Explanations and robustness in Multimodal Language Models

Pranava Madhyastha

pranava.madhyastha@city.ac.uk

CITY UNIVERSITY
LONDON

Agenda

Premise

Explaining and interpreting in Vision and Language

Multimodal language models

Unified model for Answers and Explanations

Human comprehension

Towards a Unified Model for Generating Answers and Explanations in Visual Question Answering

Chenxi Whitehouse | Tillman Weyde | PM

City, University of London

Computer Science

Premise

QA: Vision and Language domain

Given a question presented in natural language and visual information, the machine learning system has to accurately predict an answer to the question.

Eg.: What colour is the woman’s jacket?

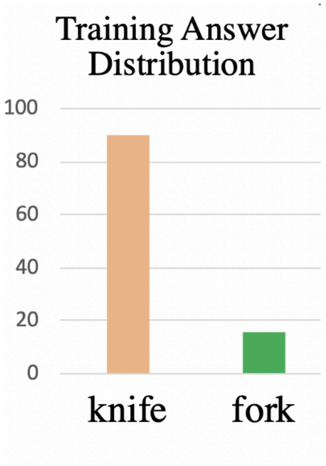
Antol, Stanislaw, et al. "Vqa: Visual question answering." Proceedings of the IEEE international conference on computer vision. 2015.

+ Explanation

Question: What utensil is pictured?

A: ?

Explanations: There is a fork on the table.



+ Knowledge grounding

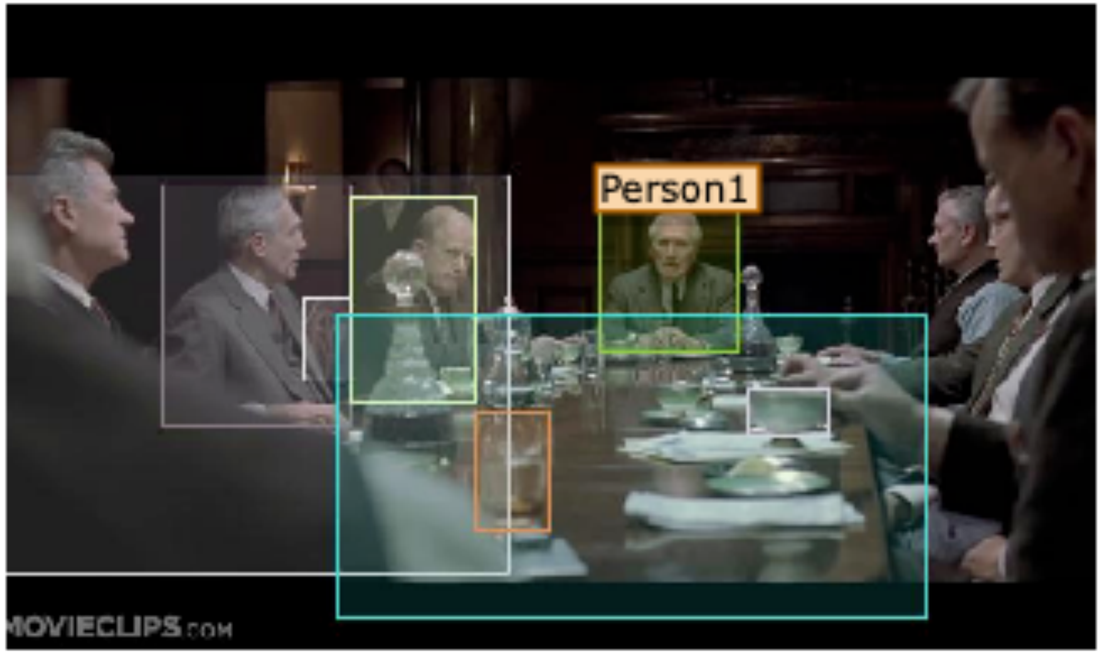
Q: Is this in an Asian country?

Explanations: Japanese words on the train and Japan is an Asian country.

+Commonsense reasoning

Q: What is Person 1 going to do?

A: Person 1 is going to lead a business meeting

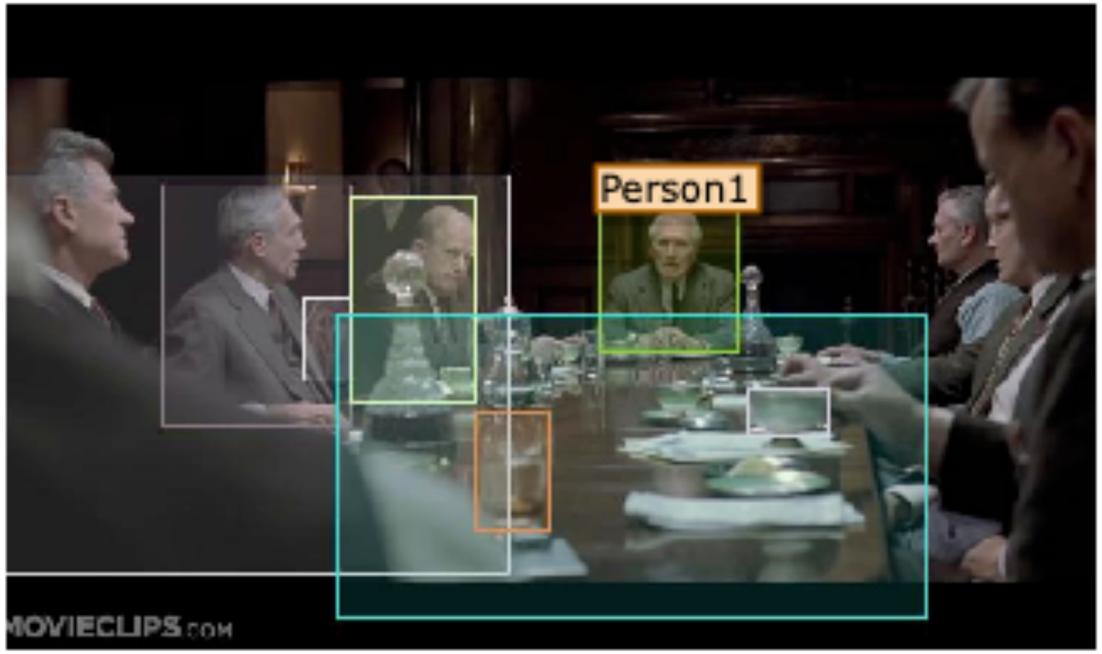


+Explanation

Q: What is Person 1 going to do?

A: Person 1 is going to lead a business meeting

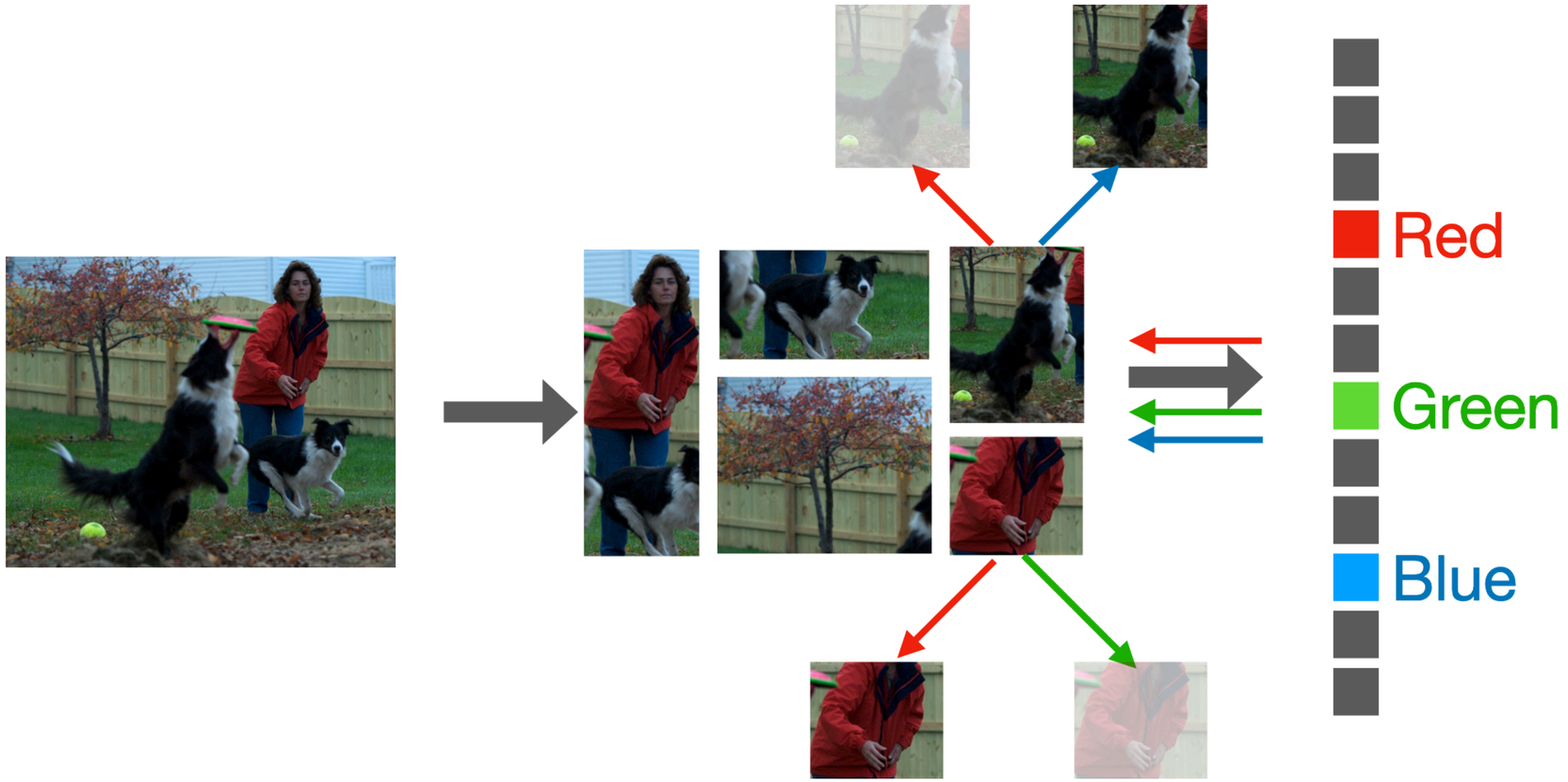
E: Person1 is at the head of a table of men in suits.



X in Vision and Language

Person1 is at the head of a table of men in suits.

Visual justifications



Ramprasaath R Selvaraju, et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In ICCV 2017

Textual justifications

Q: Is this a healthy meal?

A: No

...because it is a hot dog with a lot of toppings.

A: Yes

...because it contains a variety of vegetables on the table.

Dong Huk Park, et al.. Multimodal Explanations: Justifying Decisions and Pointing to the Evidence. In *CVPR* 2018.

Complexity

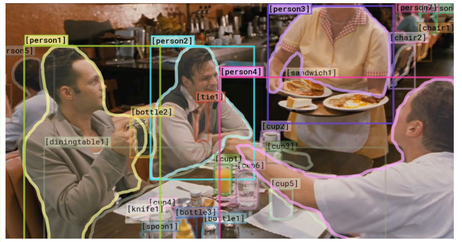
As images get complex where multiple concepts intermingle, labelling and explaining the labels becomes challenging.

Subjective experience usually is associated with explanations - there are multiple ways to arrive at an answer

Single Task
Reasoning process is not explicitly evaluated.

Does this person have 20/20 vision?

Parallel Task
Reasoning is evaluated as a separate task with no guarantee for helping the main task.



Why is [person4] pointing at [person1]?

Why is this answer right?

What should I do, according to this advertisement? [action]

Why, according to this ad should I take this action? [reason]

Ye, Keren, Mingda Zhang, and Adriana Kovashka. "Breaking shortcuts by masking for robust visual reasoning." WCACV 2021

Evaluations

Automated evaluation methods are all fallible - fail in almost deterministic ways.

Current methods: repurposing of NLG metrics

➡Reference evaluations sampled from humans

➡Procedure: Intersection of vocabulary

➡Challenges with human biases in reasoning

Challenges with labelling

Question: Is this legal or illegal?

Ground Truth Answers:
legal (6), illegal (4)

Generation: legal

Question: In which country are the transportation regulations loose enough to allow vehicles like these?

Ground Truth Answers:
india (8), china (2)

Generation: england

Question: How long does it take to cook?

Ground Truth Answers:
45 minutes (4), 20 minutes (2), 25 minutes (2), minute (2)

Generation: 1 hour

Question: What nationality is this food?

Ground Truth Answers:
american (4), mediteranian (2), greek (2), asian (2)

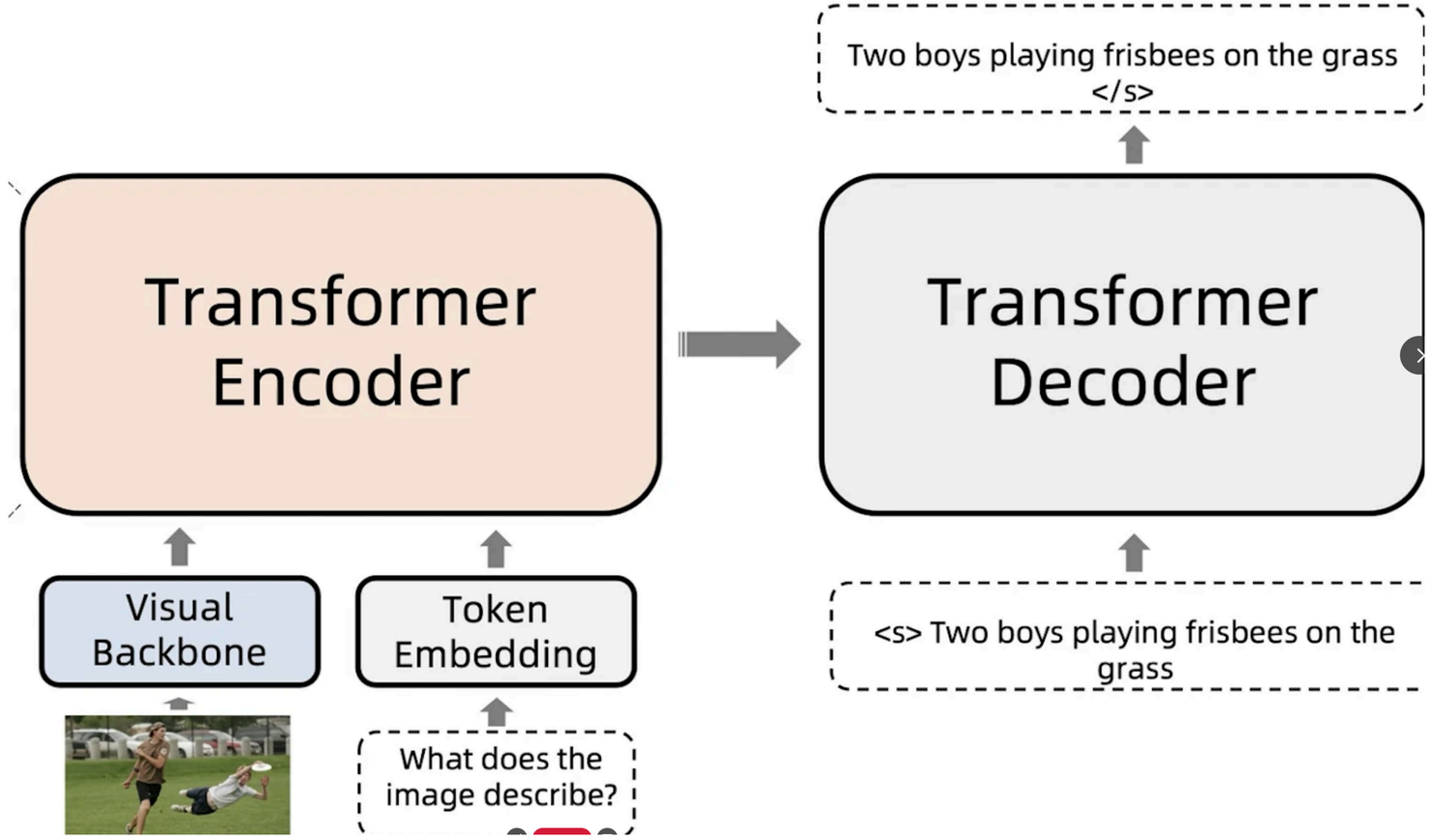
Generation: italian

Multimodal LMs

Transformer backend!

Visual information as tokens!

Transformer backbone



What happens with the images?

Image is split into sequence of patches

Which that then embedded using a pretrained vision model

Goal - consider every input as tokens

Pre-trained over a variety of data sources

Models are usually pretrained over a variety of tasks and tokens

- Language related data separately

- Language and Vision related data

- Vision tasks reformulated as with a language prompt and language like outputs

Unified model for Answers and Explanations

Premise

Agnostic explanation methods are usually not grounded in the task descriptions

Separate models usually for explaining the behaviours of the models tend to be generate usually disparate explanations

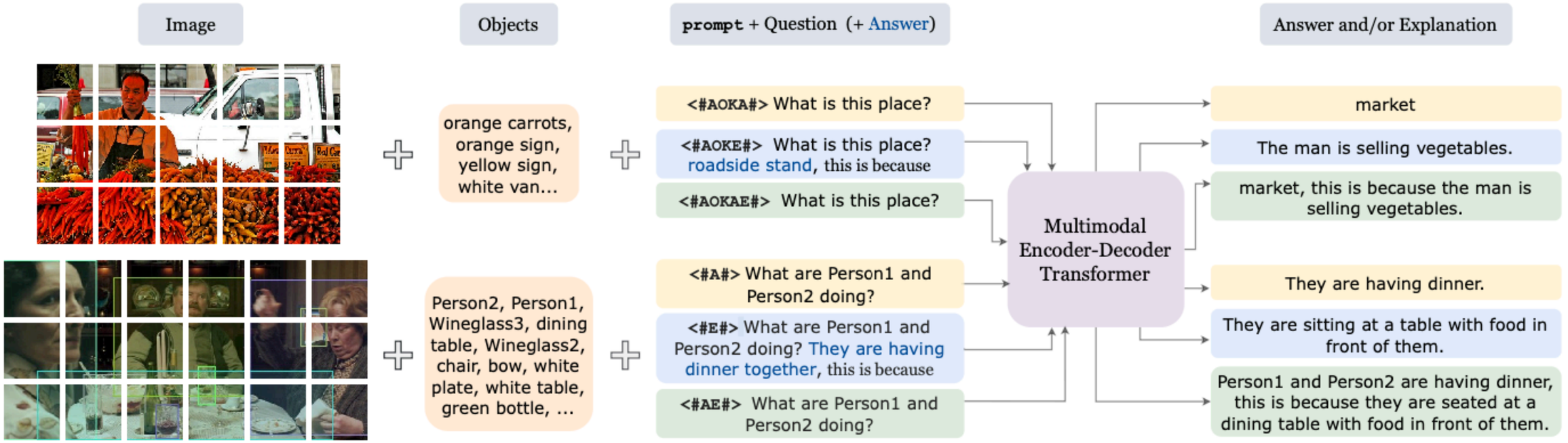
Synthetic prompt tokens

3 settings settings for any given image:

- Q→A: Answer prediction
- QA→E: Explanation generation conditioned on the answer
- Q→AE: Joint answer and explanation generation for a given question

This setup allows the model to enhance the signal of “answers” associated with

Illustration



Synthetic prompt tokens

Free form commands with “why is X the answer” do not seem to generalise with answers across domains due to the inherent ambiguities

Synthetic symbol with a uniform semantics usually allows for consistent outputs

Complementarity of explanations

Model	OK-VQA	A-OKVQA					VCR	
	direct answer	*multiple choice*			*direct answer*		*multiple choice*	BERTSCORE
	TEST	VAL (*ppl*)	VAL (*GloVe*)	TEST	VAL	TEST	VAL (*ppl*)	VAL
OFA*	40.40	24.54	56.19	47.40	48.09	39.77	33.55	64.55
$OFA_{Q->A}$	49.93	74.32	65.30	61.71	63.00	53.91	54.89	83.85
$UMAE_{ALL}$	**51.77**	**74.59**	**65.67**	**63.26**	**63.29**	**56.14**	**56.66**	**85.97**
PRIOR-BEST	54.41	–	60.30	53.70	48.60	40.70	$(77.10)^{\dagger}$	–

Key takeaway: Explanations help with prediction of consistent and robust labels

Unified model allows for better explanations

Dataset	Model	N-gram Scores					Learnt Score
		BLEU4	ROUGE-L	METEOR	CIDEr	SPICE	BERTSCORE
A-OKVQA	OFA*	0.30	4.45	3.26	4.82	4.62	68.64
	$OFA_{Q->A}$+$OFA_{QA->E}$	22.18	48.51	23.56	86.76	22.46	85.96
	$UMAE_{A-OKVQA}$	**27.61**	52.23	24.06	**104.39**	22.88	87.86
	$UMAE_{ALL}$	27.35	**52.56**	**24.83**	101.09	**23.33**	**88.21**
VCR	e-UG	4.30	22.50	11.80	32.70	12.60	79.00
	$UMAE_{VCR}$	12.25	28.87	16.67	**48.14**	27.36	81.77
	$UMAE_{ALL}$	**13.44**	**29.53**	**17.54**	47.33	**26.45**	**81.91**
VQA-X	e-UG	23.20	45.70	22.10	74.10	20.10	87.00
	$UMAE_{ALL}$	14.63	35.12	20.29	50.35	19.13	85.40

Key takeaway: dataset centric particularities are usually important as annotations are conducted under varied settings.

Question: What time of year was the picture likely taken?
Answer: fall

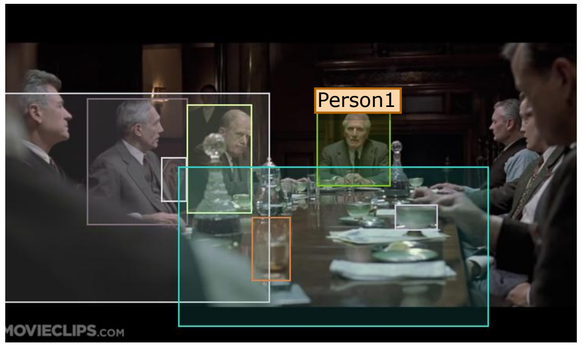
Ground Truth Explanations:
1) The child is wearing a long sleeve shirt and pants but no coat.
2) There are brown leaves on the sidewalk.
3) The time is fall.

Generated Explanations:
Beam Search: The time is fall.
Top-k: The leaves are dropping.
Nucleus: The leaves are fall.
Typical: The leaves are brown and dry.

Question: Which two words were said by both the person in black and the person in white here?
Answer: i do

Ground Truth Explanations:
1) The people got married.
2) There is a wedding cake. the smiling people in the suit and white dress are the bride and groom.
3) The photo was obviously taken at a wedding with the bride and groom at the center of it. it is traditional that they say "i do" when taking their vows.

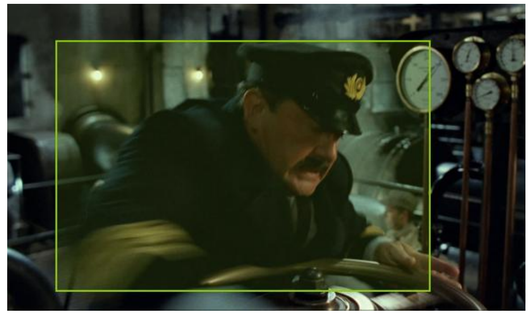
Generated Explanations:
Beam Search: They are at a wedding.
Top-k: The words are ""i"".".
Nucleus: The person in black and the person in white are in a wedding.
Typical: The people are just after getting married.



Question: What is Person1 going to do?
Answer: Person1 is going to lead a business meeting.

Ground Truth Explanation:
Person1 is at the head of a table of men in suits.

Generated Explanations:
Beam Search: Person1 is sitting at the head of the table, which is where leaders sit.
Top-k: Person1 appears to be in charge of the board room.
Nucleus: Person1 is sitting at the head of the table, which is where leaders sit during business meetings.
Typical: Person1 is sitting at the head of the table, which is often the place for leaders.



Question: What is Person1 doing?
Answer: Person1 is turning the wheel.

Ground Truth Explanation:
Person1 has his hands on the wheel and is exerting force.

Generated Explanations:
Beam Search: He has his hands on the wheel.
Top-k: Person1 is hunched over and has his hands on the wheel.
Nucleus: Person1 is leaning over the wheel and has his hand on it.
Typical: The man is leaning forward and his hand is on the wheel.

Conclusions

Training with explanations helps robustly task related predictions

Synthetic tokens with well defined task related semantics enrich multi-task capabilities

A unified model capable of prediction and explaining the predictions is better grounded in the task and dataset related intricacies.

Hiring ...

Robust inference with probabilistic answer set program scaffolds for large language models

2 post docs

Probabilistic programming

ASP/ILP/LP

Hosted at the Alan Turing Institute

Communicating explanations

Important role of human comprehension

Human comprehension is most effective when it is multimodal

We learn with multimodal signals

Changing human beliefs with explanations is difficult

Social AI and the Challenges of Human-AI Ecosystem, Pedreschi et al 2022

Are words equally surprising in audio and audio-visual comprehension?

PM | Claudia (Ye) Zhang | Gabriella Vigliocco

University College London

Psychology and Language Sciences

Expectation based theories of sentence comprehension

Sentence processing difficulty is influenced by the predictability of upcoming lexical material in context (Levy, 2008).

Previous research has typically examined the impact of the following types of context: extra-sentential information (e.g., discourse), previous sets of lexical items, and the current lexical item.

Humans rely on their accumulated linguistic knowledge, including complex grammatical structures and contextual understanding, to process sentences incrementally.

Surprisal theory

Quantifies the unexpectedness of linguistic events

An information theoretic measure

Predicts processing difficulty as a function of word probability (Hale, 2001)

Cognitive effort (linguistic unit) $\propto$ Surprisal (linguistic unit)

Operationalisation of surprisal

Surprisal is typically calculated based on the likelihood of encountering a linguistic unit under the preceding context.

This has been done previously using:

- corpus-based frequencies

- using close tasks

- information theoretic estimates (such as entropy)

- **language models**

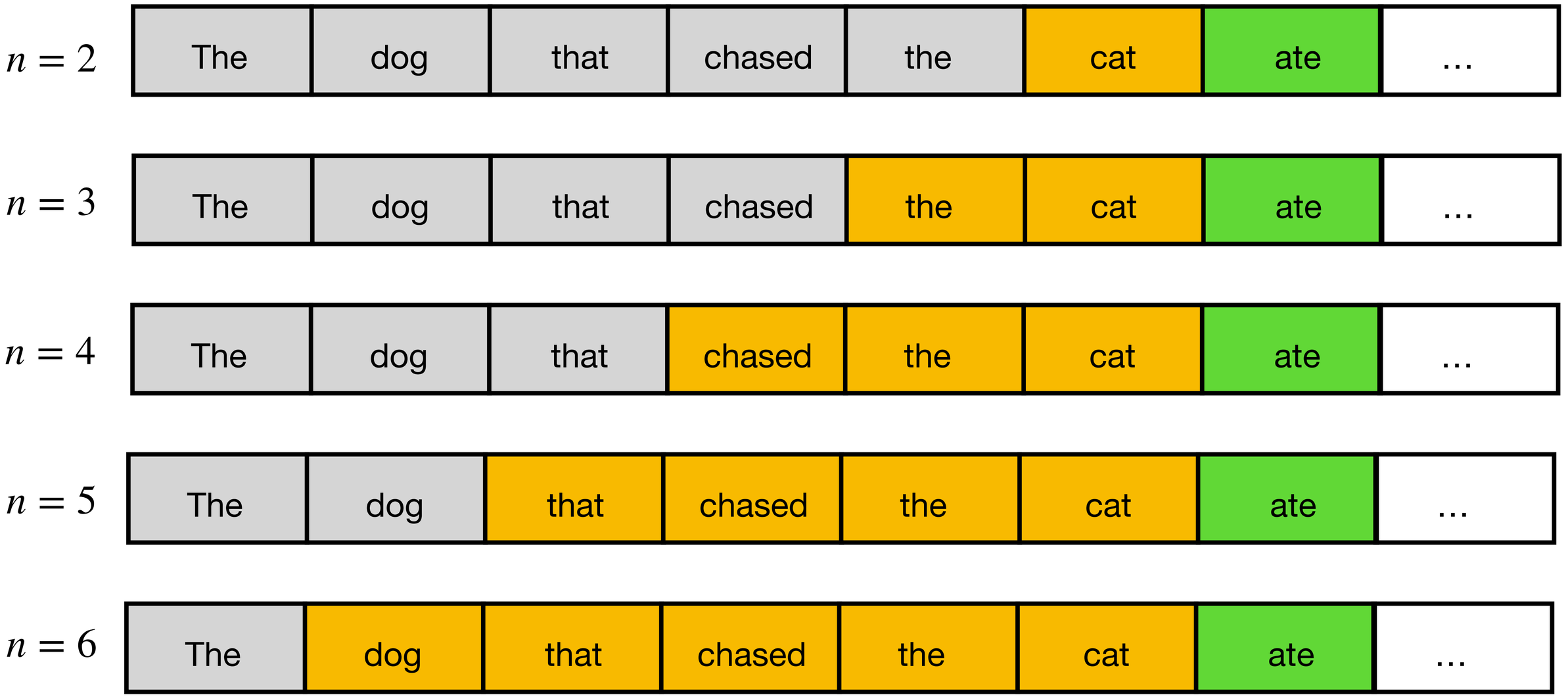
Operationalisation of surprisal

Language models are highly reliable for measuring surprisal, and have been found to strongly correlate with both behavioral and EEG-based measures of cognitive effort (Michaelov & Bergen, 2020; Meister et al., 2021).

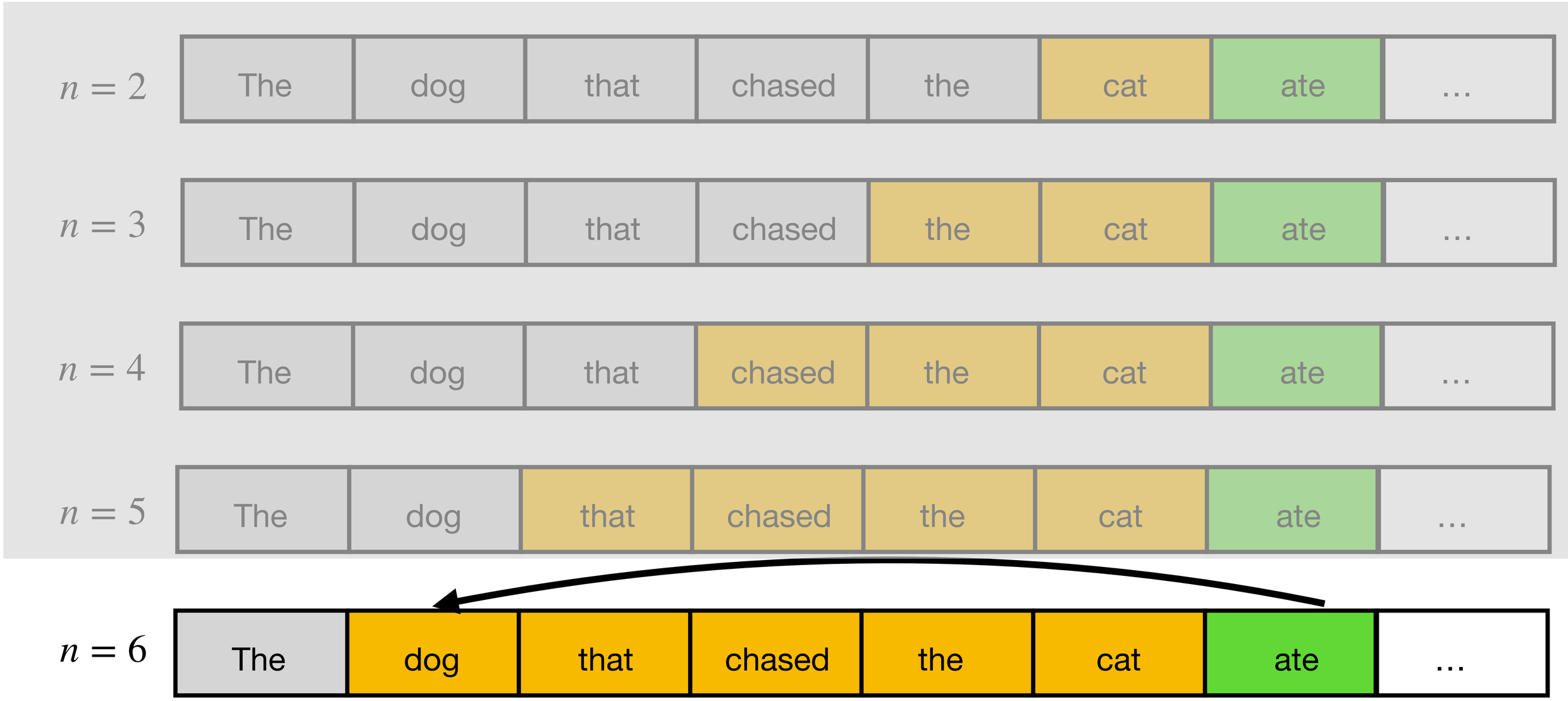
$$\text{surprisal}(\text{lexical unit}) = -\log p_\theta(\text{lexical unit} \mid \text{lexical context})$$

Estimated using language models

n-gram based models

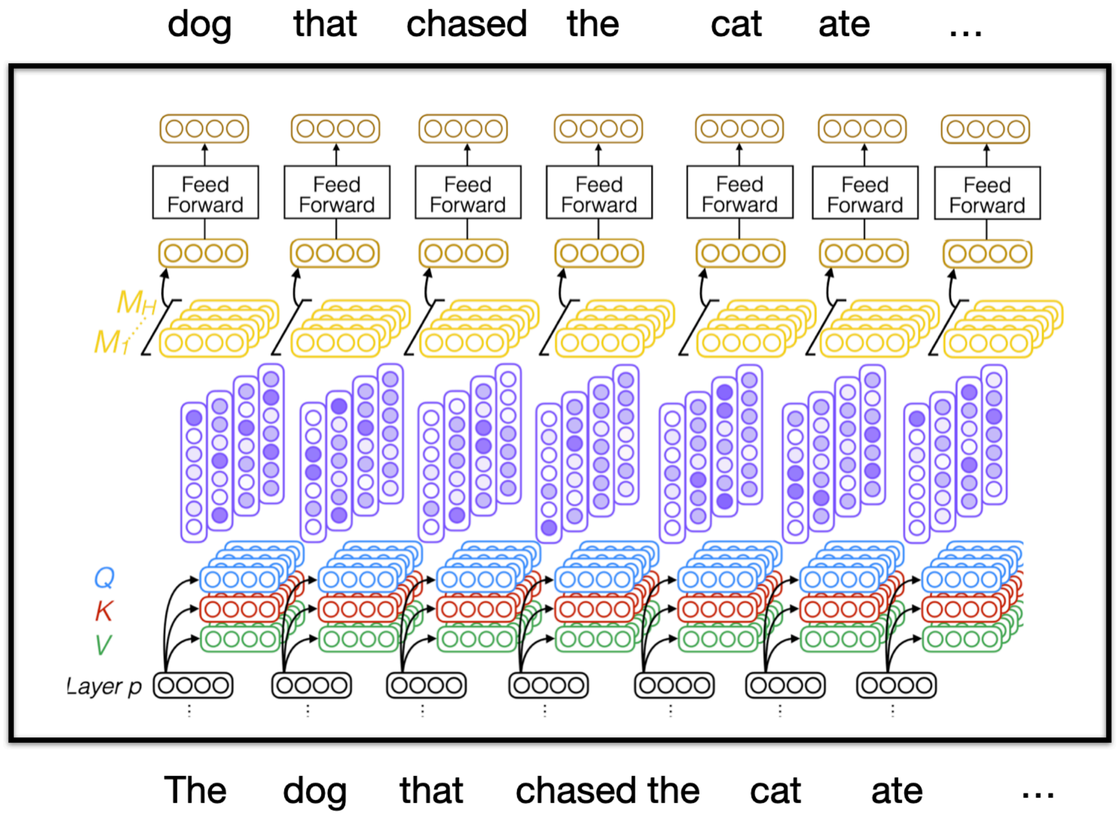


Long ranged dependencies



Transformers based models

- Capture higher order correlations (Sinha et al, 2021)
- Access to infinite context (for our datasets)
- Capable of capturing very-long dependencies
- We consider two instances of these models:
  - ✦ GPT2: a generative model that predicts the next token for a given context.
  - ✦ BERT: a predictive model that is trained in a cloze style word prediction setup.



Empirical validation of surprisal theory

Surprisal estimates from language models have been shown as a good predictor of language effort during language processing. These include:

- strong correlation with reading times (Smith and Levy 2013)

- word fixations were longer when words had high surprisal values (Demberg and Keller 2008)

- significant associated with reading times in eye-tracking data (Futrell et al. 2019)

- **highly correlated with neurophysiological signals**

Neurophysiological signals and surprisal

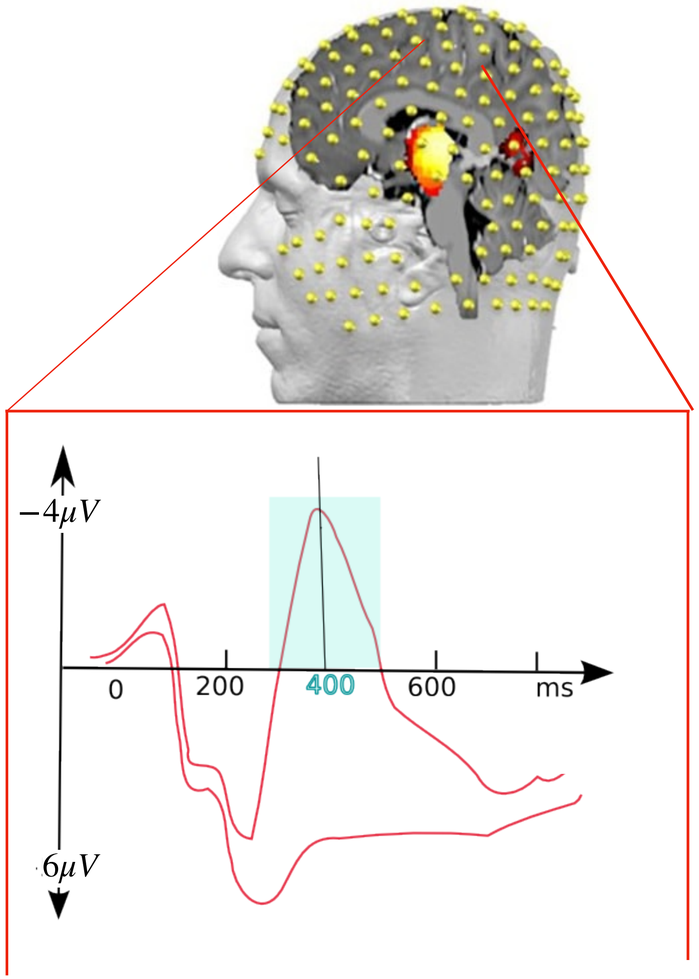
ERP studies typically examine the brain's response to words or linguistic stimuli with different levels of predictability or unexpectedness.

ERP peaking negatively at ≈400ms at the central parietal areas during language processing tasks.

Several works have shown the demonstrated the relationship between N400 and semantic processing:

- Frank et al. 2015: Surprisal (as obtained from LM) predicts n400.

- Michaelov et al. 2021: Surprisal estimates from larger models are better predictors of N400.



Audiovisual language

Language is embedded within a rich multimodal environment that includes gestures, facial expressions, body movements, visual cues, and other nonverbal elements

Multimodal information provides additional context and meaning, making communication more effective (Ankener et al., 2018; Grzyb et al., 2022; Zhang et al., 2021)

Multimodal information, such as pitch prosody, meaningful gestures and informative mouth movements, modulates the N400 signal especially for high surprisal words (Zhang et al., 2021 and Baumann & Schumacher, 2012)

Audiovisual communication

Facial expressions

Head movements

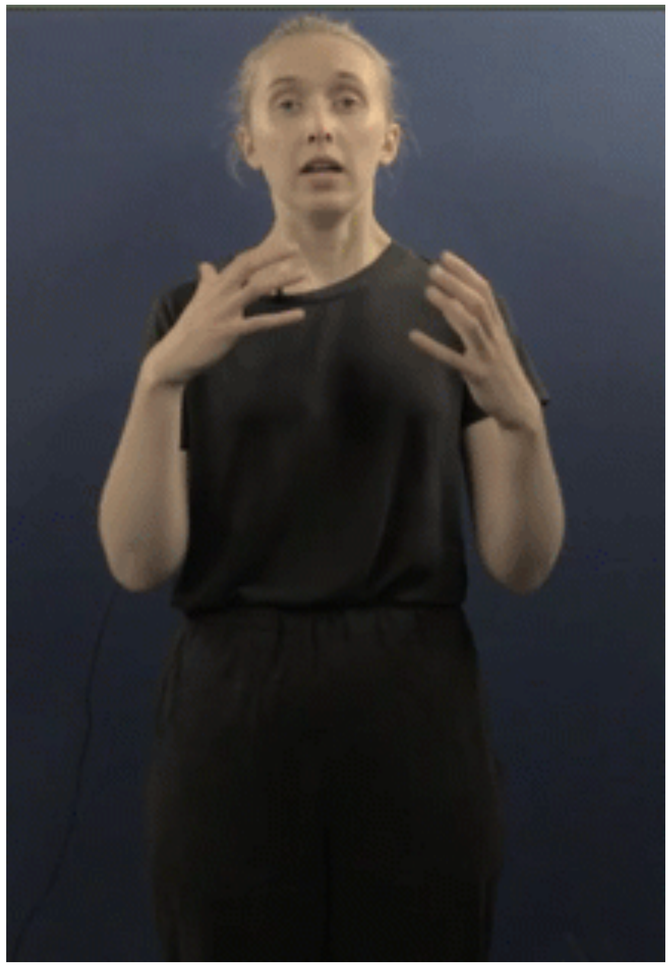
Hand gestures

Iconic gestures

Affective gestures

Illustrative gestures

. . .



Audio only vs Audiovisual communication

Previous research has mostly focussed on characterising comprehension difficulty through experiments based on lexical information alone.

Information theoretic frameworks have typically focused on information content propagated through a single channel.

However, most natural modes of communication involves the contribution from **multiple modalities**

This work

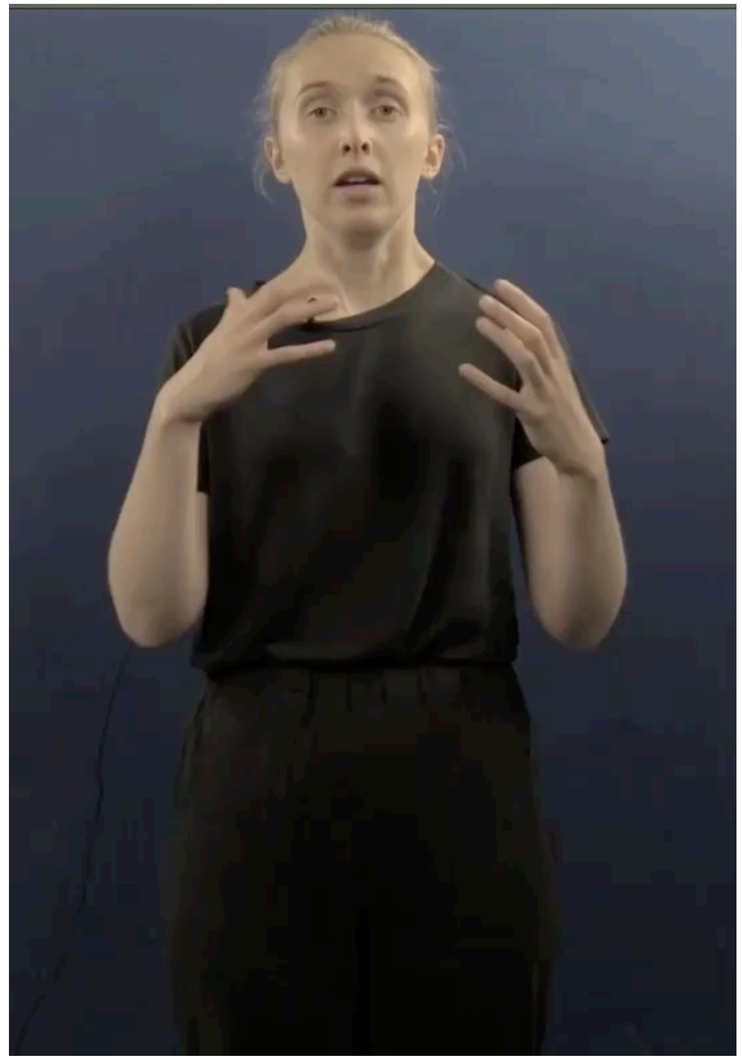
We present a controlled study to investigate the effects of visual signals (seeing the speaker) on language comprehension

We compare the effects of **audio-only and audio-visual** settings using the **same language stimuli** and analyse the changes in ERP signals

We then evaluate the effectiveness of surprisal estimates, using different language models with varying lexical context windows, in explaining cognitive effort in both unimodal and multimodal conditions

Stimuli

- 103 naturalistic passages carefully sampled passages from BNC
- Recorded by a native English-speaking actress
- Natural prosody and facial expressions.



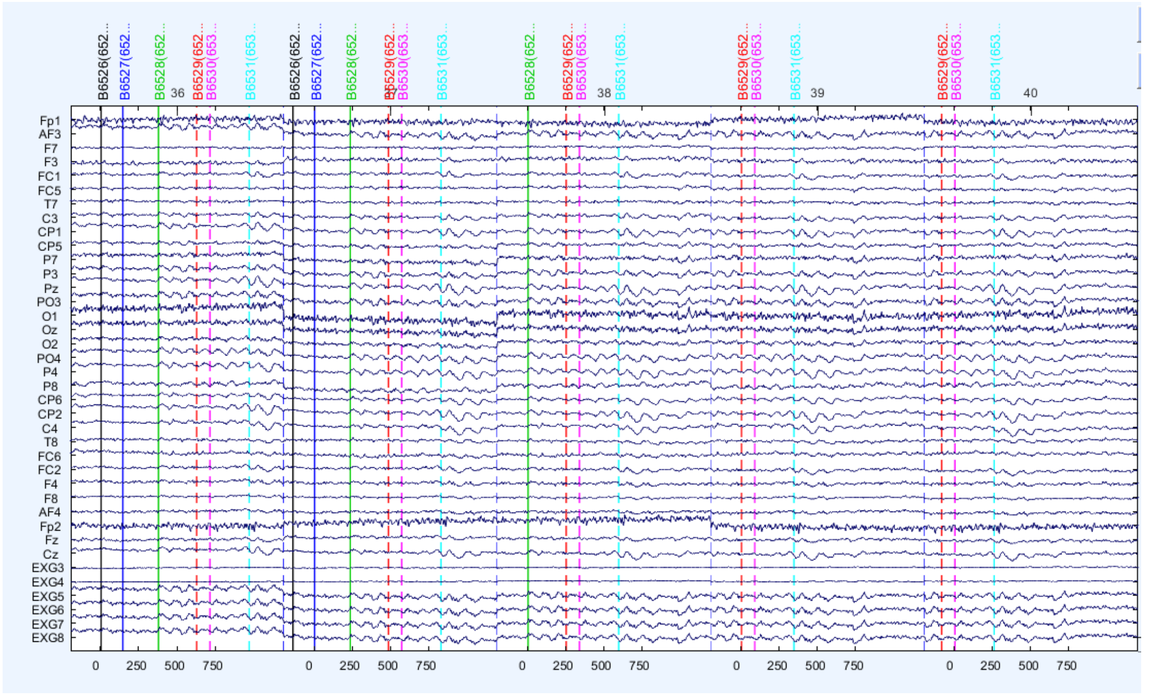
Participants

Two experimental conditions:

- Audio-only setting - where participants only listen to the speaker

- Audio-visual setting - where participants both listen to and watch the speaker

Data



EEG data was collected for both group

Identical lexical information across both settings

Surprisal estimates

We obtain surprisal estimates using log-probabilities through:

- n-gram language models
  - we vary the context windows and consider 2,3,4,5 and 6-gram models
- transformer based language models
  - access to infinite lexical context

Comparing models

Baseline models: only consist of information from the location of EEG electrodes (ROI).

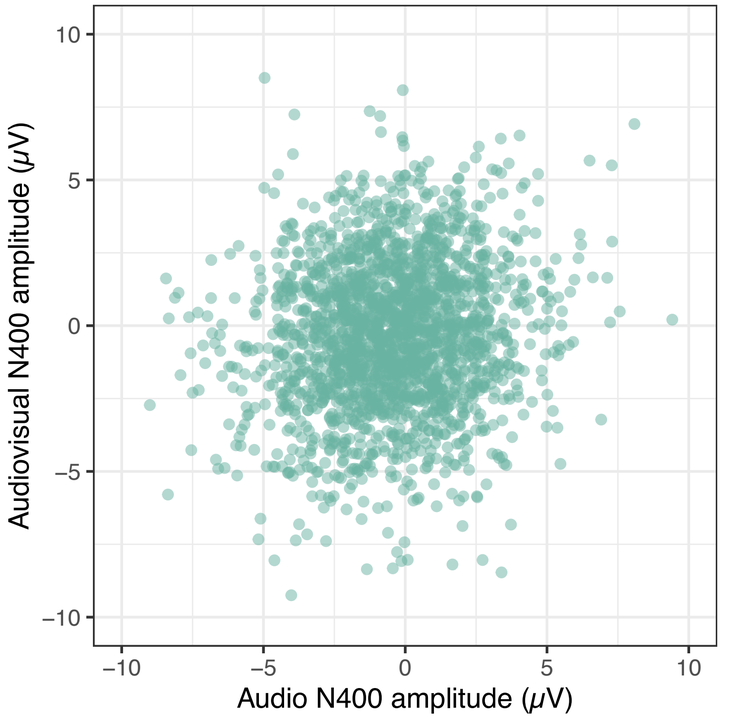
Participant, passage and electrode as random intercepts to control for individual behavioural effects.

We consider both additive and multiplicative models of surprisal

We fit a use linear mixed-effects model and consider the difference between the akaike information criterion (ΔAIC) of the models with surprisal and the baselines.

Observations: **N400** signals across the experiments

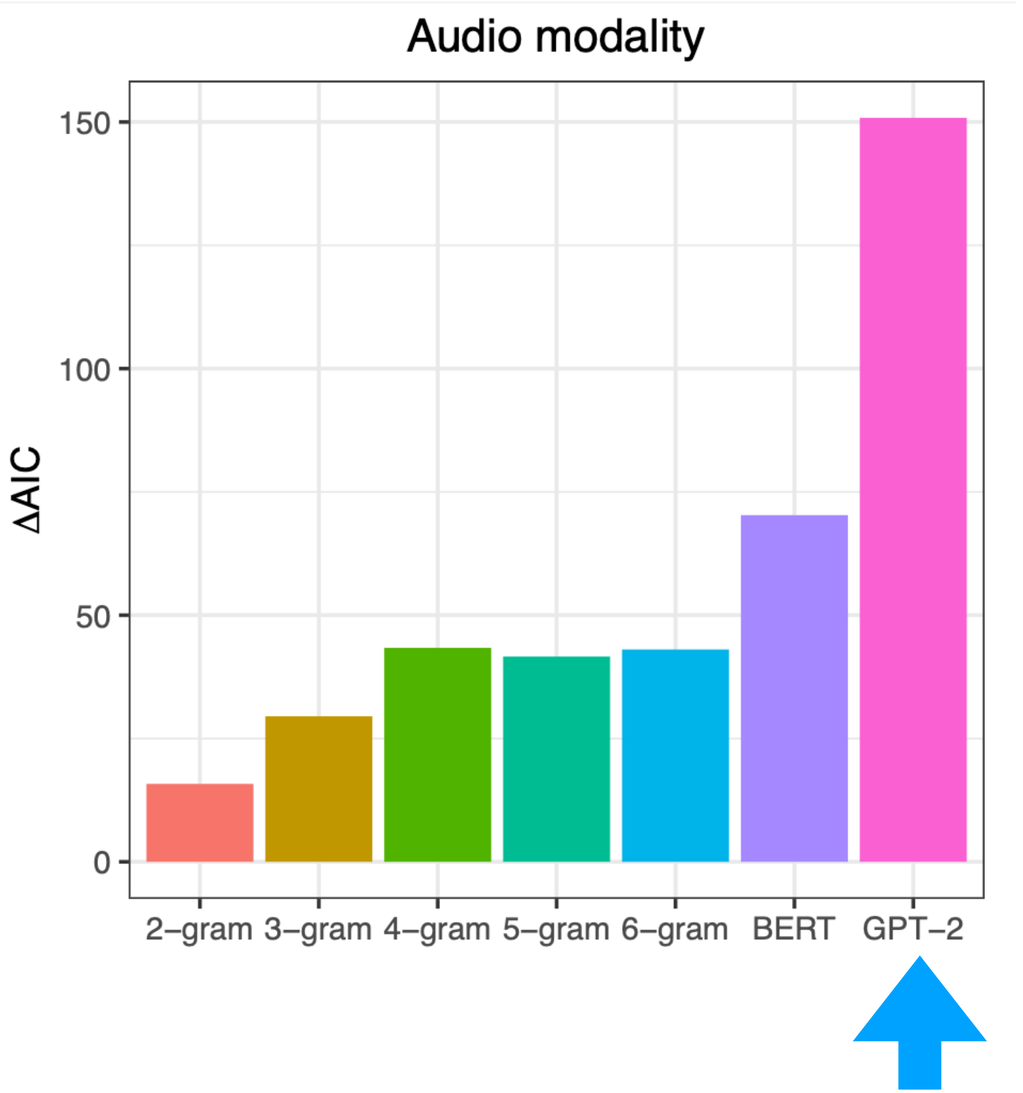
- Weak correlation
- If the lexical information were the most significant contributing factor, we would expect a stronger correlation between audio-only and audio-visual conditions since both experiments involve the same verbal stimuli.
- This indicates that multimodal signals significantly modulate N400 (more than lexical information)



$$r = 0.11 (p < 0.001)$$

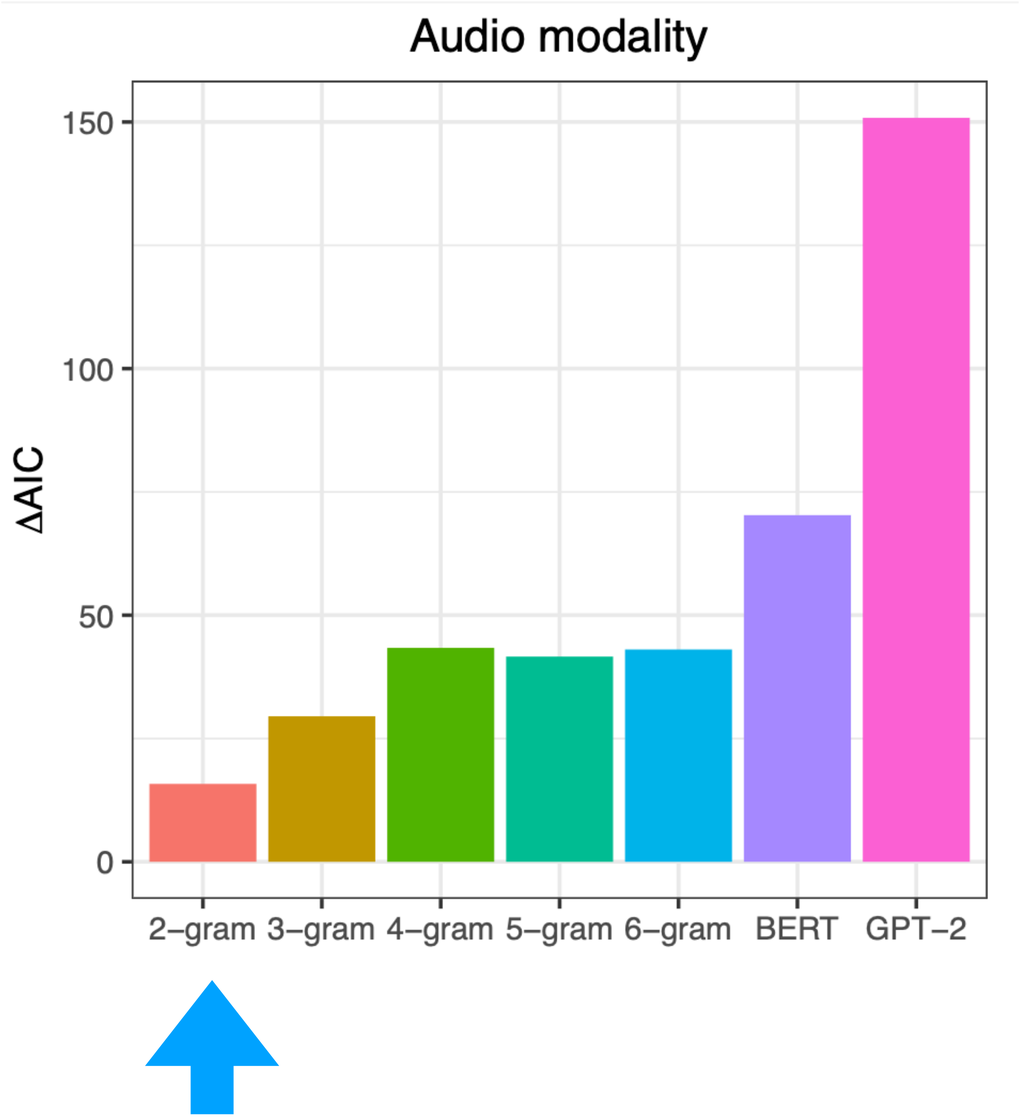
Audio only: prefer models with longer contexts

- The largest reduction in AIC compared to the baseline model is observed with GPT-2.



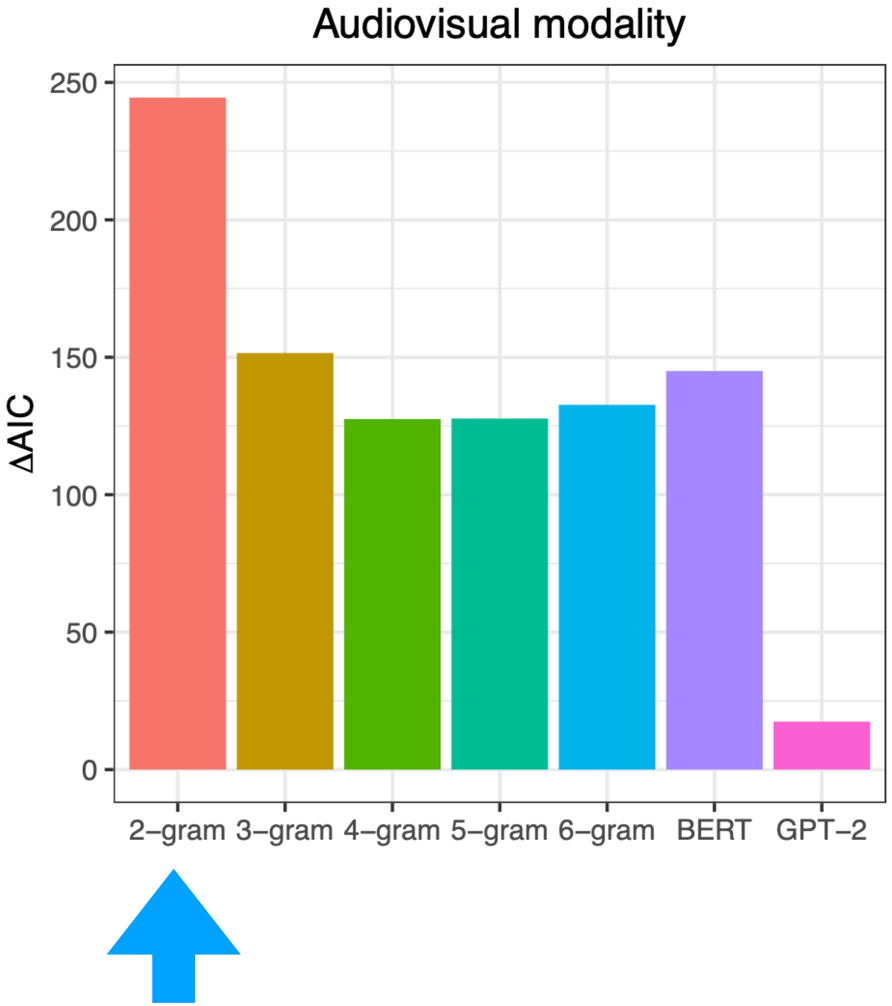
Audio only: prefer models with longer contexts

- The largest reduction in AIC compared to the baseline model is observed with GPT-2.
- While the 2-gram model shows the smallest reduction in AIC.



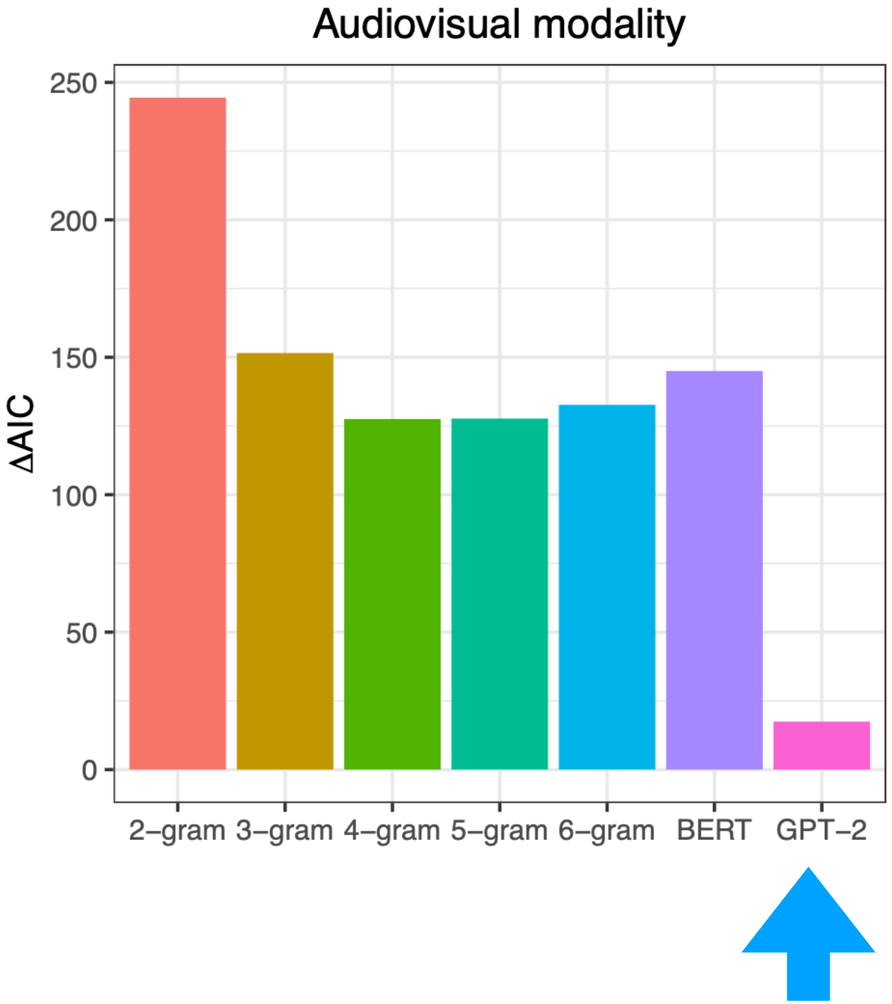
Audio-visual: prefer models with shorter contexts

- 2-gram model shows the largest reduction AIC

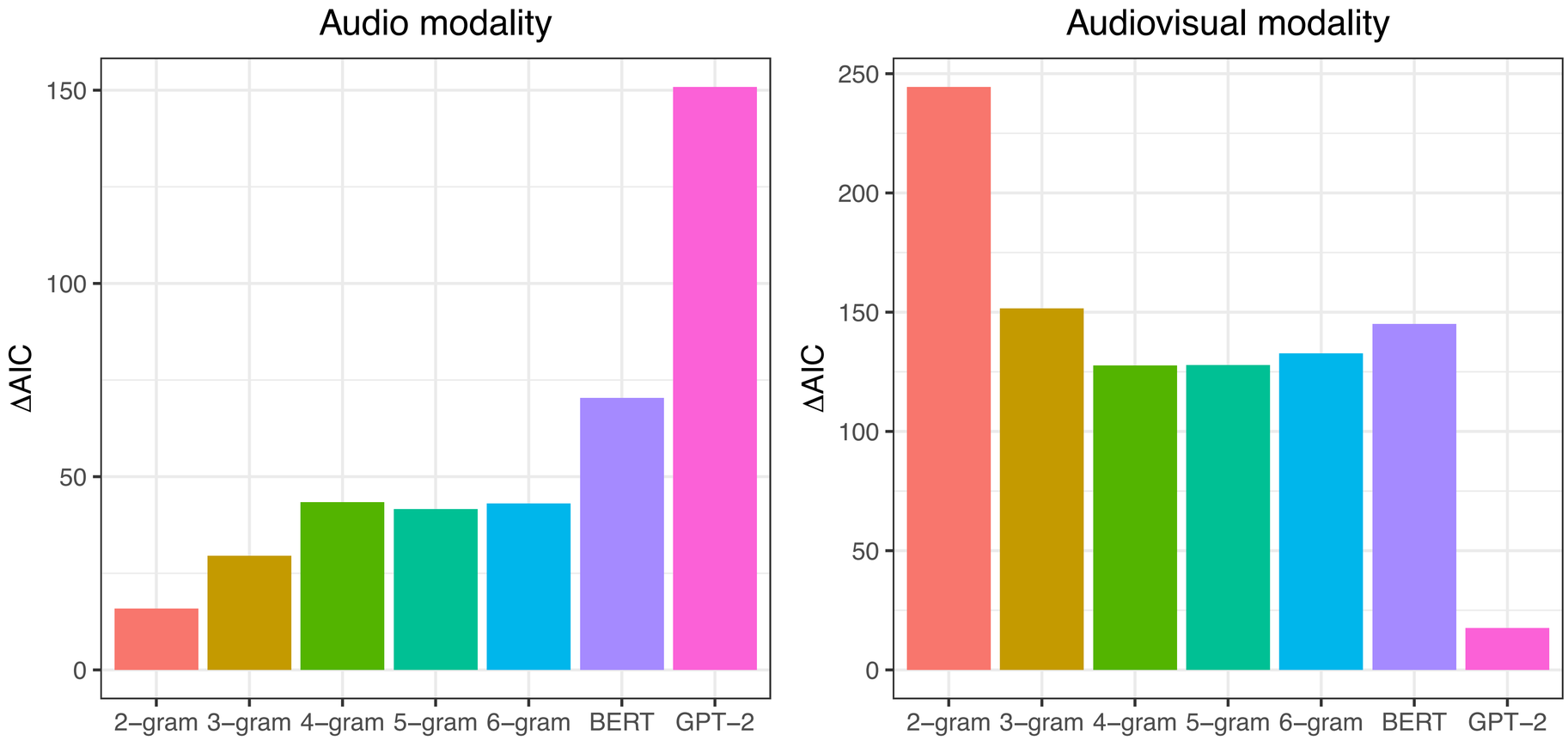


Audio-visual: prefer models with shorter contexts

- 2-gram model shows the largest reduction AIC
- On the other hand, GPT-2 which has access to the largest context window, shows lower reduction in AIC.

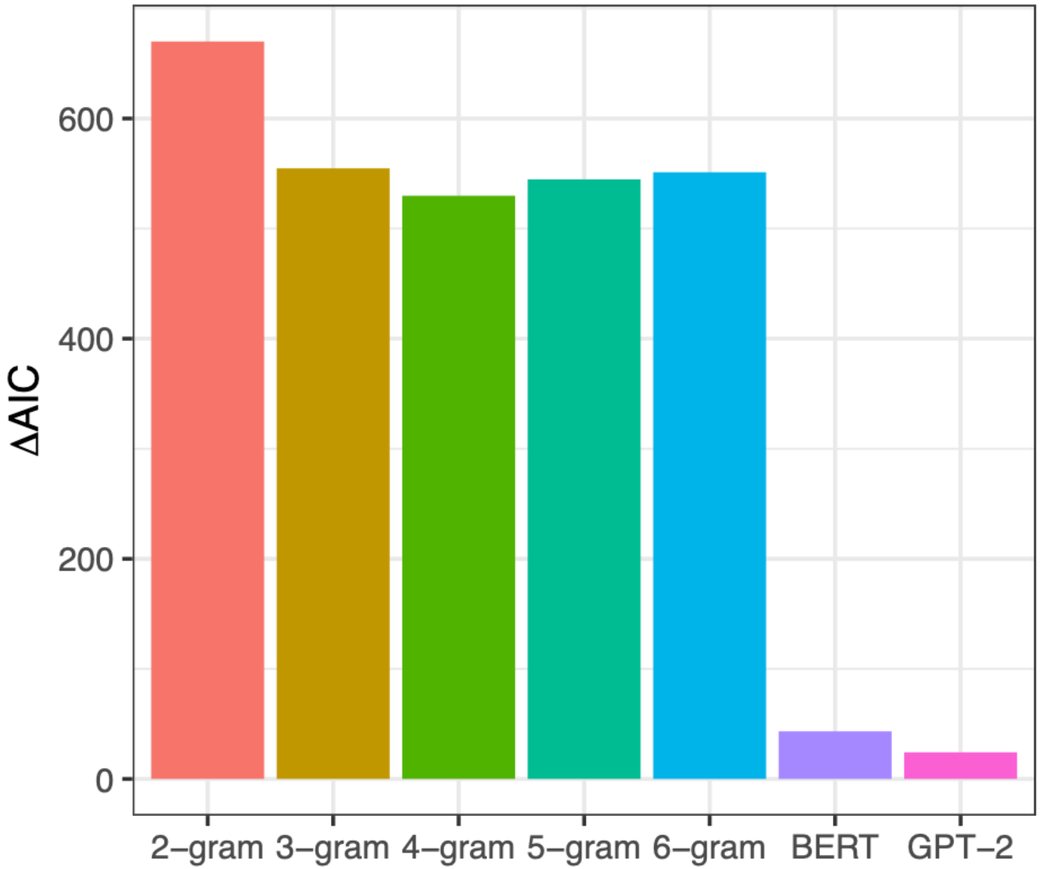


Reversal of trend



Replication

We observe similar behaviours in a setup where the audio visual stimuli consist of passages that are sampled from more realistic corpus.



Discussion

We replicate recent findings of language models with larger context windows to have better correlations with neurophysiological signals of cognitive effort in the unimodal setting (Michaelov & Bergen, 2022).

However, Language model context window plays a significant role in predicting N400 under two different conditions with the same lexical stimuli.

This study raises important questions on the importance of including multimodal channels of communication in information theoretic frameworks

Key takeaways

Under similar lexical stimuli, we observe that multimodal cues significantly modulate the N400 signals

Cognitive effort differs significantly between multimodal (audio visual) and unimodal (audio only) settings

Local lexical context plays a significant role in cognitive processing in a multimodal environment

Communicating explanations or interpretations perhaps needs to take multimodality in to effect?