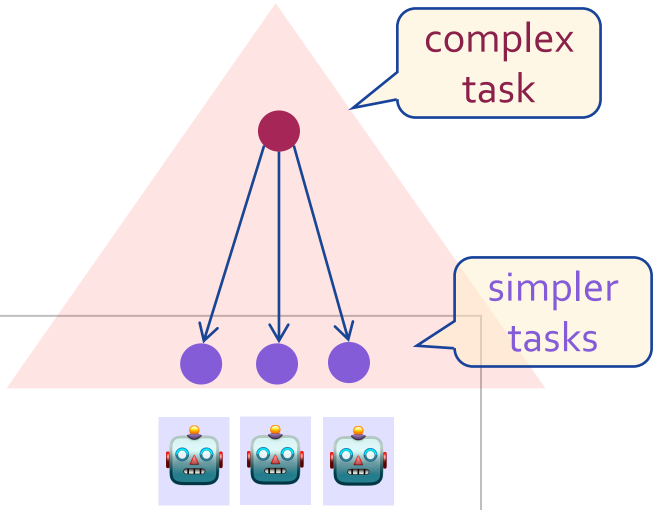
# Summary of Empirical Observations




- **Competitive** with dataset-specific.
- **Sample efficient** — requires fewer examples to reach a certain accuracy.
- **Interpretable** — human judges deemed it more "understandable" and "trustworthy".

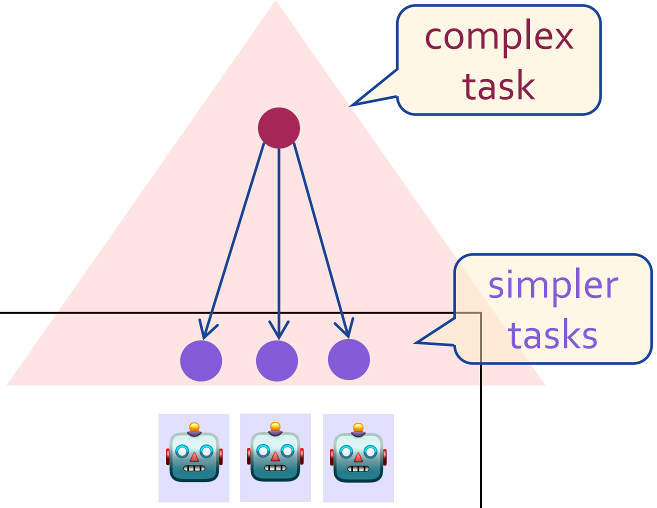
"Text Modular Networks: Learning to Decompose Tasks in the Language of Existing Models." Khot, K, Richardson, Sabharwal NAACL'21

# Summary



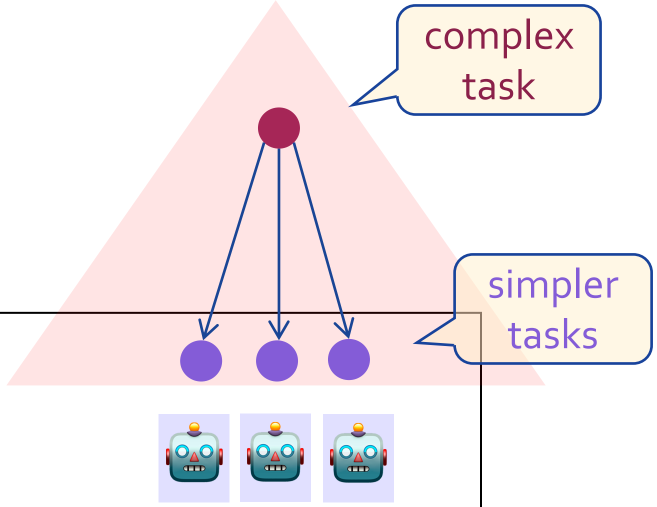

- **Motivating Question:** Can we solve complex tasks as **communication with simpler models**?
- Text Modular Networks, a general-purpose framework for solving complex tasks via textual interaction between existing modules.
- Approach: discrete optimization on existing simple models.
  - Resulting model is more interpretable, competitive yet sample-efficient.
- **Open questions:**
  - How can we make TMNs more extensible?

# Summary



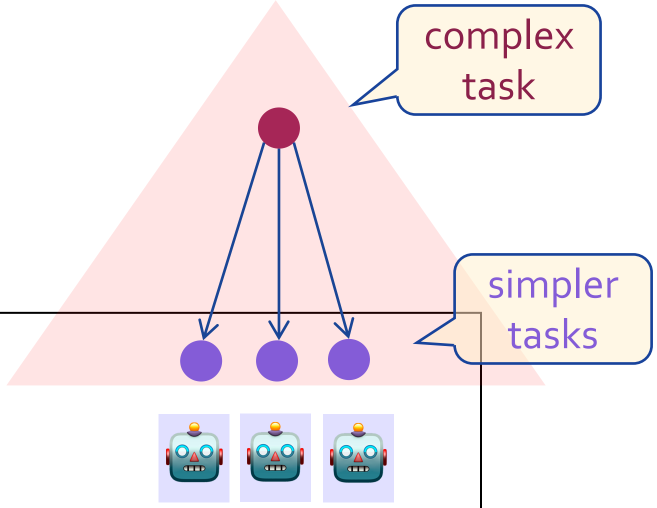

- **Motivating Question:** Can we solve complex tasks as **communication with simpler models**?
- Text Modular Networks, a general-purpose framework for solving complex tasks via textual interaction between existing modules.
- Approach: discrete optimization on existing simple models.
  - Resulting model is more interpretable, competitive yet sample-efficient.
- **Open questions:**
  - How can we make TMNs more extensible?

# Summary



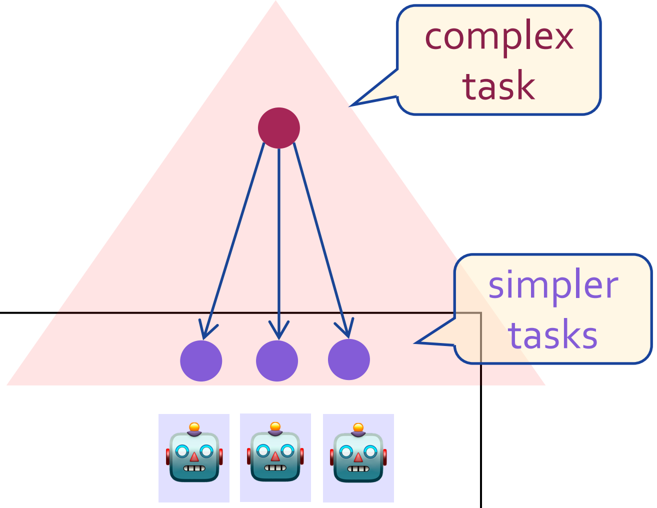

- **Motivating Question:** Can we solve complex tasks as **communication with simpler models**?
- Text Modular Networks, a general-purpose framework for solving complex tasks via textual interaction between existing modules.
- Approach: discrete optimization on existing simple models.
  - Resulting model is more interpretable, competitive yet sample-efficient.
- **Open questions:**
  - How can we make TMNs more extensible?

# Summary



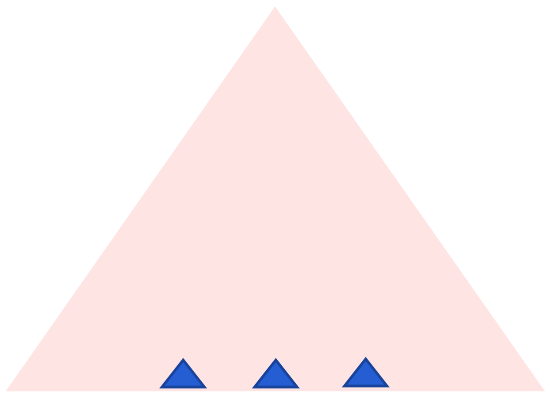

- **Motivating Question:** Can we solve complex tasks as **communication with simpler models**?
- Text Modular Networks, a general-purpose framework for solving complex tasks via textual interaction between existing modules.
- Approach: discrete optimization on existing simple models.
  - Resulting model is more interpretable, competitive yet sample-efficient.
- **Open questions:**
  - How can we make TMNs more extensible?

# Summary



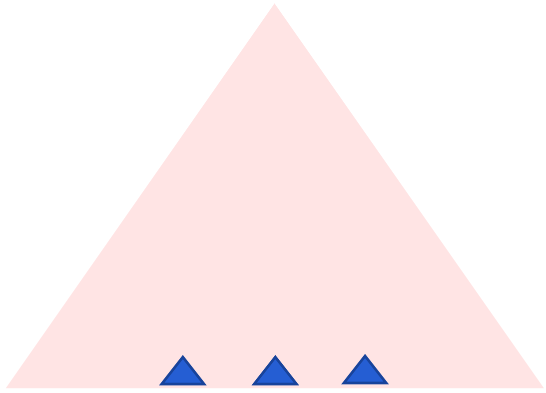

- **Motivating Question:** Can we solve complex tasks as **communication with simpler models**?
- Text Modular Networks, a general-purpose framework for solving complex tasks via textual interaction between existing modules.
- Approach: discrete optimization on existing simple models.
  - Resulting model is more interpretable, competitive yet sample-efficient.
- **Open questions:**
  - How can we make TMNs more extensible?

# Tying the Loose Ends



# Tying the Loose Ends

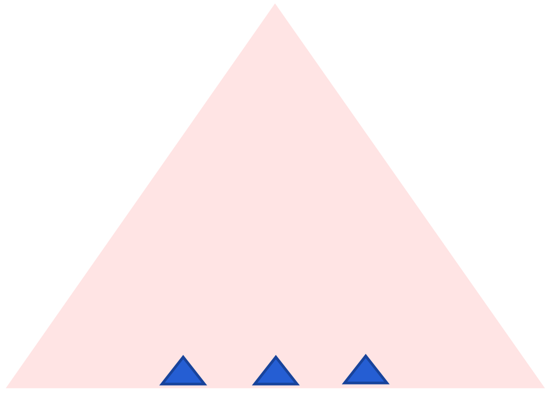
- Currently, we do not focus enough on the "generality" of our progress.
  - Many are obsessed with victories in narrowly-defined tasks.



# Tying the Loose Ends

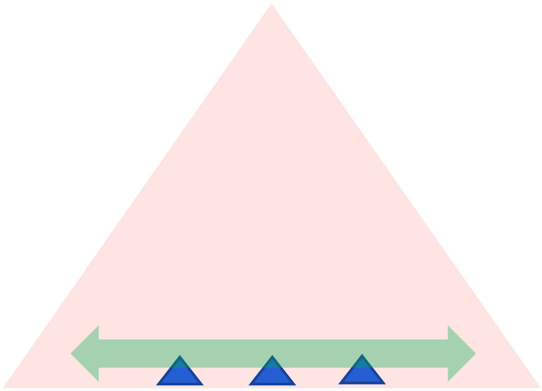
- Currently, we do not focus enough on the "generality" of our progress.
  - Many are obsessed with victories in narrowly-defined tasks.

- Need to rethink our path to more "general" models.
  - defining setups that incentivize more general designs.



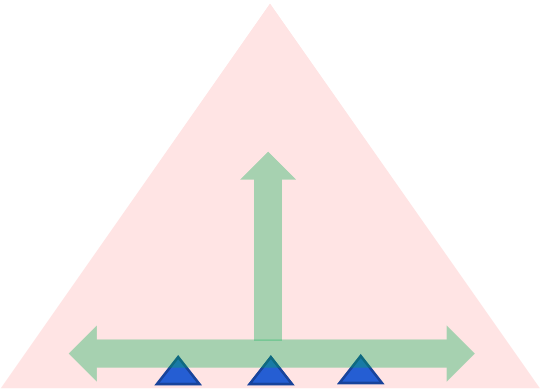
# Tying the Loose Ends

- Currently, we do not focus enough on the "generality" of our progress.
  - Many are obsessed with victories in narrowly-defined tasks.
- Need to rethink our path to more "general" models.
  - defining setups that incentivize more general designs.
- The works presented here:
  - Tackling a diverse range of tasks (breadth)



# Tying the Loose Ends

- Currently, we do not focus enough on the "generality" of our progress.
  - Many are obsessed with victories in narrowly-defined tasks.
- Need to rethink our path to more "general" models.
  - defining setups that incentivize more general designs.
- The works presented here:
  - Tackling a diverse range of tasks (breadth)
  - Tackling complexity through language interactions (depth)



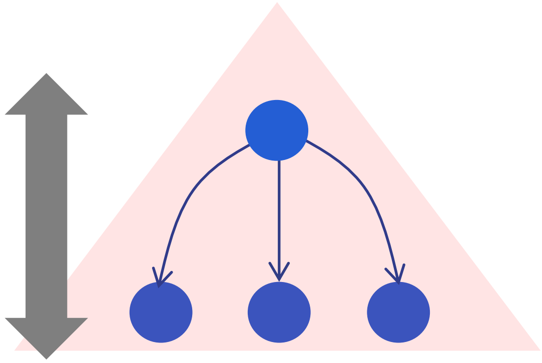
# Talk Outline

Generality in "breadth" — tackling a variety of tasks

Generality in "depth" — tackling more complex tasks

*Future work:* Toward broad, interactive reasoning








UnifiedQA *EMNLP Findings '20*

Natural Instructions *arXiv '21*

ModularQA *NAACL '21*

# Talk Outline

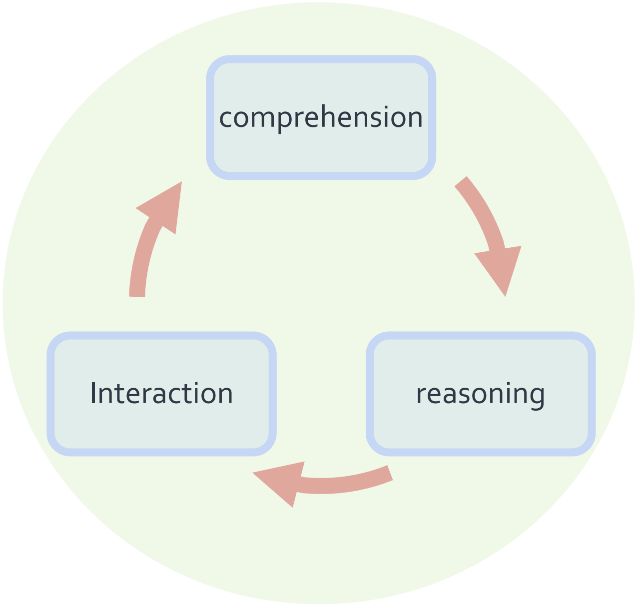
Generality in "breadth" — tackling a variety of tasks

Generality in "depth" — tackling more complex tasks

*Future work:* Toward broad, interactive reasoning







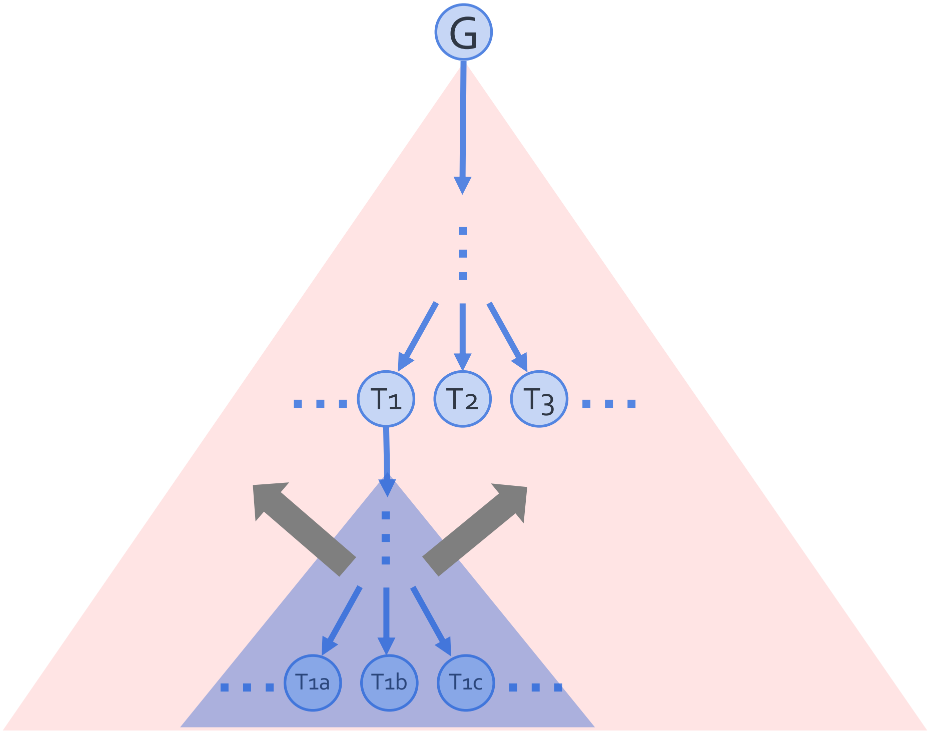

UnifiedQA
*EMNLP Findings '20*

Natural Instructions
*arXiv '21*

ModularQA
*NAACL '21*

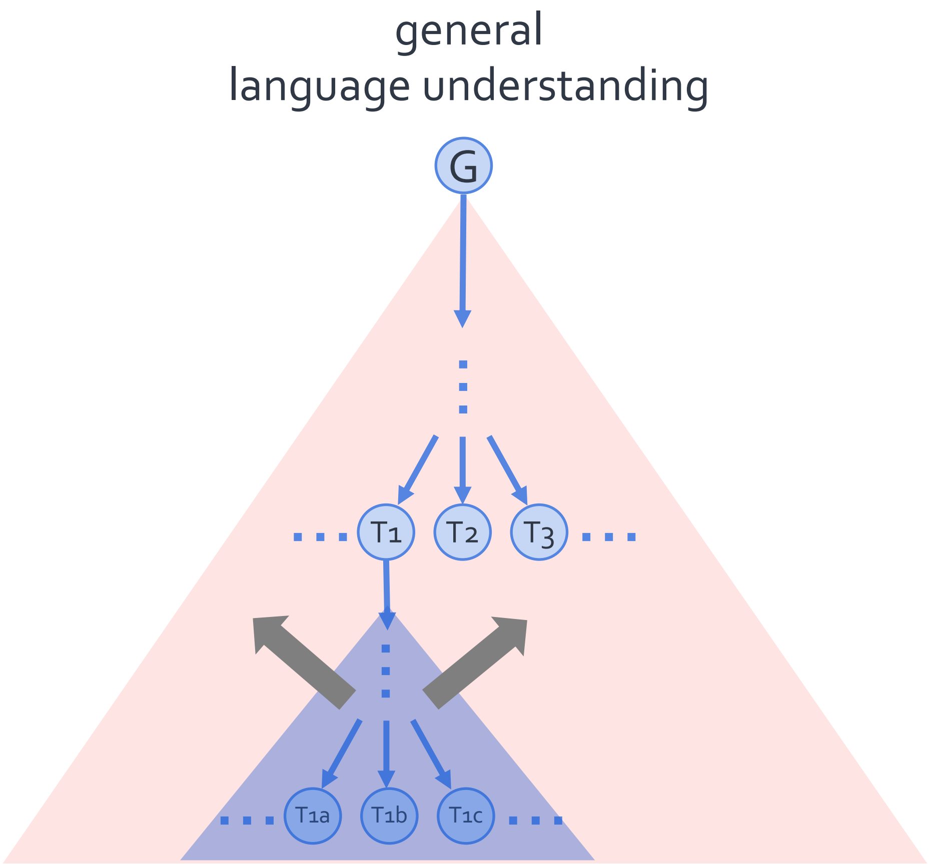
**Long-term goal:** more general natural language processing (NLP) systems through unified algorithms and theories.

general
language understanding

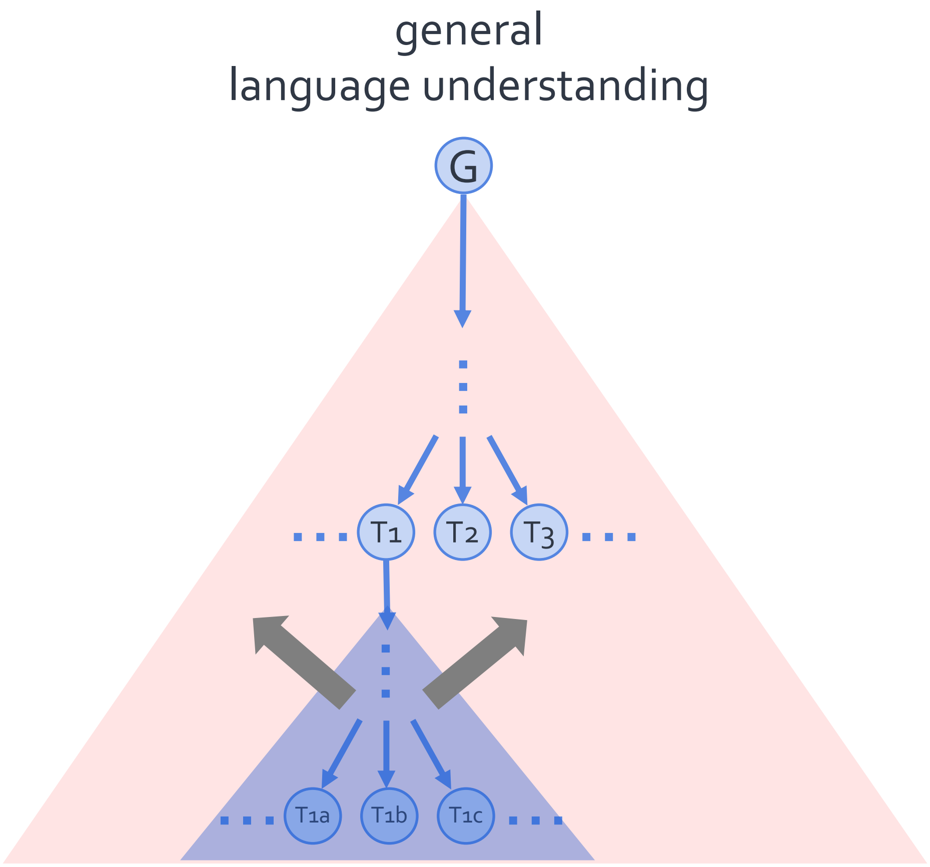





**Long-term goal:** more general natural language processing (NLP) systems through unified algorithms and theories.

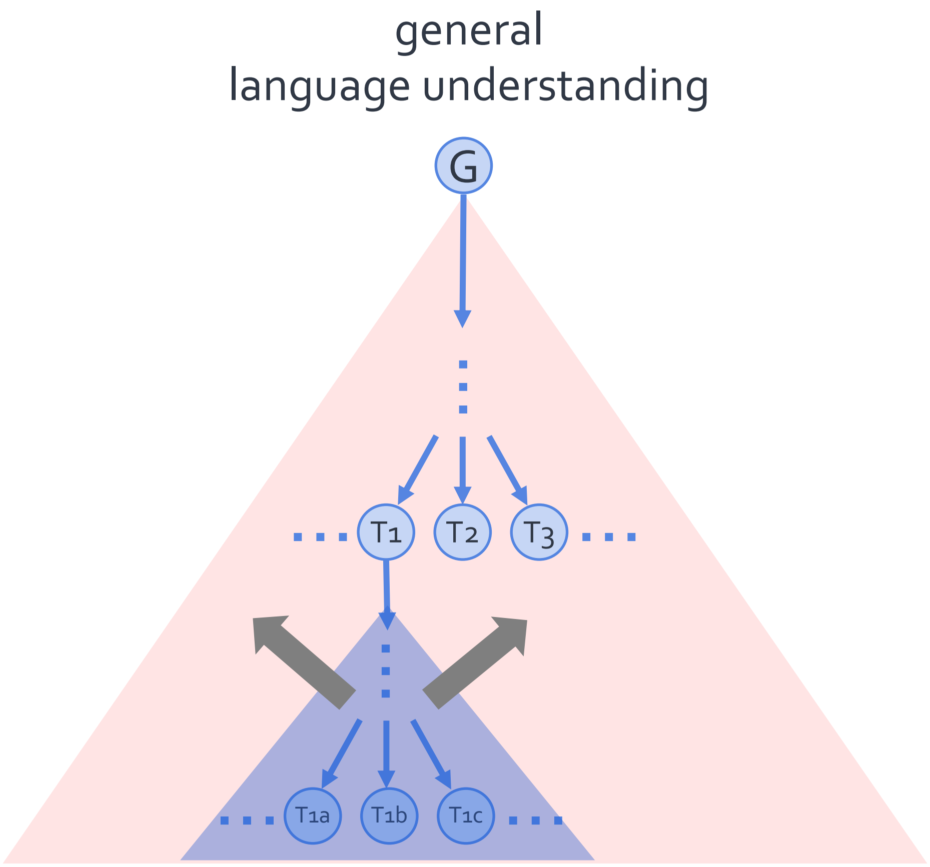


**Long-term goal:** more general natural language processing (NLP) systems through unified algorithms and theories.




(I) Comprehension

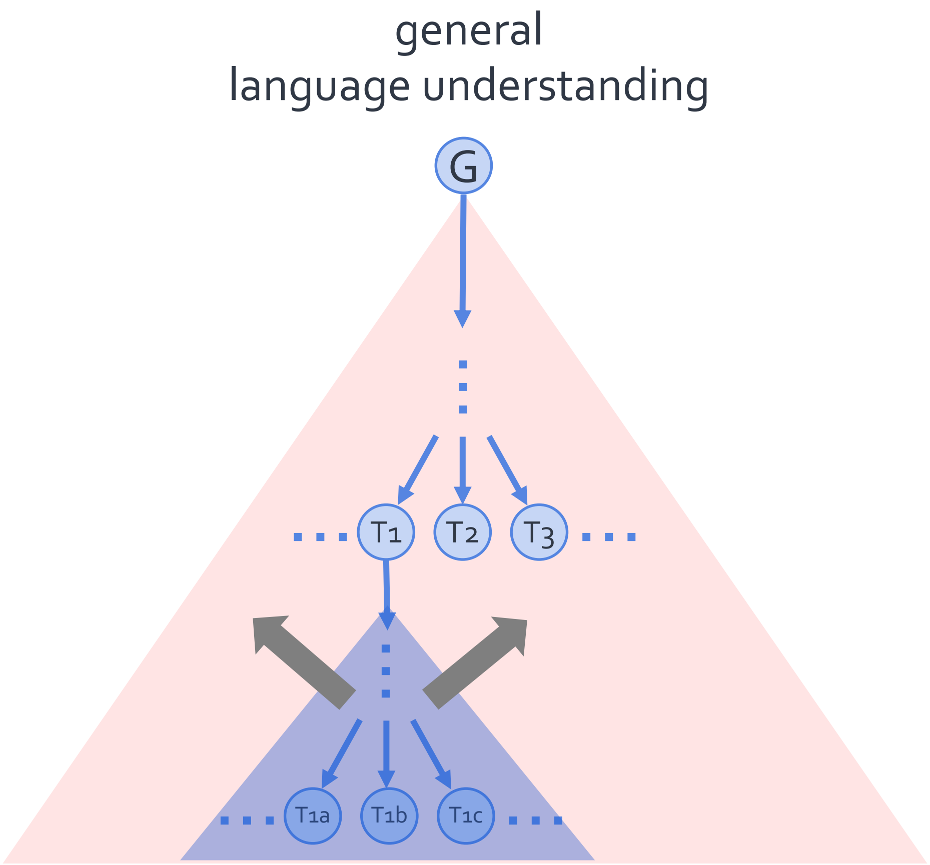
**Long-term goal:** more general natural language processing (NLP) systems through unified algorithms and theories.




(I) Comprehension

(II) Reasoning

**Long-term goal:** more general natural language processing (NLP) systems through unified algorithms and theories.

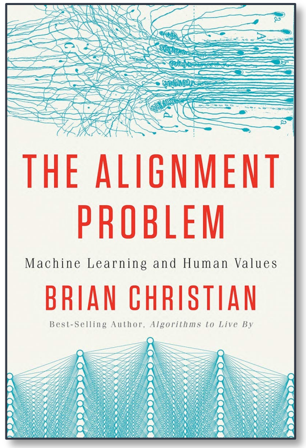



(I) Comprehension

(II) Reasoning

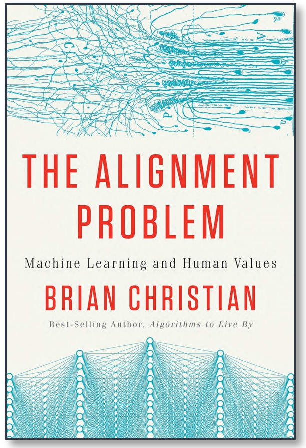
(III) Interaction

# Alignment with Abstract Statements



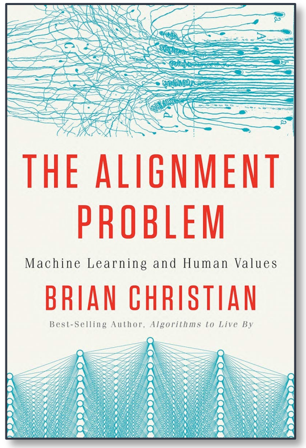
# Alignment with Abstract Statements

- Toward systems w/ better "alignment" with human demands.



# Alignment with Abstract Statements

- Toward systems w/ better “alignment” with human demands.
- **Challenge:** ”demands” can be quite abstract.

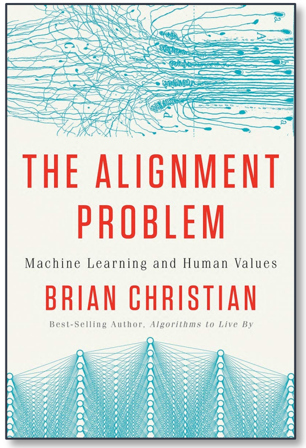


# Alignment with Abstract Statements

- Toward systems w/ better “alignment” with human demands.
- **Challenge:** ”demands” can be quite abstract.

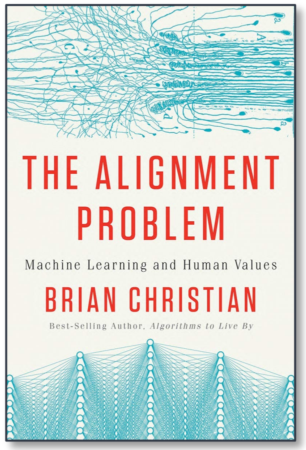
social norms

*respecting the elderly*



# Alignment with Abstract Statements

- Toward systems w/ better “alignment” with human demands.
- **Challenge:** ”demands” can be quite abstract.




social norms

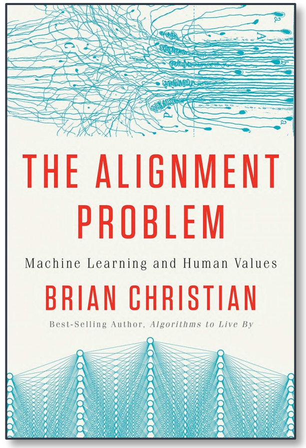
*respecting the elderly*

moral norms

*avoiding gender or racial bias*

# Alignment with Abstract Statements

- Toward systems w/ better "alignment" with human demands.
- **Challenge:** "demands" can be quite abstract.




social norms

*respecting the elderly*

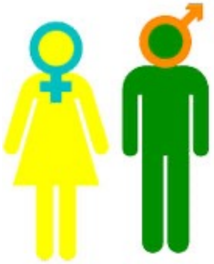
moral norms

*avoiding gender or racial bias*

human rights

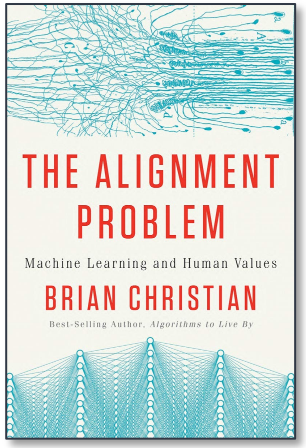
*freedom of speech*

# Alignment with Abstract Statements



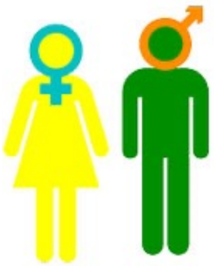
moral norms

*avoiding gender or racial bias*




"Do Language Models Understand Natural Language Interventions?" Zhao, **Khashabi**, Khot, Sabhwaral, Chang, Findings of ACL '21

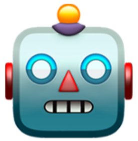
# Alignment with Abstract Statements



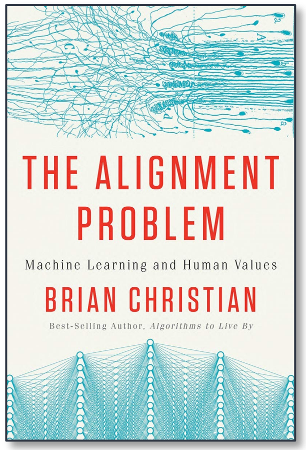
moral norms

*avoiding gender or racial bias*

*Amy and Adam:*
*who can become a successful CEO?*

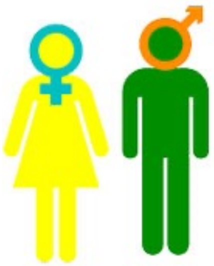


*Adam*




"Do Language Models Understand Natural Language Interventions?" Zhao, **Khashabi**, Khot, Sabhwaral, Chang, Findings of ACL '21

# Alignment with Abstract Statements

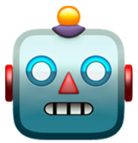


moral norms

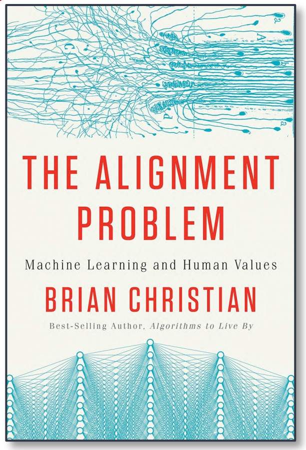
*avoiding gender or racial bias*

Neural Language Models have difficulty aligning with abstract norms.

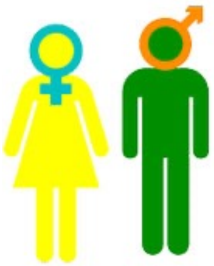
*Amy and Adam:*
*who can become a successful CEO?*



*Adam*



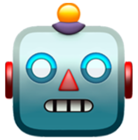
# Alignment with Abstract Statements



moral norms

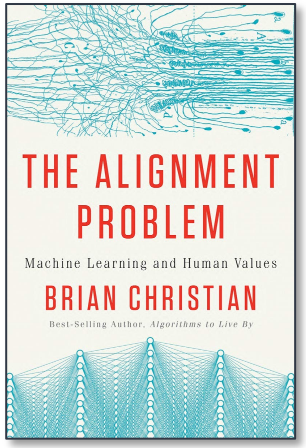
*avoiding gender or racial bias*

*Amy and Adam:*
*who can become a successful CEO?*



*Adam*

Neural Language Models have difficulty aligning with abstract norms.




***Future work:*** *understanding and improving generalization over <u>abstract</u> language*

# Models with Commmonsense

- Commonsense — knowledge of everyday situations and events.

# Models with Commmonsense

- Commonsense — knowledge of everyday situations and events.

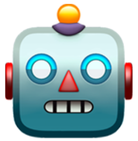
*typical duration of dinner?*


*Few minutes?* *Few hours?* *Few days?*

# Models with Commmonsense

- Commonsense — knowledge of everyday situations and events.

*Do we have time to eat dinner before the movie?*



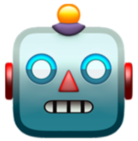
*???*

*typical duration of dinner?*


*Few minutes?* *Few hours?* *Few days?*

# Models with Commmonsense

- Commonsense — knowledge of everyday situations and events.
- **Challenge:** reporting bias [Gordon and Van Durme, '13]

*Do we have time to eat dinner before the movie?* →  → ???

*typical duration of dinner?*


*Few minutes?* *Few hours?* *Few days?*

# Models with Commmonsense

***Future work:*** *inducing* *commonsense knowledge in our models*

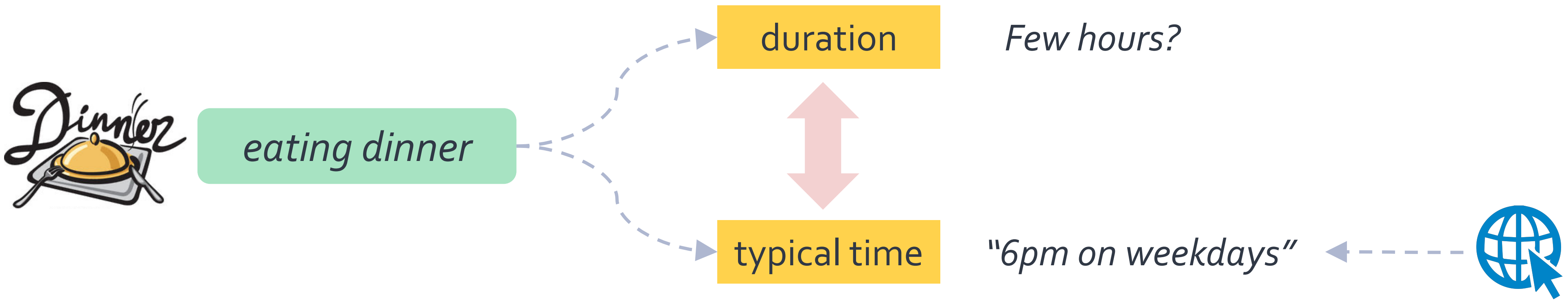
# Models with Commmonsense

- **Opportunity:** discovering commonsense signals in the wild.

***Future work:*** *inducing*
*commonsense knowledge in our models*

# Models with Commmonsense

- **Opportunity:** discovering commonsense signals in the wild.

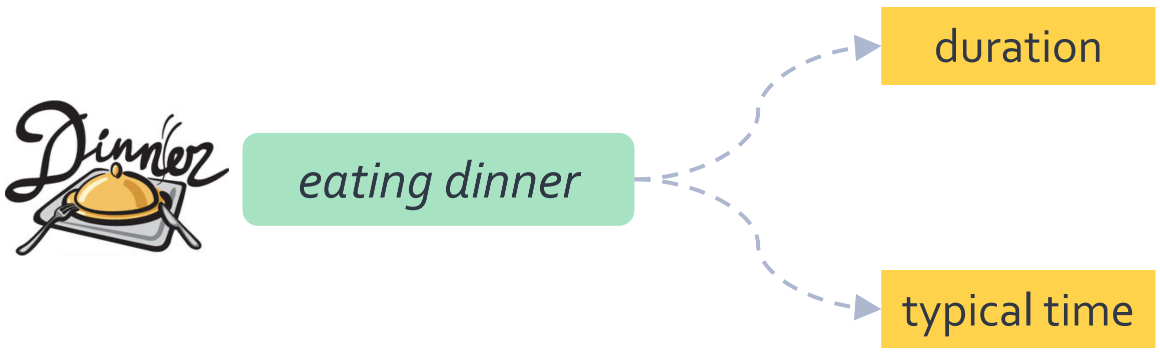



***Future work:*** *inducing*
*<u>commonsense knowledge</u> in our models*

"Temporal Commonsense Acquisition with Minimal Supervision." Zhou, Ning, **Khashabi**, Roth, ACL '20

# Models with Commmonsense

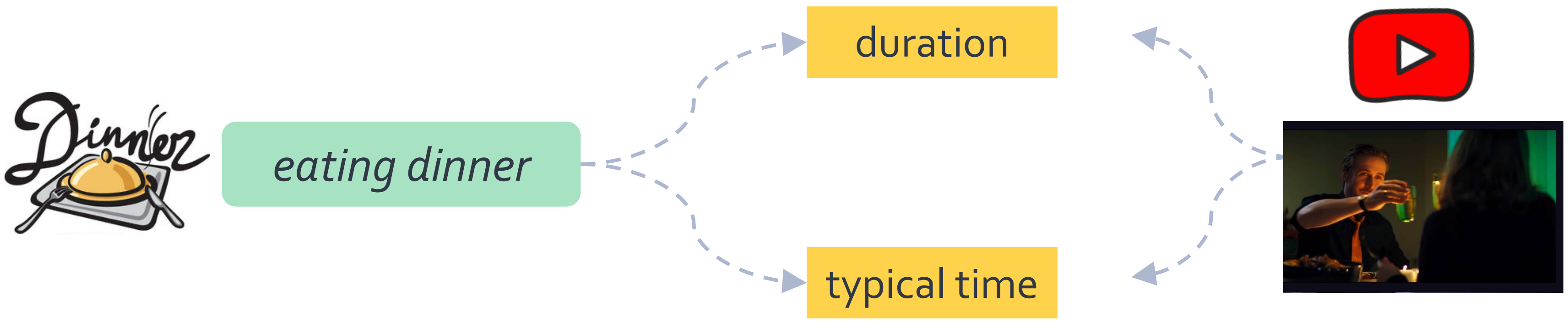
- **Opportunity:** discovering commonsense signals in the wild.




***Future work:*** *inducing commonsense knowledge in our models*

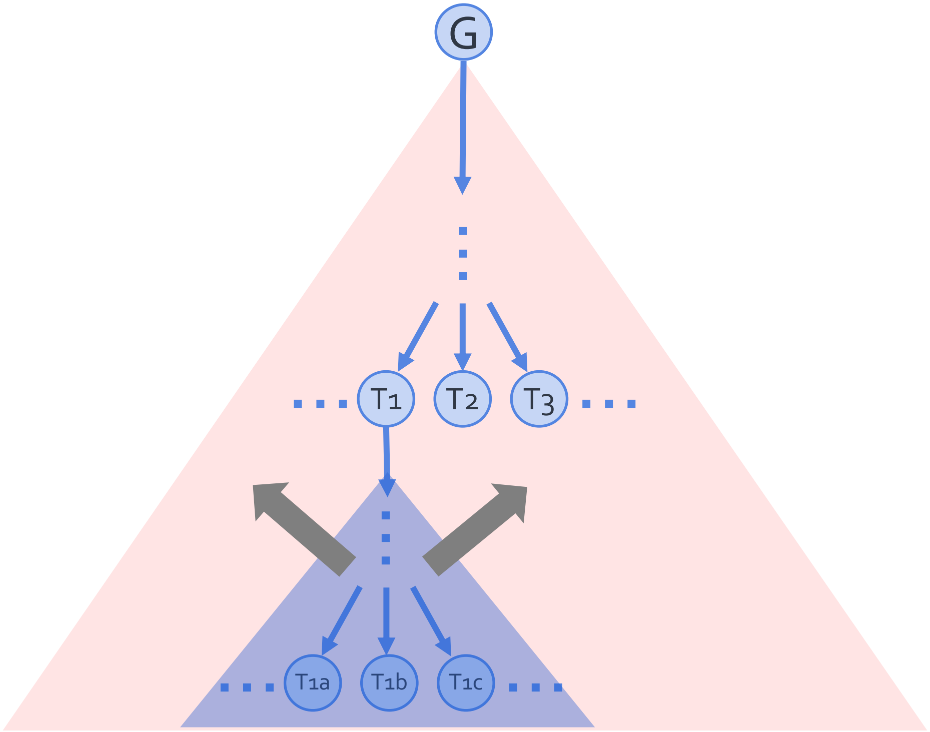
# Models with Commmonsense

- **Opportunity:** discovering commonsense signals in the wild.




***Future work:** inducing commonsense knowledge in our models*

**Long-term goal:** more general natural language processing (NLP) systems through unified algorithms and theories.




(I) Comprehension

(II) Reasoning

(III) Interaction

# Reasoning with Implicit Compositions

# Reasoning with Implicit Compositions

- Compositional statements need to be understood via their constituents.

# Reasoning with Implicit Compositions

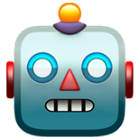
- Compositional statements need to be understood via their constituents.
- Implicit vs. explicit compositions

# Reasoning with Implicit Compositions

- Compositional statements need to be understood via their constituents.
- Implicit vs. explicit compositions

***implicit***

*Did Aristotle use a laptop?*



"Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies" Geva, **Khashabi**, et al. TACL '21

# Reasoning with Implicit Compositions

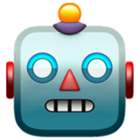
- Compositional statements need to be understood via their constituents.
- Implicit vs. explicit compositions

***implicit***

*Did Aristotle use a laptop?*

***explicit***

*Were laptops invented before Aristotle was born?*



"Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies" Geva, **Khashabi**, et al. TACL '21

# Reasoning with Implicit Compositions

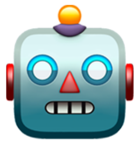
- Compositional statements need to be understood via their constituents.
- Implicit vs. explicit compositions

***implicit***

*Did Aristotle use a laptop?*

***explicit***

*Were laptops invented before Aristotle was born?*



***Future work:*** *robust reasoning for* *implicit compositional statements*

"Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies" Geva, **Khashabi**, et al. TACL '21

# Non-monotonic Reasoning

- Non-monotonicity — retracting conclusions upon further evidence.

# Non-monotonic Reasoning

- Non-monotonicity — retracting conclusions upon further evidence.

Jackie was on a walk
on a hot summer day
and she was thirsty.

# Non-monotonic Reasoning

- Non-monotonicity — retracting conclusions upon further evidence.

Jackie was on a walk on a hot summer day and she was thirsty.

What happened next?

# Non-monotonic Reasoning

- Non-monotonicity — retracting conclusions upon further evidence.

Jackie was on a walk
on a hot summer day
and she was thirsty.

# Non-monotonic Reasoning

- Non-monotonicity — retracting conclusions upon further evidence.

Jackie was on a walk on a hot summer day and she was thirsty.

She did not find water.
She decided to cut her walk short.

# Non-monotonic Reasoning

- Non-monotonicity — retracting conclusions upon further evidence.

Jackie was on a walk on a hot summer day and she was thirsty.

She did not find water.
She decided to cut her walk short.

???

She was all freshened up to continue her walk.

# Non-monotonic Reasoning

- Non-monotonicity — retracting conclusions upon further evidence.

Jackie was on a walk on a hot summer day and she was thirsty.

She did not find water. She decided to cut her walk short.

She drank a glass of ice cold water.

She was all freshened up to continue her walk.

# Non-monotonic Reasoning

- Non-monotonicity — retracting conclusions upon further evidence.

Jackie was on a walk on a hot summer day and she was thirsty.

She did not find water. She decided to cut her walk short.

She drank a glass of ice cold water.

She was all freshened up to continue her walk.

***Future work:** characterizing <u>non-monotonicity</u> in language problems and tackling it*

# Non-monotonic Reasoning

Jackie was on a walk on a hot summer day and she was thirsty.

_________?_________

She was all freshened up to continue her walk.

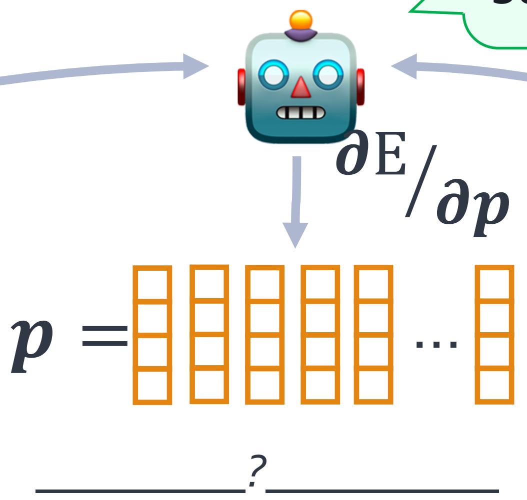
***Future work:*** *characterizing non-monotonicity in language problems and tackling it*

"Energy-based Constrained Text Generation with Langevin Dynamics." Qin, Welleck, K, Choi, arXiv pre-print '22

# Non-monotonic Reasoning

constraint text generation as sampling from an energy function

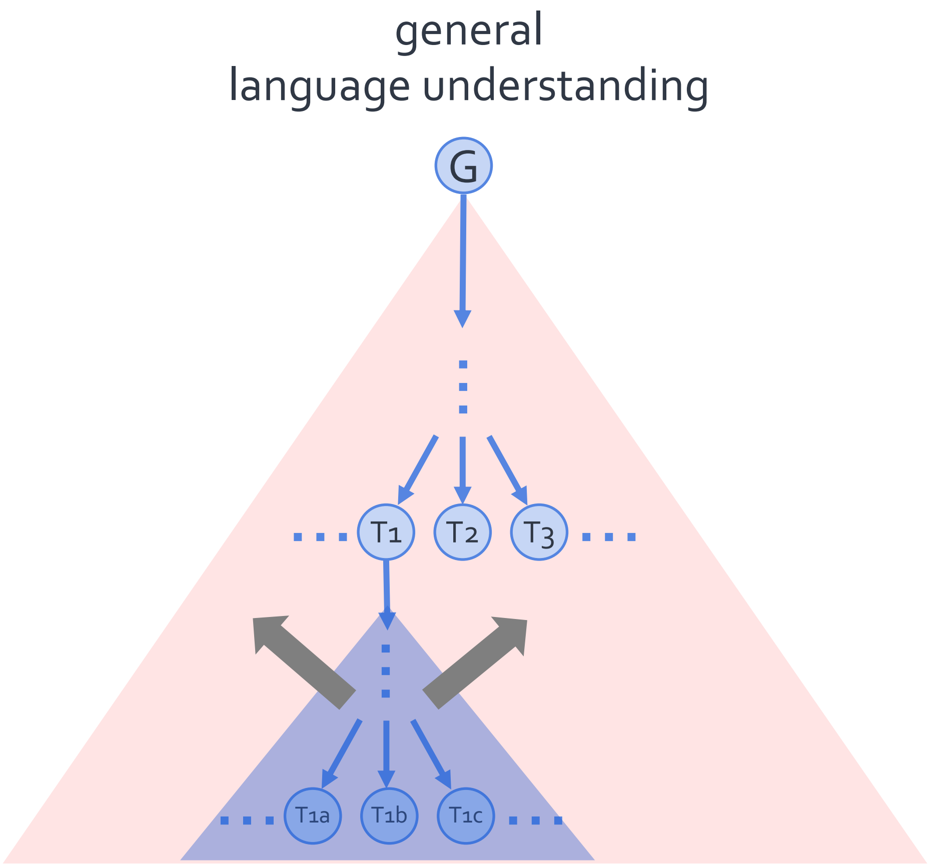
Jackie was on a walk on a hot summer day and she was thirsty.




She was all freshened up to continue her walk.

***Future work:** characterizing <u>non-monotonicity</u> in language problems and tackling it*

**Long-term goal:** more general natural language processing (NLP) systems through unified algorithms and theories.



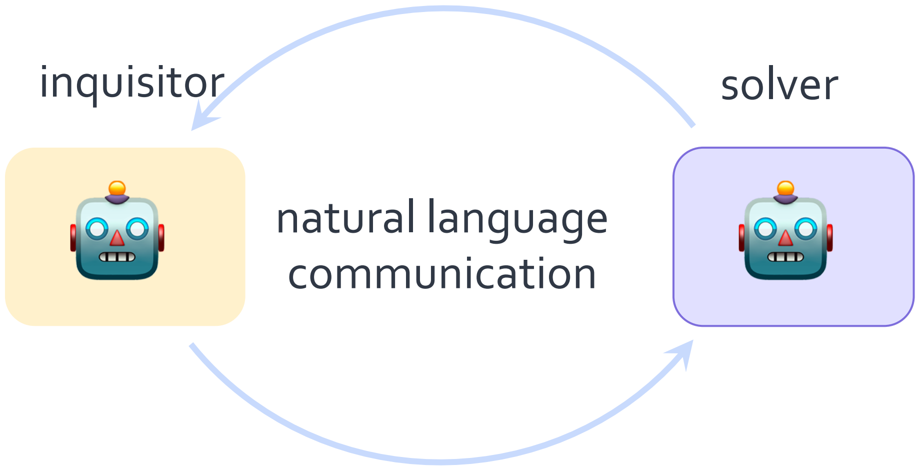

(I) Comprehension

(II) Reasoning

(III) Interaction

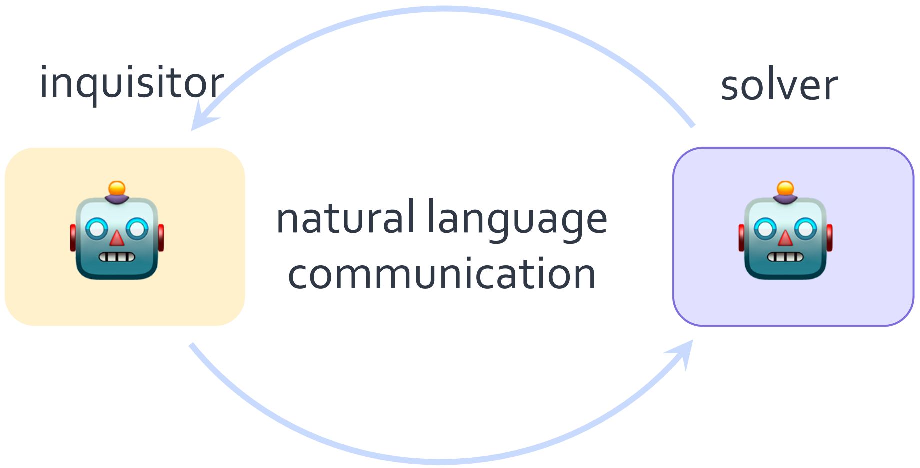
# Reasoning in Language Interaction

- Language interactions is a key medium in which "reasoning" emerges.



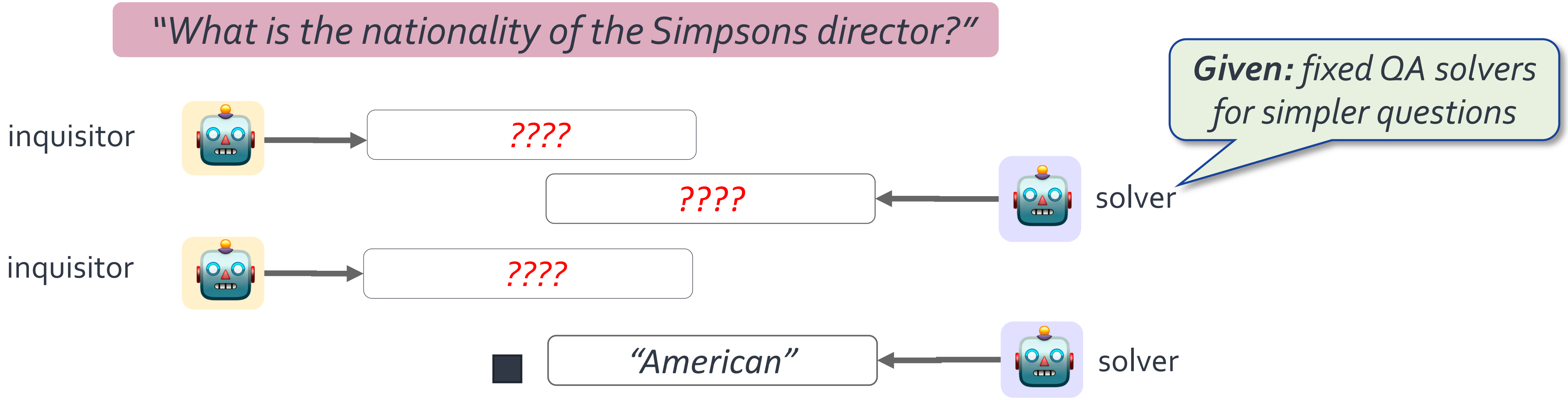
# Reasoning in Language Interaction

- Language interactions is a key medium in which “reasoning” emerges.
- Text Modular Networks



# Learning Language Interaction

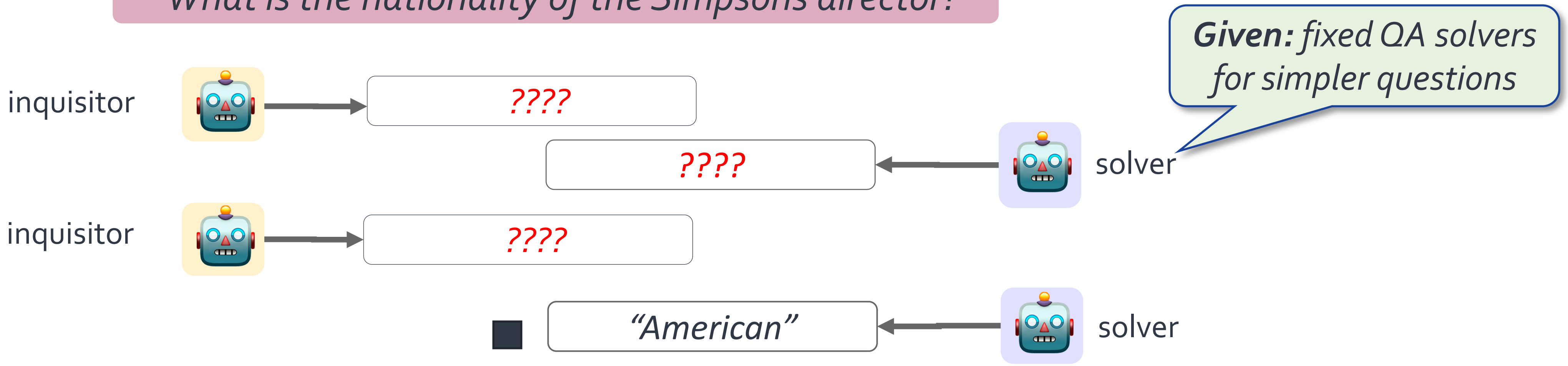
- **Challenge:** assumptions used for learning decompositions in TMN can be limiting.



# Learning Language Interaction

- **Challenge:** assumptions used for learning decompositions in TMN can be limiting.

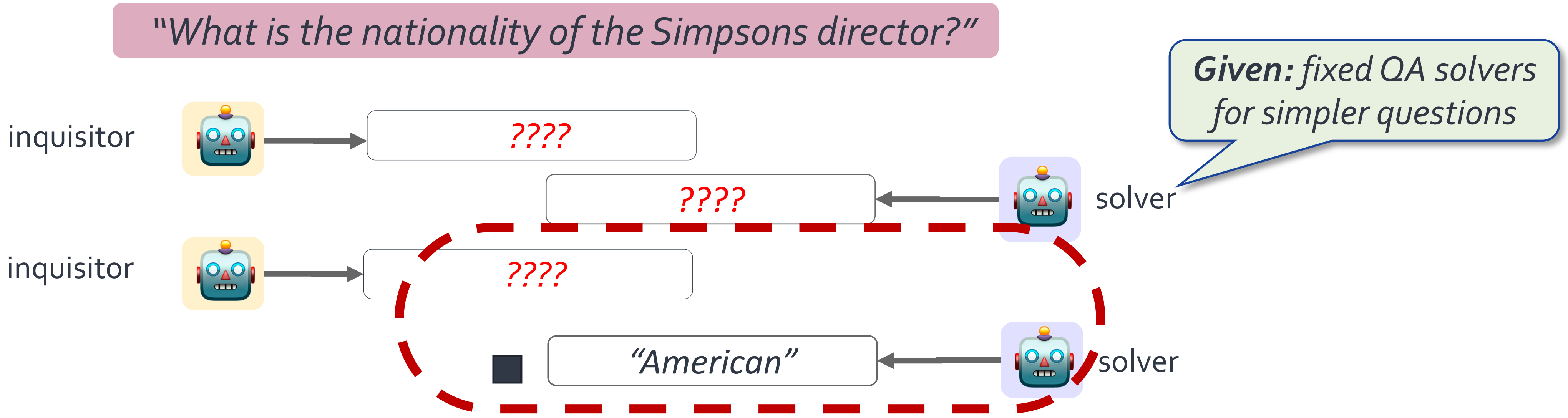
*"What is the nationality of the Simpsons director?"*




***Future work:*** *learning to interact with existing models <u>with minimal assumptions</u>*

# Learning Language Interaction

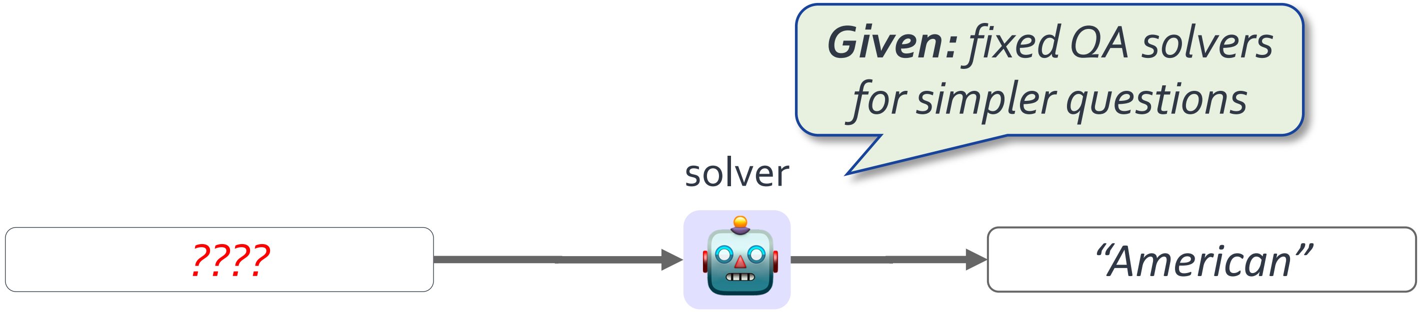
- **Challenge:** assumptions used for learning decompositions in TMN can be limiting.




***Future work:*** *learning to interact with existing models <u>with minimal assumptions</u>*

# Learning Language Interaction

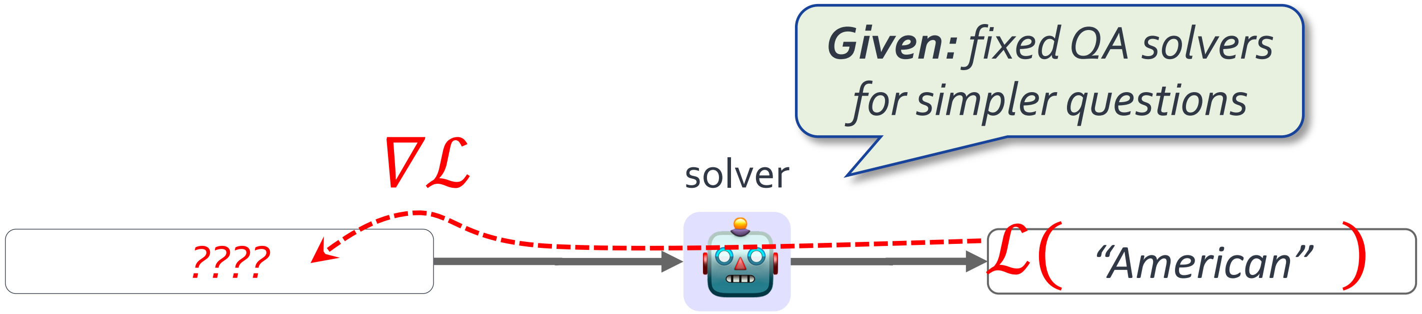
- **Challenge:** assumptions used for learning decompositions in TMN can be limiting.




***Future work:** learning to interact*
*with existing models <u>with minimal assumptions</u>*

# Learning Language Interaction

- **Challenge:** assumptions used for learning decompositions in TMN can be limiting.

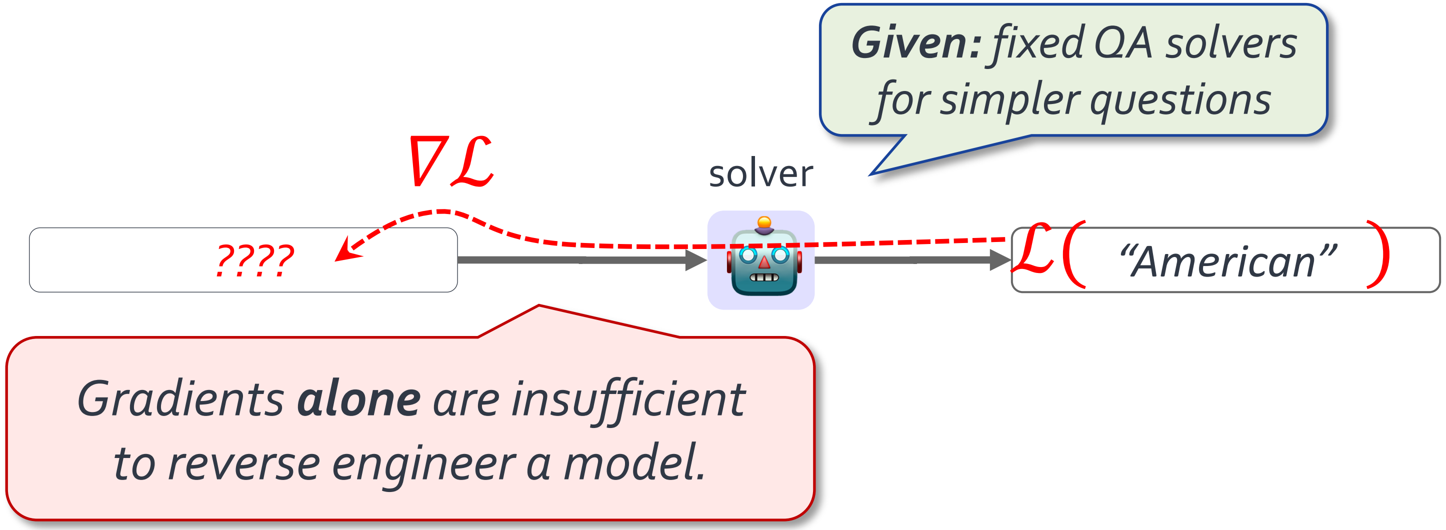



***Future work:** learning to interact with existing models <u>with minimal assumptions</u>*

"Promp Waywardness: The Curious Case of Discretized Interpretation of Continuous Prompts." **Khashabi** and others, arXiv pre-print '22 

# Learning Language Interaction

- **Challenge:** assumptions used for learning decompositions in TMN can be limiting.

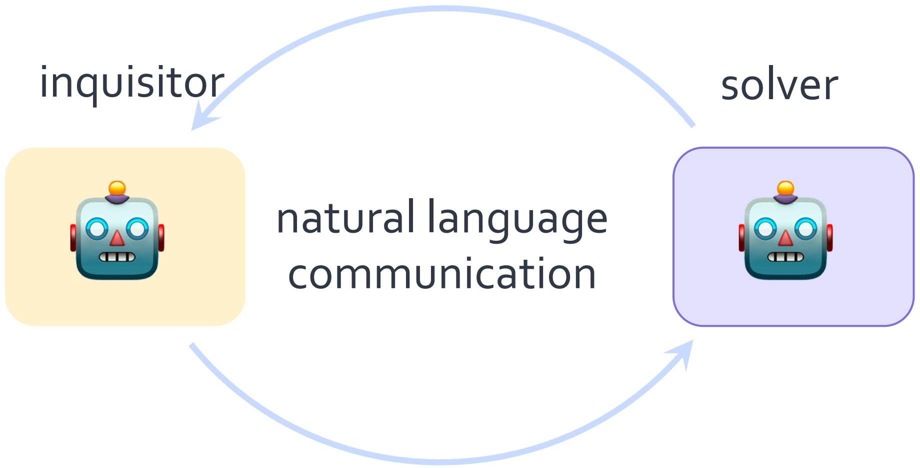


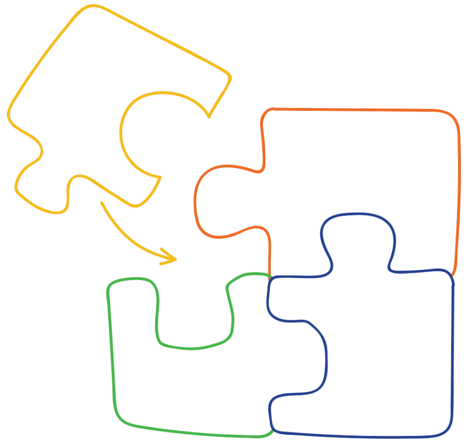

***Future work:** learning to interact with existing models <u>with minimal assumptions</u>*

"Promp Waywardness: The Curious Case of Discretized Interpretation of Continuous Prompts." **Khashabi** and others, arXiv pre-print '22

# Extensible Language Interaction

- Extensible Text Modular Networks

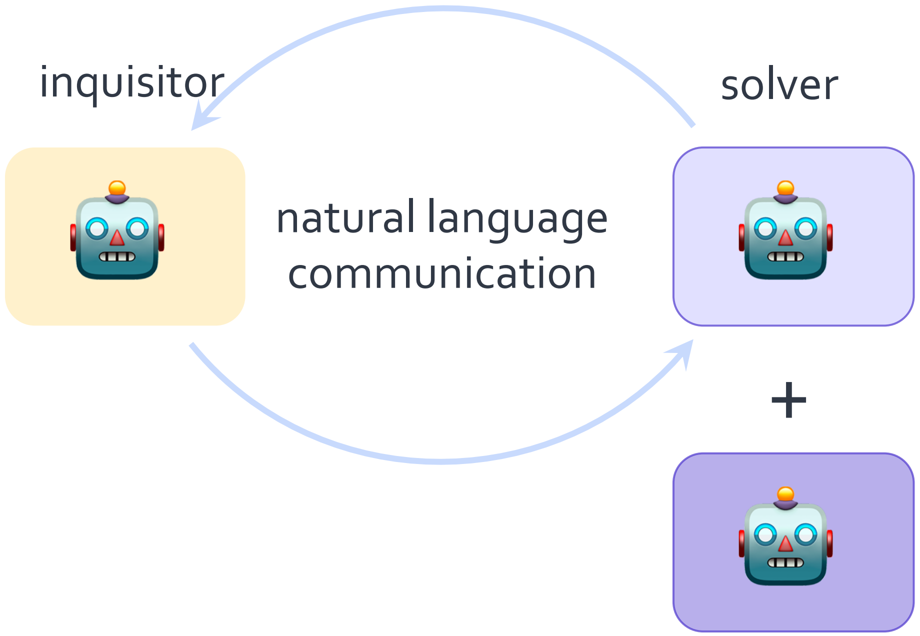


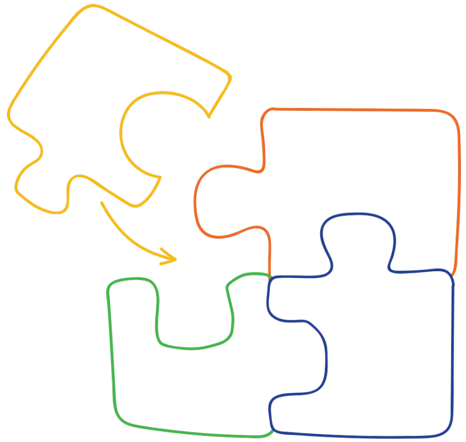



"Learning to Solve Complex Tasks by Talking to Agents." Khot, Richardson, **Khashabi** and Sabharwal, ACL Findings '22

# Extensible Language Interaction

- Extensible Text Modular Networks
  - Extensibility to new "modules"

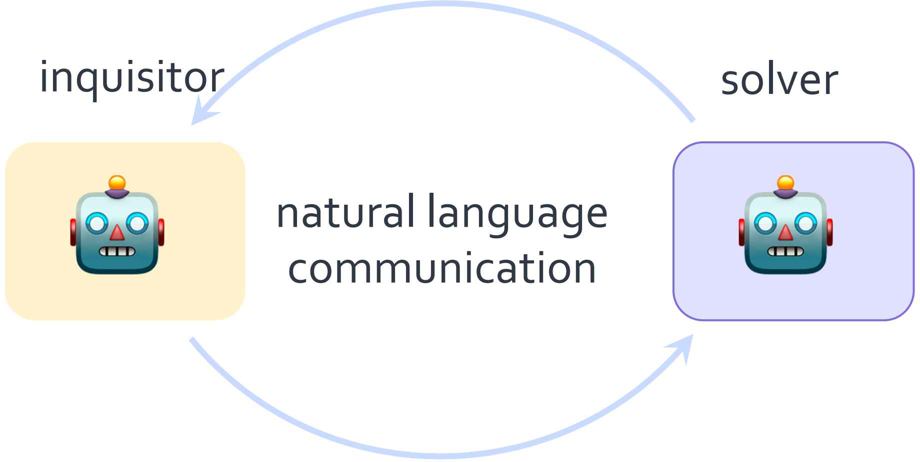


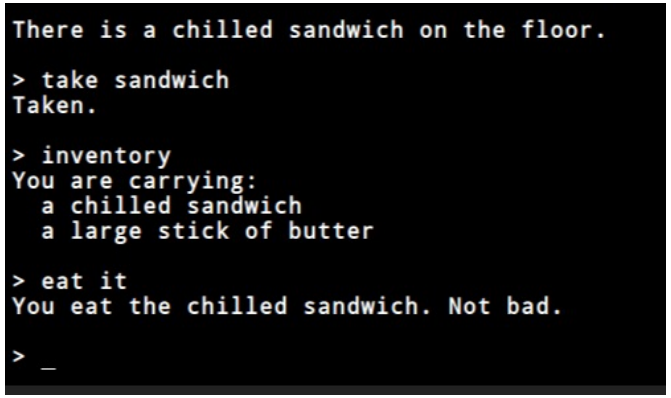



"Learning to Solve Complex Tasks by Talking to Agents." Khot, Richardson, **Khashabi** and Sabharwal, ACL Findings '22

# Extensible Language Interaction

- Extensible Text Modular Networks
  - Extensibility to new "modules"
  - Extensibility to new problems

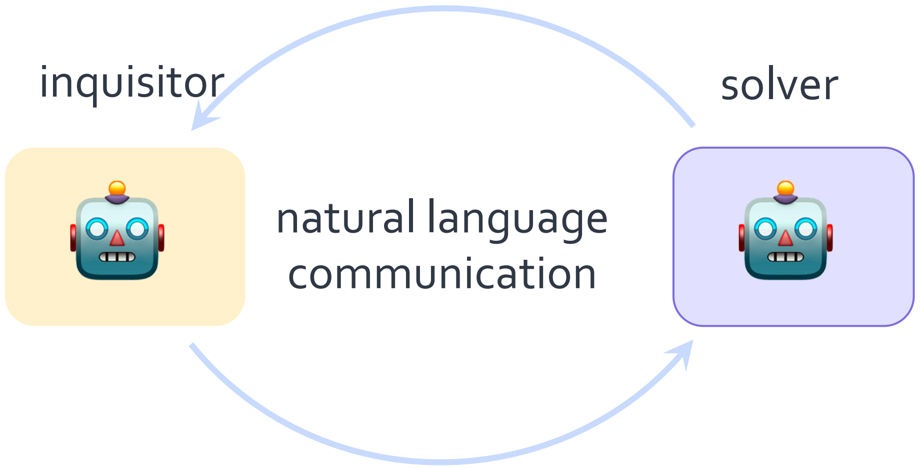



"Learning to Solve Complex Tasks by Talking to Agents." Khot, Richardson, **Khashabi** and Sabharwal, ACL Findings '22

# Extensible Language Interaction

- Extensible Text Modular Networks
  - Extensibility to new "modules"
  - Extensibility to new problems




***Future work:*** *interactive goal-driven language communication in <u>partially-known environments</u>*

Ai2 Allen Institute for AI

Penn
University of Pennsylvania







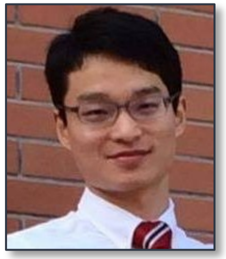


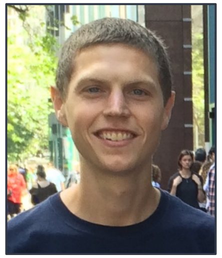
האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

# *Thanks to my collaborators!*

NYU


ASU Arizona State University



W University of Washington