

# Exploring few-shot deception detection via GPT-3's in-context learning

Tanushree Banerjee

# Roadmap for today

Introduce Diplomacy

Past work

Why Diplomacy?

Baselines: Peskov et al. 2020

GPT-3 and in-context learning

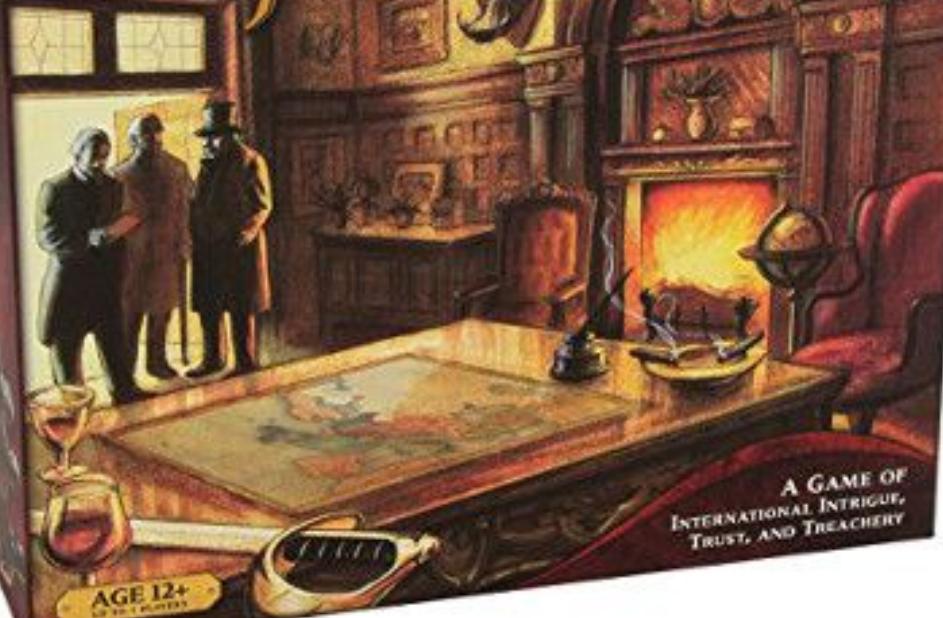
Experiments so far

Short, medium and long term vision

Brainstorming session

AH  
Academy

# Diplomacy



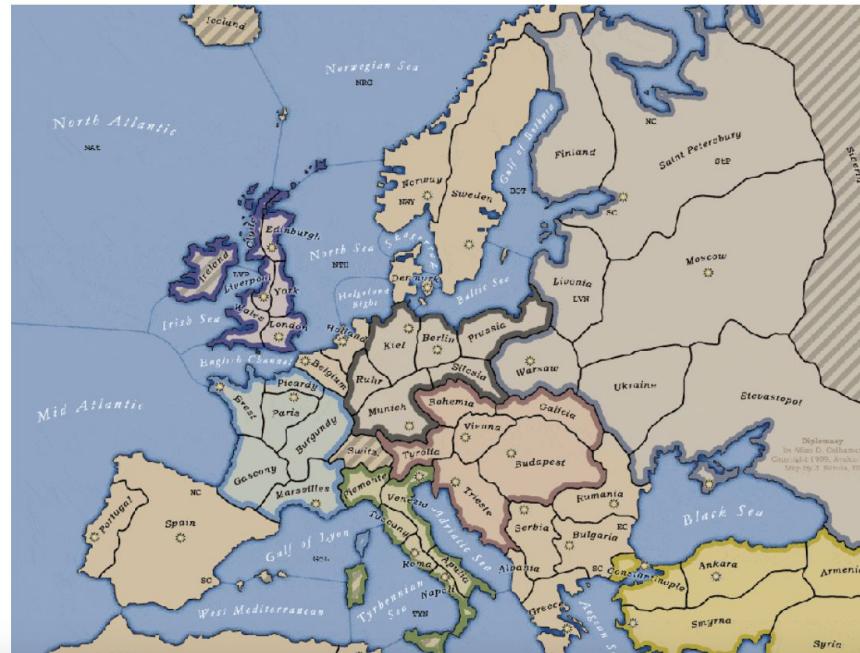
A GAME OF  
INTERNATIONAL INTRIGUE,  
TRUST, AND TREACHERY

AGE 12+  
4 TO 6 PLAYERS

CityLab  
Design

# Diplomacy: The Map That Ruined a Thousand Friendships

Allan Calhamer's brilliant geographic legacy.



ARGUMENT

# The Game That Ruins Friendships and Shapes Careers

For me, Diplomacy is an addictive quarantine hobby.  
For my high school frenemy, it was training for the  
Trump administration.

---

OCTOBER 23, 2020, 7:00 AM



Sunday, 28 March 2021

AI is learning to play Diplomacy, and it's pretty good at it



# No Press Diplomacy: Modeling Multi-Agent Gameplay

---

**Philip Paquette**<sup>1</sup>

[pcpaquette@gmail.com](mailto:pcpaquette@gmail.com)

**Yuchen Lu**<sup>1</sup>

[luyuchen.paul@gmail.com](mailto:luyuchen.paul@gmail.com)

**Steven Bocco**<sup>1</sup>

[stevenbocco@gmail.com](mailto:stevenbocco@gmail.com)

**Max O. Smith**<sup>3</sup>

[max.olan.smith@gmail.com](mailto:max.olan.smith@gmail.com)

**Satya Ortiz-Gagné**<sup>1</sup>

[s.ortizgagne@gmail.com](mailto:s.ortizgagne@gmail.com)

**Jonathan K. Kummerfeld**<sup>3</sup>

[jkummerf@umich.edu](mailto:jkummerf@umich.edu)

**Satinder Singh**<sup>3</sup>

[baveja@umich.edu](mailto:baveja@umich.edu)

**Joelle Pineau**<sup>2</sup>

[jpineau@cs.mcgill.ca](mailto:jpineau@cs.mcgill.ca)

**Aaron Courville**<sup>1</sup>

[aaron.courville@gmail.com](mailto:aaron.courville@gmail.com)

June 2020 on  
ArXiv

# Learning to Play No-Press Diplomacy with Best Response Policy Iteration

---

Thomas Anthony\*, Tom Eccles\*, Andrea Tacchetti, János Kramár, Ian Gemp,  
Thomas C. Hudson, Nicolas Porcel, Marc Lanctot, Julien Pérolat, Richard Everett,  
Roman Werpachowski, Satinder Singh, Thore Graepel and Yoram Bachrach

DeepMind

ICLR 2021

# HUMAN-LEVEL PERFORMANCE IN NO-PRESS DIPLOMACY VIA EQUILIBRIUM SEARCH

**Jonathan Gray\*, Adam Lerer\*, Anton Bakhtin, Noam Brown**

Facebook AI Research

{jsgray, alerer, yolo, noambrown}@fb.com

“No press”: communication disabled

Long-term goal: play Diplomacy “with press”

ACL 2020

# **It Takes Two to Lie: One to Lie, and One to Listen**

**Denis Peskov, Benny Cheng**

**Ahmed Elgohary, Joe Barrow**

Computer Science, University of Maryland

{dpeskov, bcheng96, elgohary, jdbarrow}@umd.edu cristian@cs.cornell.edu

**Cristian Danescu-Niculescu-Mizil**

Information Science

Cornell University

**Jordan Boyd-Graber**

iSchool, Language Science, UMIACS, LSC

University of Maryland

jbg@umiacs.umd.edu



# What is Diplomacy?

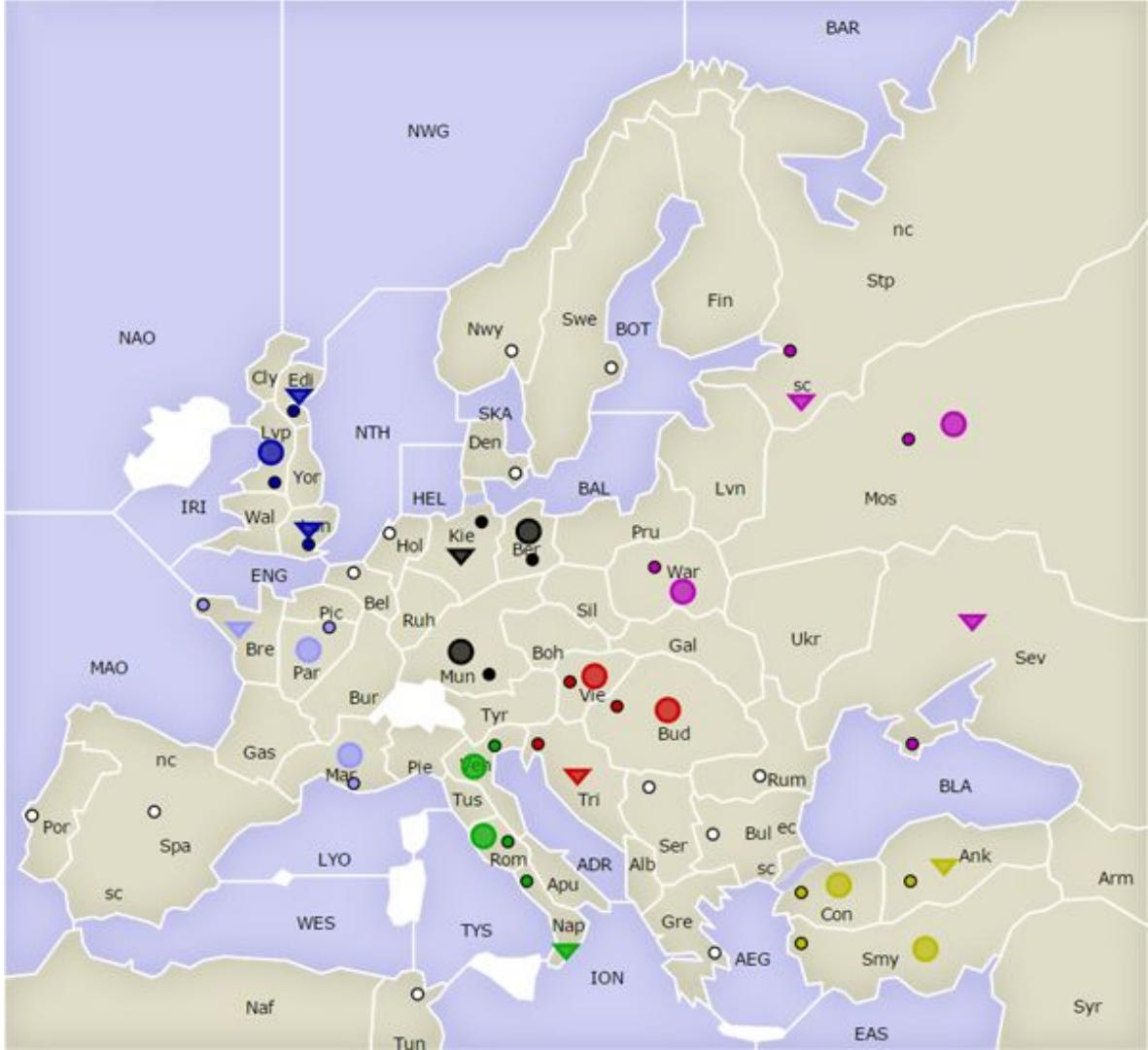
7 players

Each player controls multiple units

Each turn move units simultaneously

Conflict winner = superior force

# Coordination of moves for success

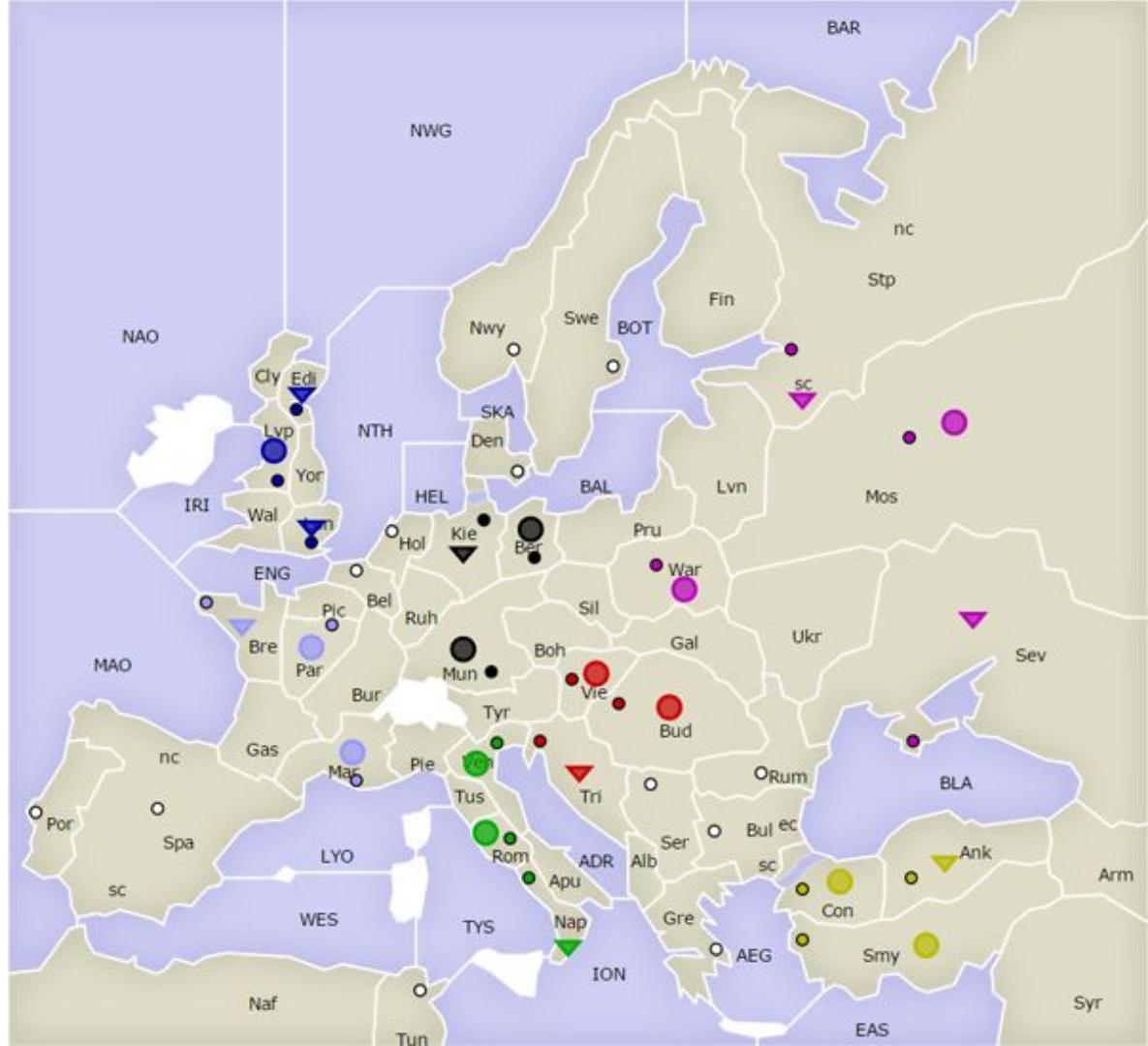


One of the best  
board games of all  
time...

Combination of strategy,  
tactics, negotiation

The complete absence of  
any random luck

Today, >10000 people  
play diplomacy



Rules = Anything goes

A zero-sum game...

# Why Diplomacy?

Understand the language of deception

Mix of competition and collaboration to win

Accentuate dilemmas from multi-agent interactions

Large combinatorial action space

# The dataset

Double annotations by both sender and receiver

Illuminate difference between a deceptive and truthful statement

~17000 messages

12 games

Each message: one word to multiple paragraphs

Specialised user base

Each game could last for > 1 month

Message	Sender's intention	Receiver's percep.
If I were lying to you, I'd smile and say "that sounds great." I'm honest with you because I sincerely thought of us as partners.	Lie	Truth
You agreed to warn me of unexpected moves, then didn't ... You've revealed things to England without my permission, and then made up a story about it after the fact!	Truth	Truth
... I have a reputation in this hobby for being sincere. Not being duplicitous. It has always served me well. ... If you don't want to work with me, then I can understand that ...	Lie	Truth
<i>(Germany attacks Italy)</i>		
Well this game just got less fun	Truth	Truth
For you, maybe	Truth	Truth

# Dataset fields

Speakers

Absolute message index

Messages

Relative\_message index

Sender labels

Seasons

Receiver labels

Years

Game score

Game id

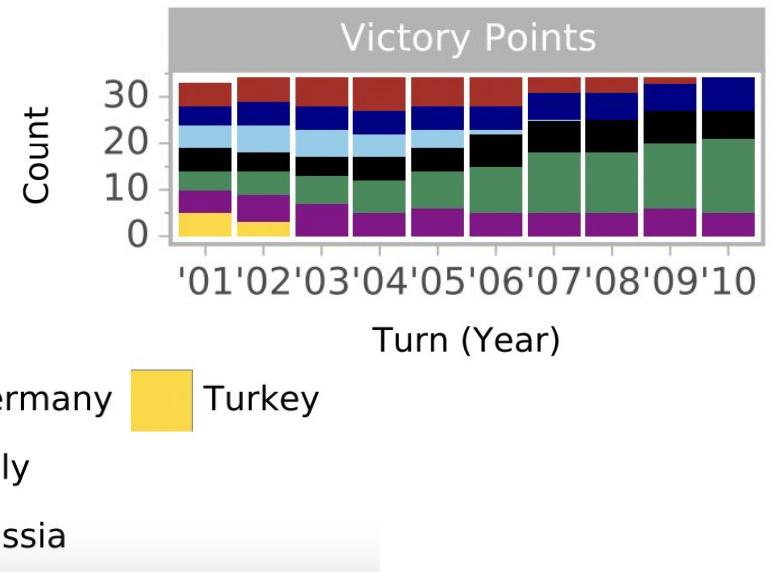
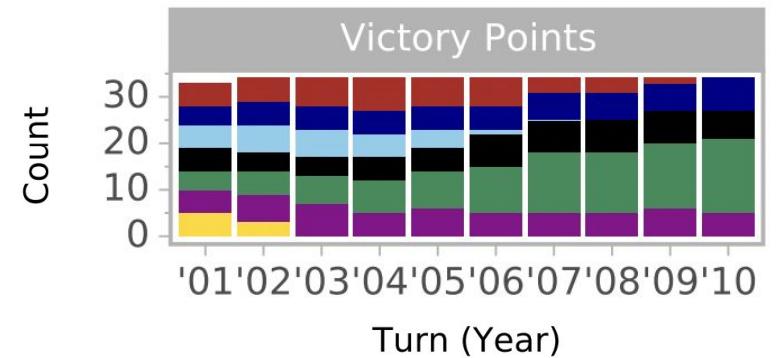
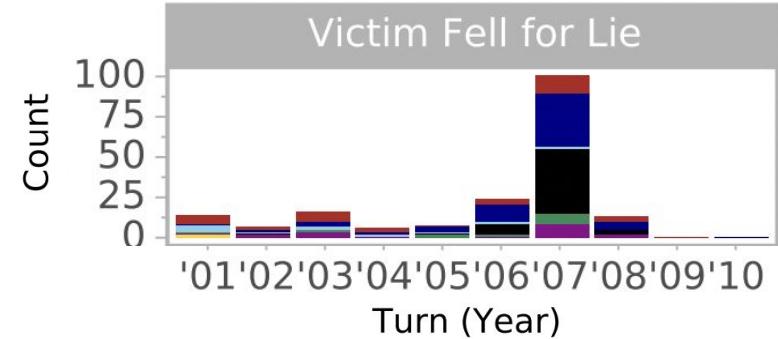
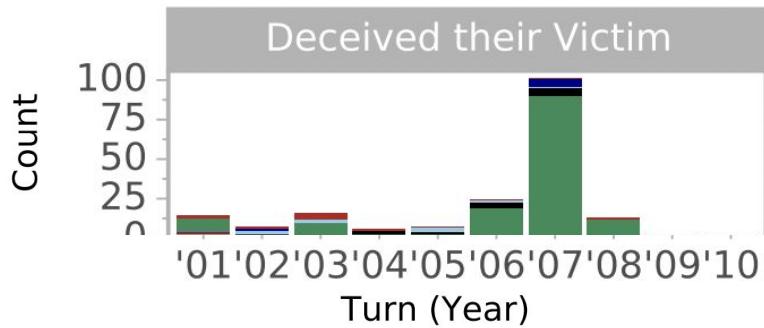


## DiplomacyGame12\_1901\_fall.json

```
{"sc": "\n      \n\n\n\n      Austria 3\n      \n\n\n\n      England 3\n      \n\n\n\n      France 3\n\n      Germany 3\n      \n\n\n\n      Italy 3\n      \n\n\n\n      Russia 4\n      \n\n\n\n      Turkey\n3\n      \n\n\n\n      \n\n\n\n      Orders \n\n\n", "orders": {"Turkey": {"Arm": {"to": "Sev", "type": "MOVE", "result": "SUCCEEDS", "result_reason": "Attack strength is greater"}, "Bul": {"type": "HOLD", "result": "SUCCEEDS", "result_reason": "Unchallenged"}, "Con": {"to": "AEG", "type": "MOVE", "result": "SUCCEEDS", "result_reason": "Attack strength is greater"}, "England": {"NTH": {"to": "Nwy", "result_reason": "Unchallenged", "type": "CONVOY", "result": "SUCCEEDS", "from": "Yor"}, "Lon": {"to": "ENG", "type": "MOVE", "result": "FAILS", "result_reason": "Attack strength is not greater than the prevent strength"}, "Yor": {"to": "Nwy", "type": "MOVE", "result": "SUCCEEDS", "result_reason": "Attack strength is greater"}, "France": {"Bur": {"to": "Mar", "type": "MOVE", "result": "SUCCEEDS", "result_reason": "Attack strength is greater"}, "Bre": {"to": "ENG", "type": "MOVE", "result": "FAILS", "result_reason": "Attack strength is not greater than the prevent strength"}, "Mar": {"to": "Spa", "type": "MOVE", "result": "SUCCEEDS", "result_reason": "Attack strength is greater"}, "Austria": {"Alb": {"to": "Gre", "type": "MOVE", "result": "SUCCEEDS", "result_reason": "Attack strength is greater"}, "Ser": {"to": "Gre", "result_reason": "Unchallenged", "type": "SUPPORT", "result": "SUCCEEDS", "from": "Alb"}, "Vie": {"to": "Tri", "type": "MOVE", "result": "FAILS", "result_reason": "Attack strength is not greater than the prevent strength"}, "Germany": {"Mun": {"to": "Bur", "type": "MOVE", "result": "SUCCEEDS", "result_reason": "Attack strength is greater"}, "Den": {"to": "Swe", "type": "MOVE", "result": "SUCCEEDS", "result_reason": "Attack strength is greater"}, "Kie": {"to": "Hol", "type": "MOVE", "result": "SUCCEEDS", "result_reason": "Attack strength is greater"}, "Italy": {"Ven": {"to": "Tri", "type": "MOVE", "result": "FAILS", "result_reason": "Attack strength is not greater than the prevent strength"}, "ION": {"to": "Tun", "type": "MOVE", "result": "SUCCEEDS", "result_reason": "Attack strength is greater"}, "Pie": {"to": "Tyr", "type": "MOVE", "result": "SUCCEEDS", "result_reason": "Attack strength is greater"}, "Russia": {"BLA": {"type": "HOLD", "result": "SUCCEEDS", "result_reason": "Unchallenged"}, "BOT": {"type": "HOLD", "result": "SUCCEEDS", "result_reason": "Unchallenged"}, "Ukr": {"type": "HOLD", "result": "SUCCEEDS", "result_reason": "Unchallenged"}, "War": {"type": "HOLD", "result": "SUCCEEDS", "result_reason": "Unchallenged"}}, "territories": {"Par": "France", "Smy": "Turkey", "Lon": "England", "Nap": "Italy", "Ven": "Italy", "Mun": "Germany", "War": "Russia", "Mar": "France", "Mos": "Russia", "Kie": "Germany", "Tri": "Austria", "Lvp": "England", "Stp": "Russia", "Ber": "Germany", "Con": "Turkey", "Sev": "Russia", "Rom": "Italy", "Vie": "Austria", "Edi": "England", "Ank": "Turkey", "Bud": "Austria", "Bre": "France"}}}
```



A NEW DECEPTION DATASET



Austria    England    France    Germany    Italy    Russia    Turkey

# Study set up

Found players to design study and play as participants

Recruited participants with no experience

Compensation for participating and completing game

Incentive to tell most amount of lies and win game

Good players asked to play again in future games

Custom discord bot to record messages + annotate sent and received messages

Online platform: Backstabber

england

## New Message

Message

Hey italy! good luck this game. I'm guessing you and Austria will be pals, you and France will be rivals?

Dated

Spring, 1901

Do you think the sender is telling the truth?

1 2

6:39 PM **italy** Well good luck to you too! No idea yet who is a friend. Have you heard anything interesting?

1 2

6:39 PM **BOT Diplomacy** You lied to your opponent.

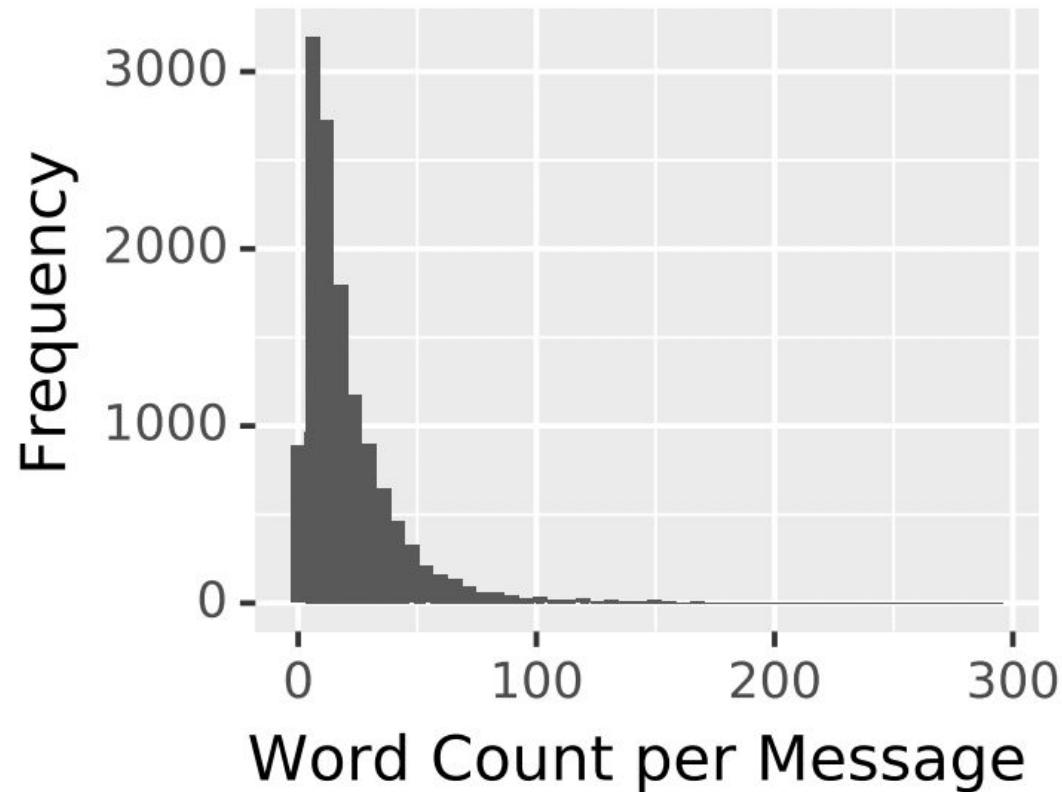
What is a lie?

“Typically, when [someone] lies [they] say what [they] know to be false in an attempt to deceive the listener”

# Dataset statistics

## Long messages...

Average message length: 21 words



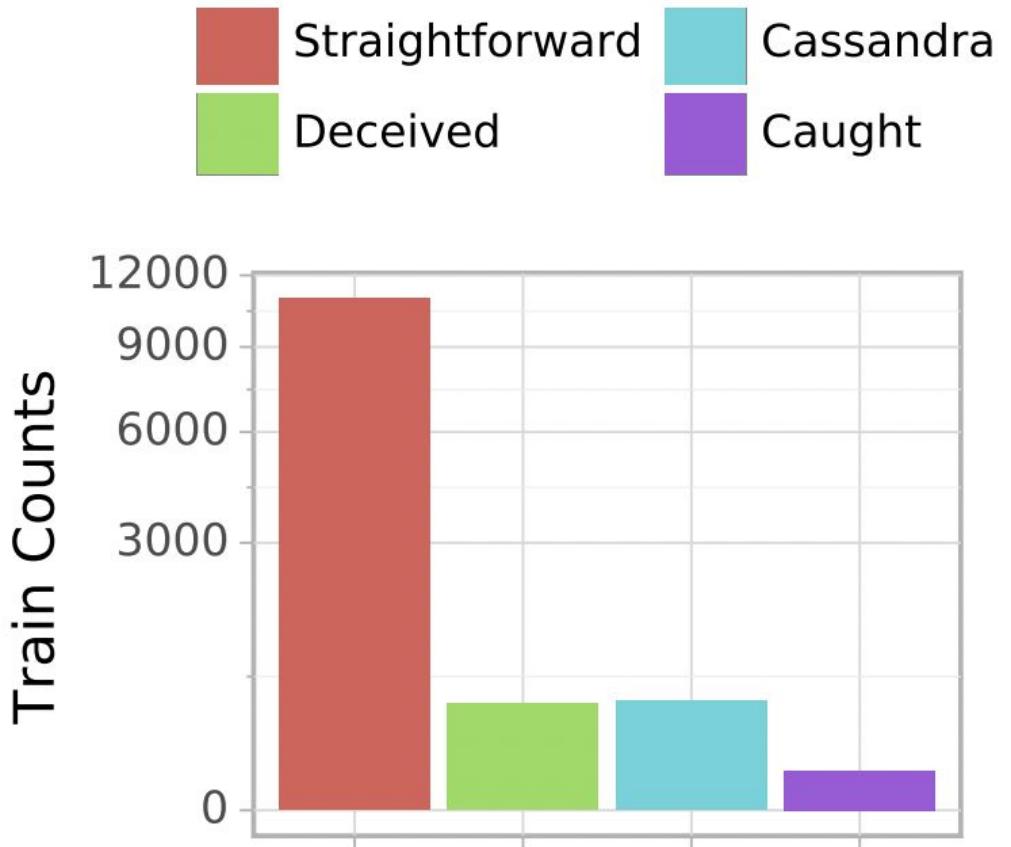
Unequal class distribution...

95% message truthful

⇒ detecting a lie is a difficult task

<b>Category</b>	<b>Value</b>
Message Count	13,132
ACTUAL LIE Count	591
SUSPECTED LIE Count	566
Average # of Words	20.79

Lies often not  
caught...



Sender's intention		Receiver's perception	
		Truth	Lie
Truth	<p><b>Straightforward</b> Salut! Just checking in, letting you know the embassy is open, and if you decide to move in a direction I might be able to get involved in, we can probably come to a reasonable arrangement on cooperation. Bonne journee!</p>	<p><b>Cassandra</b> I don't care if we target T first or A first. I'll let you decide. But I want to work as your partner. ... I literally will not message anyone else until you and I have a plan. I want it to be clear to you that you're the ally I want.</p>	
Lie	<p><b>Deceived</b> You, sir, are a terrific ally. This was more than you needed to do, but makes me feel like this is really a long term thing! Thank you.</p>	<p><b>Caught</b> So, is it worth us having a discussion this turn? I sincerely wanted to work something out with you last turn, but I took silence to be an ominous sign.</p>	

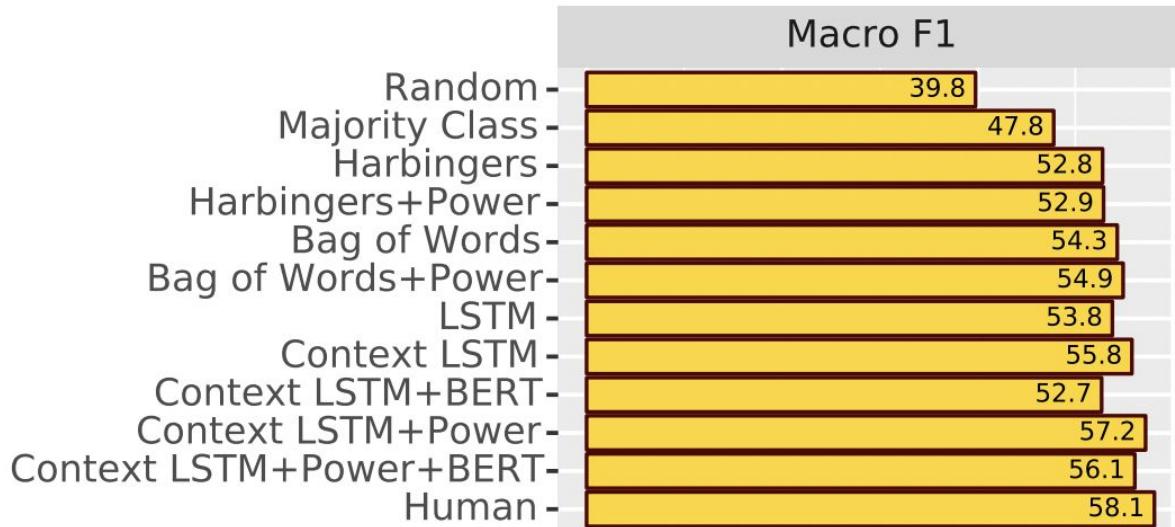
		Receiver's perception	
		Truth	Lie
Sender's intention	Truth	<b>Straightforward</b> Salut! Just checking in, letting you know the embassy is open, and if you decide to move in a direction I might be able to get involved in, we can probably come to a reasonable arrangement on cooperation. Bonne journee!	<b>Cassandra</b> I don't care if we target T first or A first. I'll let you decide. But I want to work as your partner. ... I literally will not message anyone else until you and I have a plan. I want it to be clear to you that you're the ally I want
	Lie	<b>Deceived</b> You, sir, are a terrific ally. This was more than you needed to do, but makes me feel like this is really a long term thing! Thank you.	<b>Caught</b> So, is it worth us having a discussion this turn? I sincerely wanted to work something out with you last turn, but I took silence to be an ominous sign.

**Caught** So, is it worth us having a discussion this turn? I sincerely wanted to work something out with you last turn, but I took silence to be an ominous sign.

Metric: Macro f1 , lying f1

**Sanity checks:**  
Random, majority class

Majority class: Shows dataset imbalance



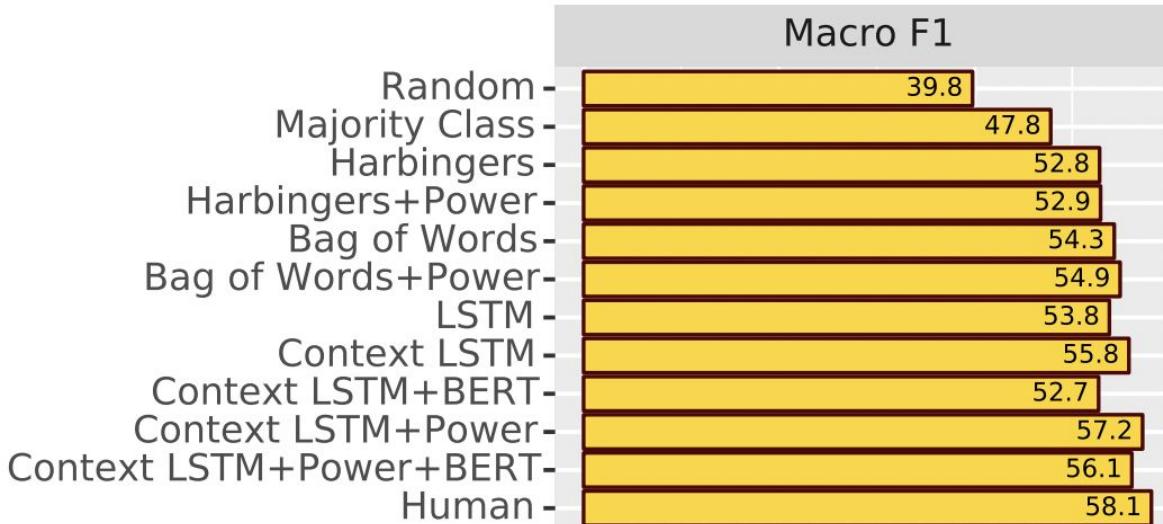
# Bag of words logistic regression

Associated with lies:

- words related to sincerity: sincerely, frankly
- words used in apologies, accusations, fallout, alternatives

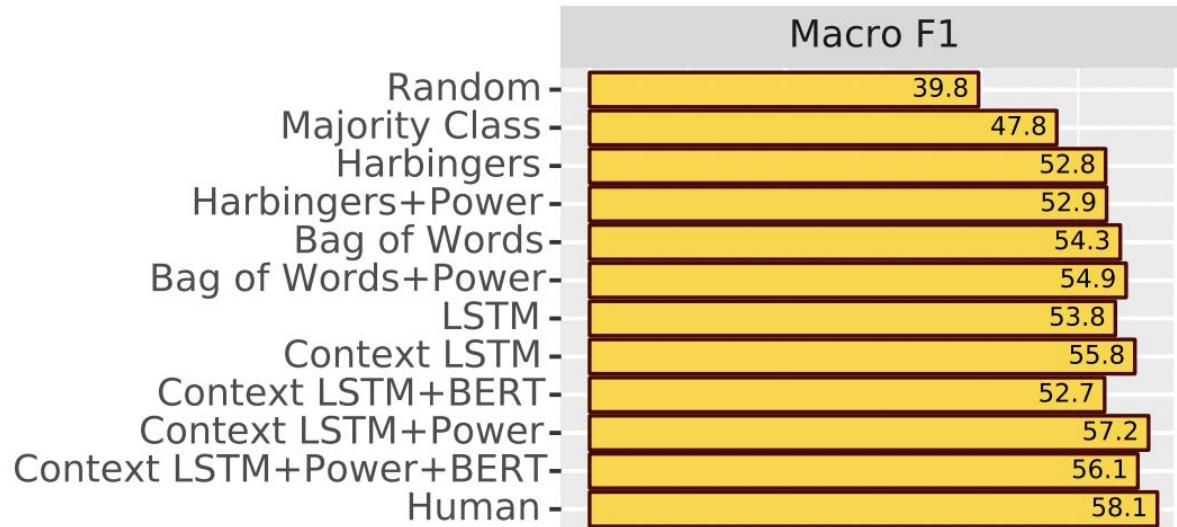
Associated with truthful statements:

- Casual words: dude
- words associated with reconnaissance: FYI
- words associated with time



# Harbingers logistic regression

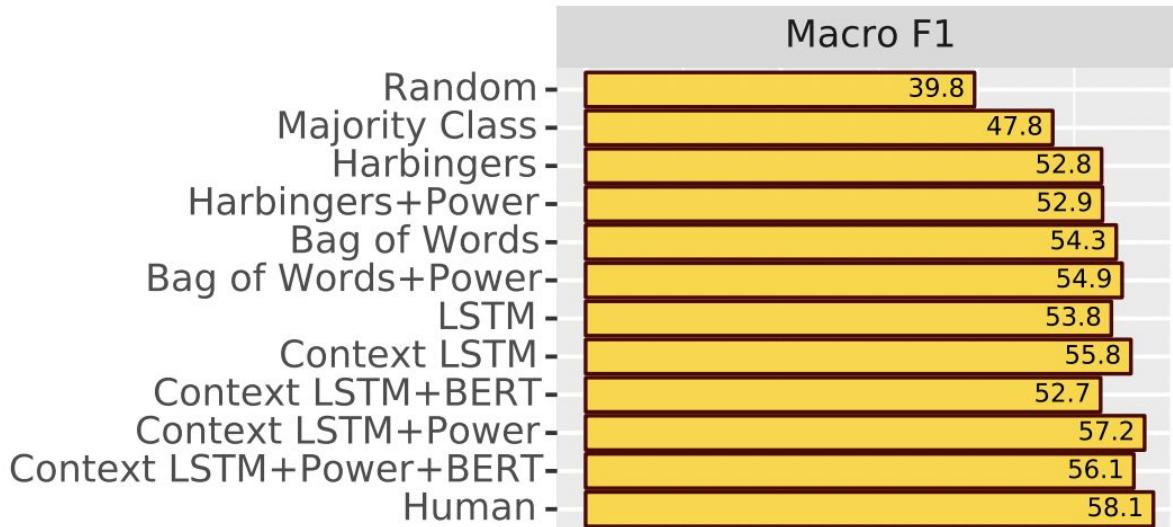
Word lists that cover topics often used in interpersonal communication—claims, subjectivity, premises, contingency, comparisons, expansion



# Power imbalance

Difference between number of supply centers under the control of the two players

Incorporated as a feature in the logistic regression models



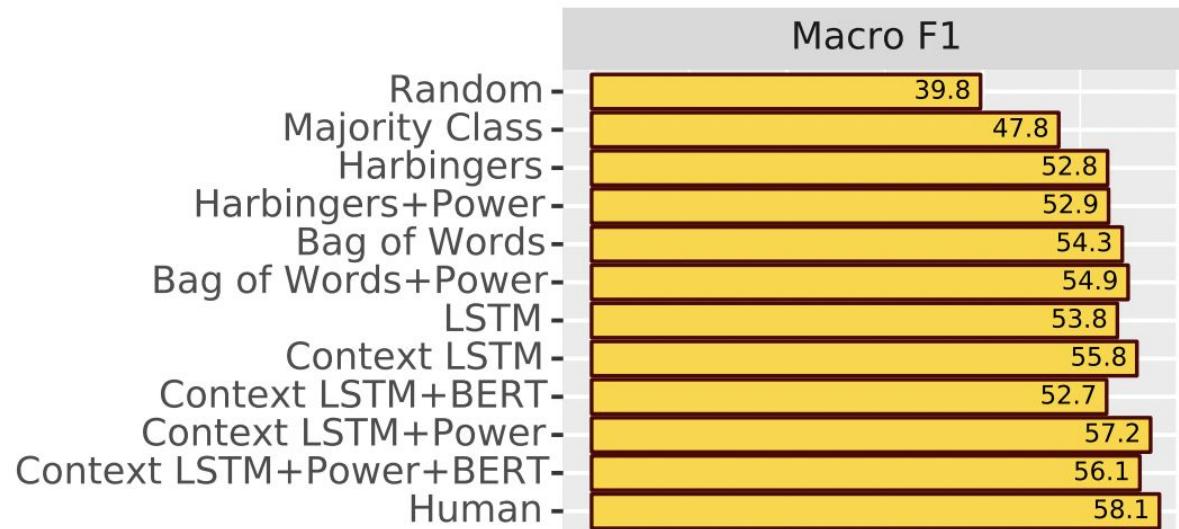
# Neural models

Baseline: LSTM, no context

Extension: Incorporate past context, power (best model)

Fine-tuning BERT embeddings: no improvement

Most gain comes from message itself, not additional information



## Summary: Peskov et al. 2020

Train baseline and neural models to detect deception using this data

Best model approaches human performance

However, both humans and machines failed to detect most lies

# Takeaways

Detecting a lie is difficult for both humans and machines

- Since lies follow an imbalanced class distribution

Press data can be used for building a bot that has a strategic approach

- Human in the loop set up which does better than humans alone

How well would a large language model be able to detect deception?



**Language Models are Few-Shot Learners**

# How to adapt pretrained LLMs for deception detection?

1. Train from scratch
2. Pre-train + fine-tune
3. In-context learning

1

sea otter =&gt; loutre de mer

← example #1



gradient update



1

peppermint =&gt; menthe poivrée

← example #2



gradient update



• • •



1

plush giraffe =&gt; girafe peluche

← example #N

gradient update

1

cheese =&gt; .....

← prompt

# In-context learning

## Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description  
2 cheese => ..... ← prompt
```

---

## One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description  
2 sea otter => loutre de mer ← example  
3 cheese => ..... ← prompt
```

---

## Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description  
2 sea otter => loutre de mer ← examples  
3 peppermint => menthe poivrée  
4 plush girafe => girafe peluche  
5 cheese => ..... ← prompt
```

# In-context learning

## Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

- ```
1 Translate English to French: ← task description  
2 cheese => ..... ← prompt
```

---

## One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

- ```
1 Translate English to French: ← task description  
2 sea otter => loutre de mer ← example  
3 cheese => ..... ← prompt
```

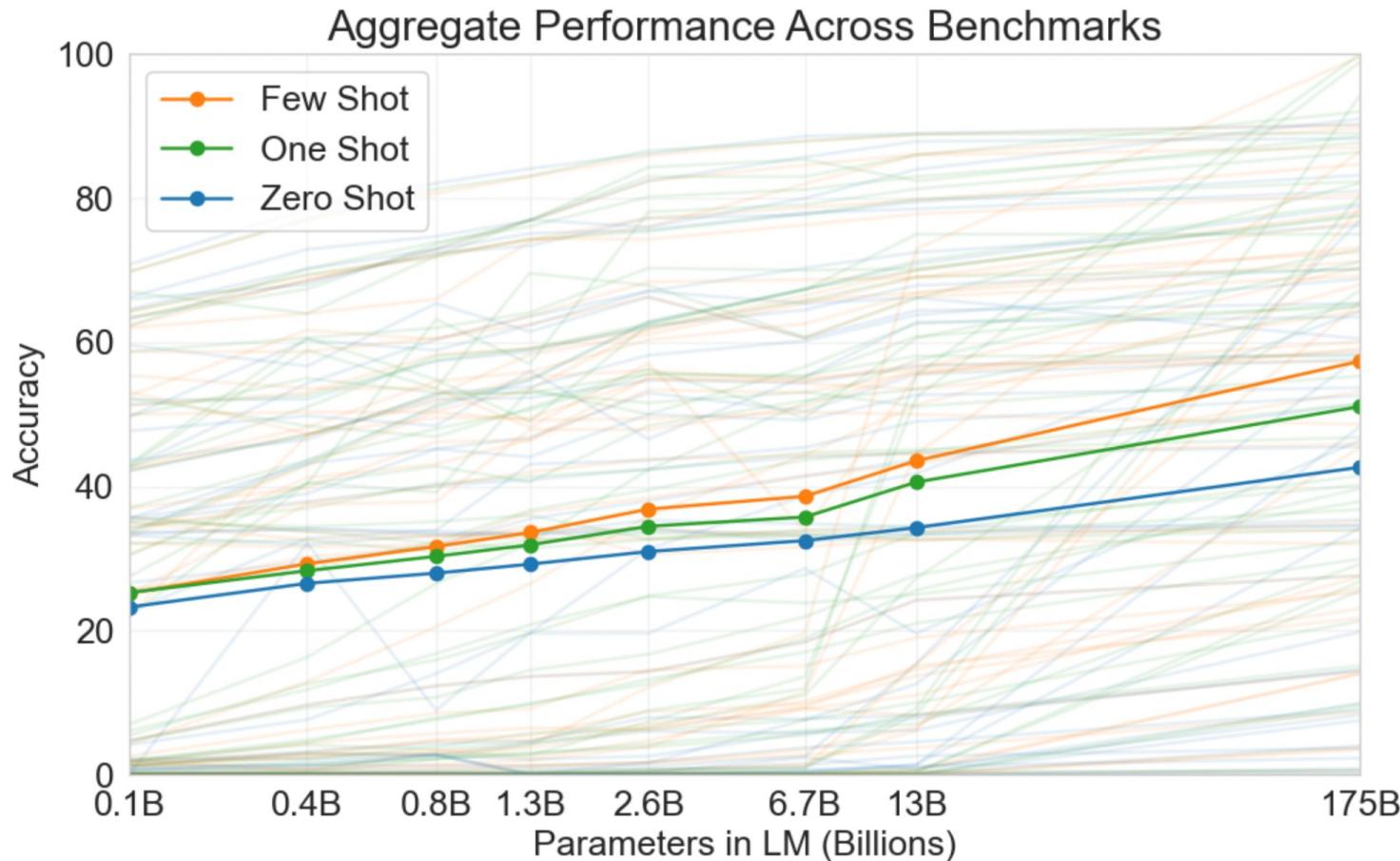
---

## Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

- ```
1 Translate English to French: ← task description  
2 sea otter => loutre de mer ← examples  
3 peppermint => menthe poivrée  
4 plush girafe => girafe peluche  
5 cheese => ..... ← prompt
```

# Few shot learning improves with scale



# Why in-context learning?

Academically interesting

Practically relevant with GPT-3

- Effective with 0-16 examples
- One model for many tasks

**ICML 2021**

---

---

# **Calibrate Before Use: Improving Few-Shot Performance of Language Models**

**Tony Z. Zhao<sup>\*1</sup> Eric Wallace<sup>\*1</sup> Shi Feng<sup>2</sup> Dan Klein<sup>1</sup> Sameer Singh<sup>3</sup>**

# Contextual calibration

Step 1: Estimate the bias

Insert “content-free” test input

Input: Subpar acting. Sentiment: negative  
Input: Beautiful film. Sentiment: positive  
Input: **N/A** Sentiment:

Get model’s prediction

|                 |      |
|-----------------|------|
| <i>positive</i> | 0.65 |
| <i>negative</i> | 0.35 |

Step 2: Counter the bias

“Calibrate” predictions with affine transformation

$$\hat{\mathbf{q}} = \text{softmax}(\mathbf{W}\hat{\mathbf{p}} + \mathbf{b})$$

↑  
Calibrated probs      ↑  
Original probs

Fit  $\mathbf{W}$  and  $\mathbf{b}$  to cause uniform prediction for “N/A”

$$\mathbf{W} = \begin{pmatrix} \frac{1}{0.65} & 0 \\ 0 & \frac{1}{0.35} \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

## Structure of prompt

Italy: ... Germany: ... Italy: ... We are Germany. Was Italy lying to Germany? Yes

Germany: ... Germany: ... Italy: ... We are Germany. Was Italy lying to Germany? No

Italy: ... Germany: ... Italy: ... We are Germany. Was Italy lying to Germany? Yes

Italy: ... Italy: ... Italy: ... We are Germany. Was Italy lying to Germany? No

Italy: ... Germany: ... Italy:... We are Germany. Was Italy lying to Germany?

# Finding good prompts: method

Context window size = 2

In context examples: 4 x examples from train set

Dev set: random sample of 100 examples from original dev set

- Find best threshold (that gives highest macro F1) for difference in probability of “Yes” token and “No” token on dev set (after contextual calibration)

# Finding best prompt: Result on dev set

Best prompt: We are [RECEIVER]. Was [SPEAKER] lying to [RECEIVER]?

Macro F1: 0.603

Lying F1: 0.655

Confusion matrix: {"straightforward": 24, "caught": 37, "deceived": 13, "cassandra": 26}

Notes:

- Replacing specific country names with “us/them” does not cause much improvement
- Variation in F1 scores across prompts reduces after calibration (generally in 0.5-0.6 range)
- Small changes in prompt (“I think… I believe…”) cause a lot of variation in performance

# Finding best prompt: evaluation on full test set

Macro F1: 0.500

Lying F1: 0.274

Best threshold: 0.487

Confusion matrix: straightforward: 1387, caught: 202, deceived: 354, cassandra: 714

# Ensemble approach: part 1

1. Attributes that constitute a lie (based on literature referenced by Peskov et al. 2020): authority, scarcity, likability, reciprocity
2. Tested 2-3 prompts for each of the above attributes
  - a. Ground truth label: sender labels (for lies) and prediction indicates if the model thinks that attribute is displayed in the message.
  - b. Thus perhaps a measure of how well the attributes correlated/correspond to messages that are lies?
3. Chose prompt that gave best macro F1
4. Prompts for each attribute are tested on the same sample of 100 examples from the dev set

# Ensemble approach part 1 results

Best F1s (macro, lying) for each attribute:

1. Authority: 0.521, 0.617
2. Scarcity: 0.499, 0.479
3. Likability: 0.495, 0.545
4. Reciprocity: 0.510, 0.524

# Ensemble approach part 1 results

Best prompt for each attribute:

1. Authority: We are RECEIVER. Is SPEAKER using authority to persuade RECEIVER?
2. Scarcity: We are RECEIVER. Is SPEAKER using scarcity to persuade RECEIVER?
3. Likability: We are RECEIVER. Is SPEAKER displaying likability?
4. Reciprocity: We are RECEIVER. Is SPEAKER reciprocating RECEIVER?

# Ensemble approach part 2: method

1. Choose the best prompts for each attribute
2. Get predictions for each attribute
3. Take the difference in the log probs of the yes and no token after calibration as an entry in a feature vector representing each message
  - a. each entry represents each attribute, so feature vector is of length 4 since we are considering 4 attributes: authority, likability, scarcity, reciprocity
4. Train an MLP (2 hidden layers) with input as 100 examples from dev set, and ground truth labels are the sender labels for those examples
5. Evaluate the trained MLP on the full test set
  - a. after getting the feature vectors using the same method as step 3 for each example in the test set, which are used as the input to the MLP
  - b. Note: no thresholding etc. is done

## Ensemble approach part 2: results on full test set

Macro F1: 0.430

Lying F1: 0.270

Train separate models for each player  
+ comparison with BERT+ context LSTM model from  
Peskov et al. 2020

Method (GPT-3)

- In context examples contain 4 examples with latest message in each example sent by the winner of the game
- Dev set: only examples where latest message is sent by winner of the game
- Test set: only examples where latest message is sent by winner of the game

Method (BERT+context LSTM)

- Train on full test set, dev set is the same as the original dev set
- Test set: only examples where latest message is sent by winner of the game

# BERT + context LSTM

```
1 {'lying_f1': 0.23529411764705882, 'macro_f1': 0.5286059629331185}
2 {'lying_f1': 0.0, 'macro_f1': 0.45454545454545453}
3 {'lying_f1': 0.0, 'macro_f1': 0.4576271186440678}
5 {'lying_f1': 0.0, 'macro_f1': 0.4476190476190476}
6 {'lying_f1': 0.0, 'macro_f1': 0.4285714285714286}
7 {'lying_f1': 'n/a', 'macro_f1': 1.0}
8 {'lying_f1': 0.0, 'macro_f1': 0.49056603773584906}
9 {'lying_f1': 0.0, 'macro_f1': 0.44537815126050423}
10 {'lying_f1': 0.0, 'macro_f1': 0.45161290322580644}
11 {'lying_f1': 0.0, 'macro_f1': 0.45378151260504196}
4 {'lying_f1': 0.1333333333333333, 'macro_f1': 0.49965635738831615}
12 {'lying_f1': 0.0, 'macro_f1': 0.42748091603053434}
```

# GPT-3

```
1 f1_macro: 0.4580619017676239 ; f1_lying: 0.2125340599455041
2 f1_macro: 0.4641010913268236 ; f1_lying: 0.0
3 f1_macro: 0.4722662440570523 ; f1_lying: 0.0
5 f1_macro: 0.424390243902439 ; f1_lying: 0.0
6 f1_macro: 0.488031914893617 ; f1_lying: 0.125
7 f1_macro: 0.4813863928112965 ; f1_lying: 0.15789473684210528
8 f1_macro: 0.5229000691185565 ; f1_lying: 0.12244897959183673
9 f1_macro: 0.47006369426751593 ; f1_lying: 0.0
10 f1_macro: 0.46282973621103124 ; f1_lying: 0.0
11 f1_macro: 0.5408208955223881 ; f1_lying: 0.16
4 f1_macro: 0.5055700862152475 ; f1_lying: 0.2724014336917563
12 f1_macro: 0.5203208556149733 ; f1_lying: 0.18181818181818182
```

ArXiv, January 2022

# Chain of Thought Prompting Elicits Reasoning in Large Language Models

---

**Jason Wei    Xuezhi Wang    Dale Schuurmans    Maarten Bosma**  
**Brian Ichter    Fei Xia    Ed H. Chi    Quoc V. Le    Denny Zhou**

Google Research, Brain Team  
`{jasonwei,dennyzhou}@google.com`

## Standard Prompting

### Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

### Model Output

A: The answer is 27. 

## Chain of Thought Prompting

### Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls.  $5 + 6 = 11$ . The answer is 11.

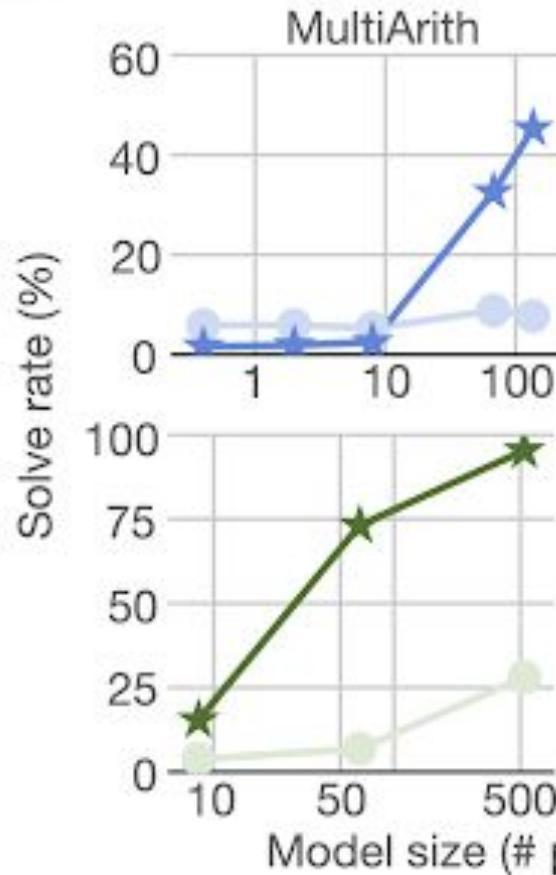
Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

### Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had  $23 - 20 = 3$ . They bought 6 more apples, so they have  $3 + 6 = 9$ . The answer is 9. 

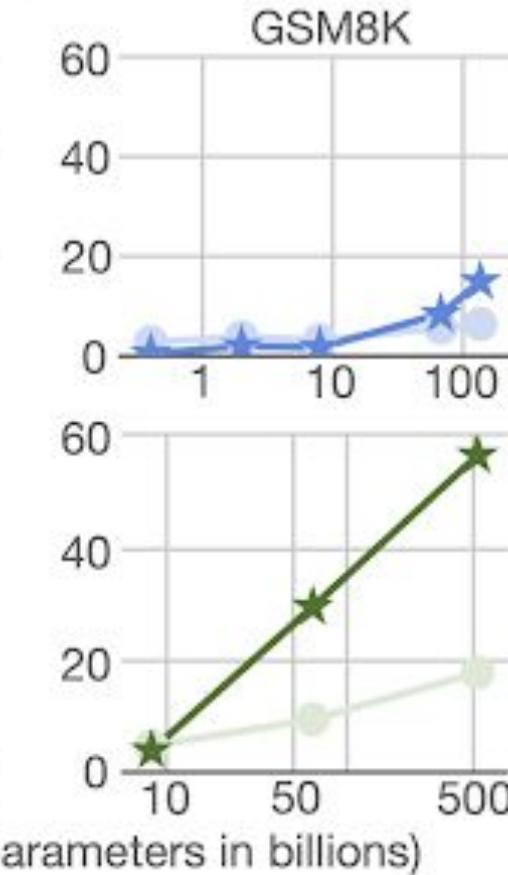
**LaMDA:**

- standard prompting
- chain of thought prompting



**PaLM:**

- standard prompting
- chain of thought prompting



# Large Language Models are Zero-Shot Reasoners

---

**Takeshi Kojima**

The University of Tokyo

t.kojima@weblab.t.u-tokyo.ac.jp

**Shixiang Shane Gu**

Google Research, Brain Team

**Machel Reid**

Google Research\*

**Yutaka Matsuo**

The University of Tokyo

**Yusuke Iwasawa**

The University of Tokyo

### (a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

*(Output) The answer is 8. X*

### (c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: The answer (arabic numerals) is

*(Output) 8 X*

### (b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls.  $5 + 6 = 11$ . The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

*(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are  $16 / 2 = 8$  golf balls. Half of the golf balls are blue. So there are  $8 / 2 = 4$  blue golf balls. The answer is 4. ✓*

### (d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: **Let's think step by step.**

*(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓*

**【1st prompt】  
Reasoning Extraction**

Q: On average Joe throws 25 punches per minute. A fight lasts 5 rounds of 3 minutes. How many punches did he throw?

**A: Let's think step by step.**



LLM



In one minute, Joe throws 25 punches.

In three minutes, Joe throws  $3 * 25 = 75$  punches.

In five rounds, Joe throws  $5 * 75 = 375$  punches.

**【2nd prompt】  
Answer Extraction**

Q: On average Joe throws 25 punches per minute. A fight lasts 5 rounds of 3 ...  
A: Let's think step by step.

In one minute, Joe throws 25 punches. ... In five rounds, Joe throws  $5 * 75 = 375$  punches. .

**Therefore, the answer (arabic numerals) is**



LLM



375.

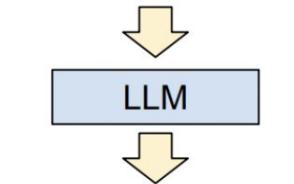
# 2 stage prompting for consistency check

1st prompt:

(Italy: ... Germany: ... Italy: ...) x4

Italy: ... Austria: ... Italy: ...

We are Austria. Is Italy making  
consistent statements?



No

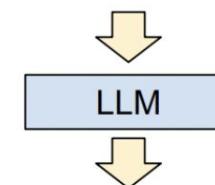
2nd prompt:

(Italy: ... Germany: ... Italy: ...) x4

Italy: ... Austria: ... Italy: ...

We are Austria. Is Italy making consistent  
statements? No

Is Italy lying to Austria?



Yes

## Other ideas

Incorporate power: “The game score delta between [SPEAKER] and [RECEIVER] is [x].”

Generate annotations for likeability, assertiveness etc.

Try to predict opponents actions k steps into the future (say k = 5)

# Large Language Models for Few Shot DAIDE Translations (Sander Schulhoff, Sept 2022)

DAIDE: language with three letter acronyms for Diplomacy moves (e.g. HLD: hold)

Experiment with GPT-3's ability to translate between DAIDE and English

Result: impressive 0 shot and few shot performance

Applications suggested:

1. English message annotation: evaluate stances of players, classify positive or negative relations between players
2. Generate different tones of speech in English from DAIDE
3. Incorporate with GameScrapes data somehow?

Here are some words from a language called DAIDE, as well as their english translations:

HLD: hold

MTO: move to

SUP: support to hold

units are defined as (country unit\_type province)

a support to move order looks like: (unit) SUP (unit) MTO province

Here are some translation examples from English to DAIDE:

[Russia to England] Can your army in Warsaw support my army in Ukraine?: PRP ((ENG AMY WAR) SUP (RUS AMY UKR))

[Germany to Austria] Can your fleet on the Baltic Sea support my army in Sweden?: PRP ((AUS FLT BAL) SUP (GER AMY SWE))

[France to Italy] Can your fleet in the Adriatic Sea convoy my army in Apulia to Trieste?:  
PRP ((ITA FLT ADR) CVY (FRA AMY APU) CTO TRI)

Now translate the following from DAIDE to English:

## Short term vision

Develop the best pipeline for leveraging GPT-3's in-context learning ability to detect deception

## Medium term vision

Gain confidence that GPT-3 is able to detect linguistic signals for deception

After gaining confidence, gain motivation to investigate the use of GPT-3's in context learning ability in generating annotations for stance, etc.

# Long term vision

Evaluate the extent/ability of large language models to detect nuanced aspects of language such as deception which consists of more complicated long-range dependencies

Any ideas/suggestions?

Thank you!

# References

[Language Models are Few-Shot Learners](#)

[Calibrate Before Use: Improving Few-Shot Performance of Language Models](#)

[Chain of Thought Prompting Elicits Reasoning in Large Language Models](#)

[Large Language Models are Zero-Shot Reasoners](#)

[It Takes Two to Lie: One to Lie, and One to Listen](#)

[Learning to Play No-Press Diplomacy with Best Response Policy Iteration](#)

[No Press Diplomacy: Modeling Multi-Agent Gameplay](#)

[HUMAN-LEVEL PERFORMANCE IN NO-PRESS DIPLOMACY VIA EQUILIBRIUM SEARCH](#)

Large Language Models for Few Shot DAIDE Translations (Sander Schulhoff, Sept 2022)

# References

[Diplomacy - Games4Sustainability](#)

[Diplomacy: The Map That Ruined a Thousand Friendships - Bloomberg](#)

[Diplomacy Is A Game That Makes Careers And Ruins Friendships](#)

[AI is learning to play Diplomacy, and it's pretty good at it | Enterprise](#)

[The Simpsons - It Takes Two To Lie](#)

[Backstabbr](#)

[It Takes Two to Lie: One to Lie and One to Listen](#)