

Applying Language Models to Archival Information Extraction

Today

- Part I: What do I mean by archival information extraction?
- Part II: How do I use LLMs for this?

Part I

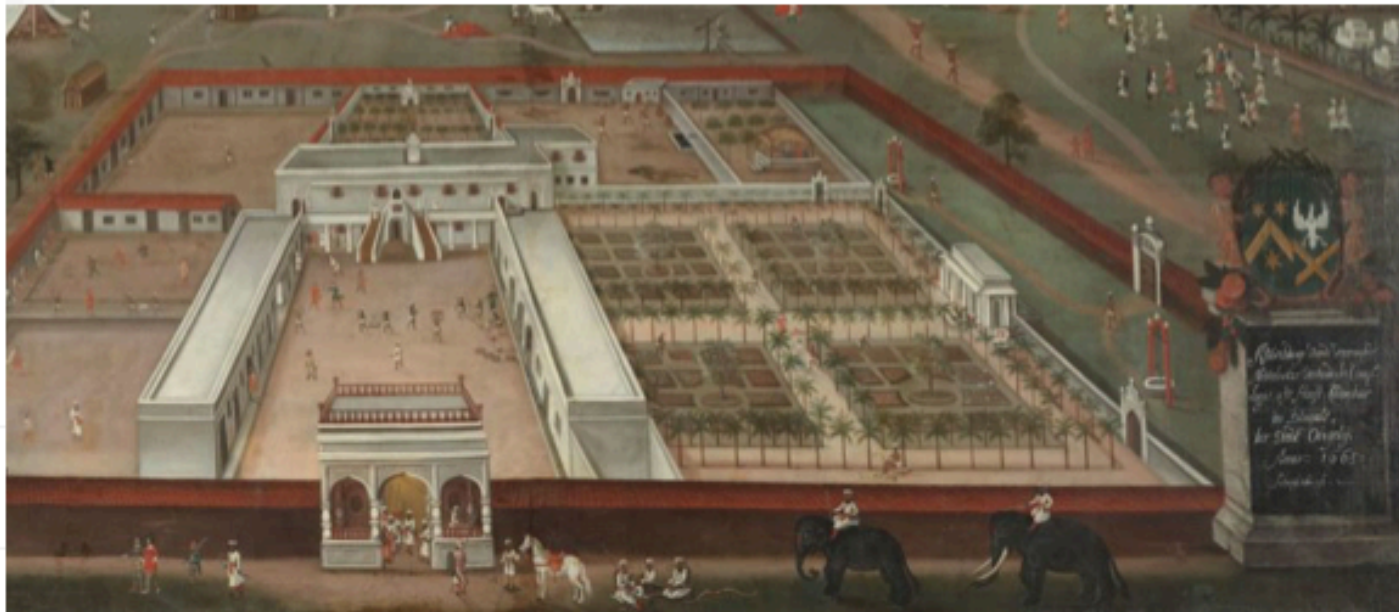
GLOBALISE

GLOBALISE

Project +

Blog & Newsletter +

Collaborations +



GLOBALISE

Unlocking the history of early globalisation and
colonialism for researchers and the general public

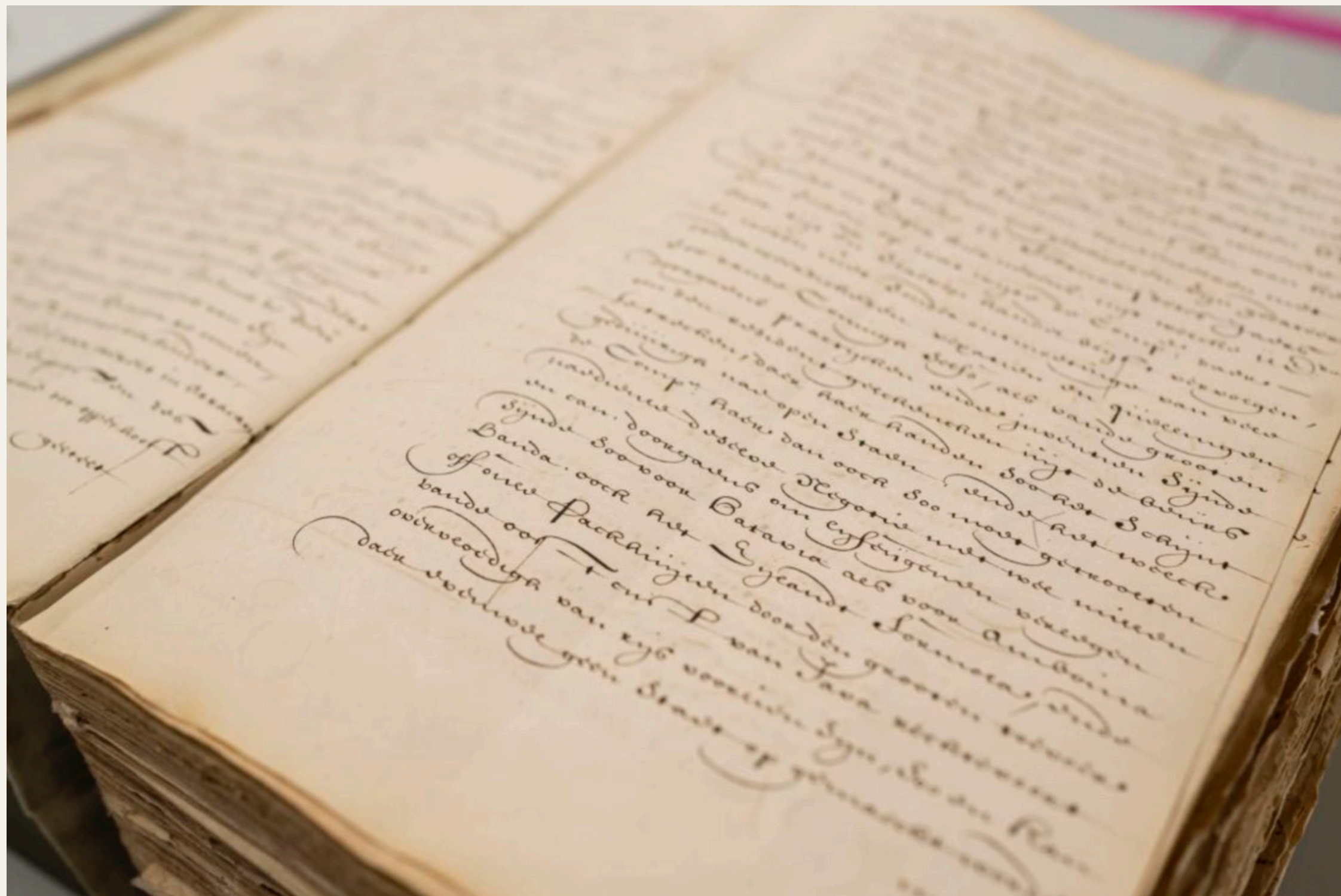
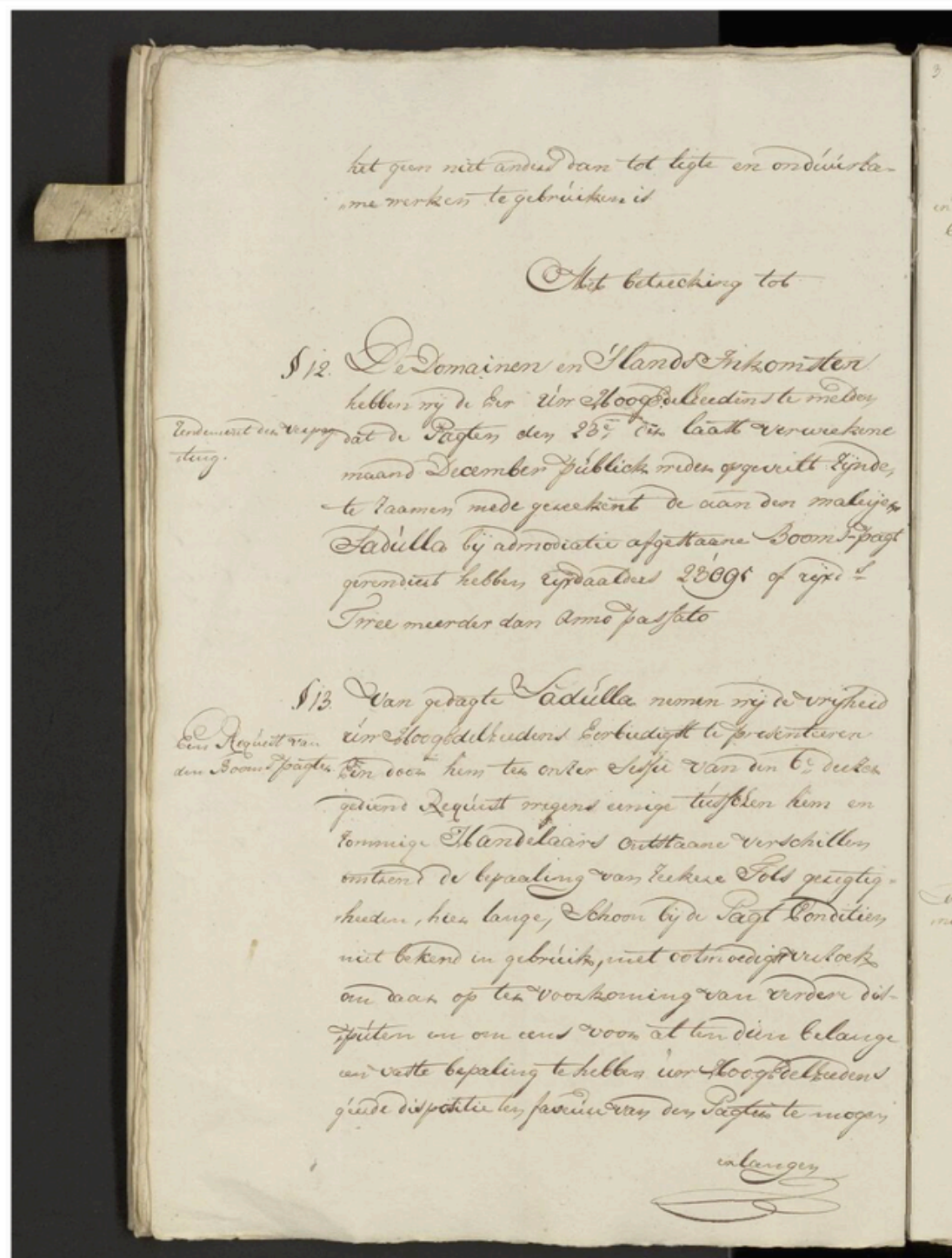


Photo Dave Straatmeyer



Transcription

Standard

Diplomatic

„ting. Een Request van het geen niet anders dan tot ligte en onduurtae= „me werken te gebruiken is. Met betrecking tot § 12. De Domainen en 'Slands Inkomsten. hebben wij de Eer uw: Hoog Edelheedens te melden rendement der verpag dat de Pagten den 23„e der laatst verweekene maand December publick weder opgeveilt zijnde, te zaamen mede gerekent de aan den maleijer Sadulla bij admodiatie afgestaane Booms-pagt gerendeert hebben ryxdaalders 23095 of rijxd=s Twee meerder dan Anno passato § 13. van gedagte Sadulla nemen wij de vrijheid uw Hoog Edelheedens Eerbiedigst te presenteren den Booms-pagter Een door hem ter onzer sessie van den 6„e deezer gediend Request wegens eenige tusschen hem en zommige Handelaars ontsaane verschillen omtrend de bepaaling van zekere Tols geregtig= „heeden, hier lange, Schoon bij de Pagt Conditien niet bekend in gebruik, met ootmoedigst verzoek om daar op ter voorkoming van verdere dis= „puten en om eens voor al ten dien belange een vaste bepaling te hebben uw Hoog Edelheedens g'eerde dispositie ten faveure van den Pagter te mogen

'Event Extraction'

Since our latest message left Batavia with the ships Frederick Hendrik, Hollandia, 't Wapen van Delft 's landts Hollandia, e and De Galias on the 12th of November last year and came through the Sunda Strait on the 15th of the same month, did, God bless, on the 17th and the 21th from Taiwan and Japan successively arrive de Erasmus and de Vreede, together loaded with 810 picul raw Chinese silk and 332 picul copper

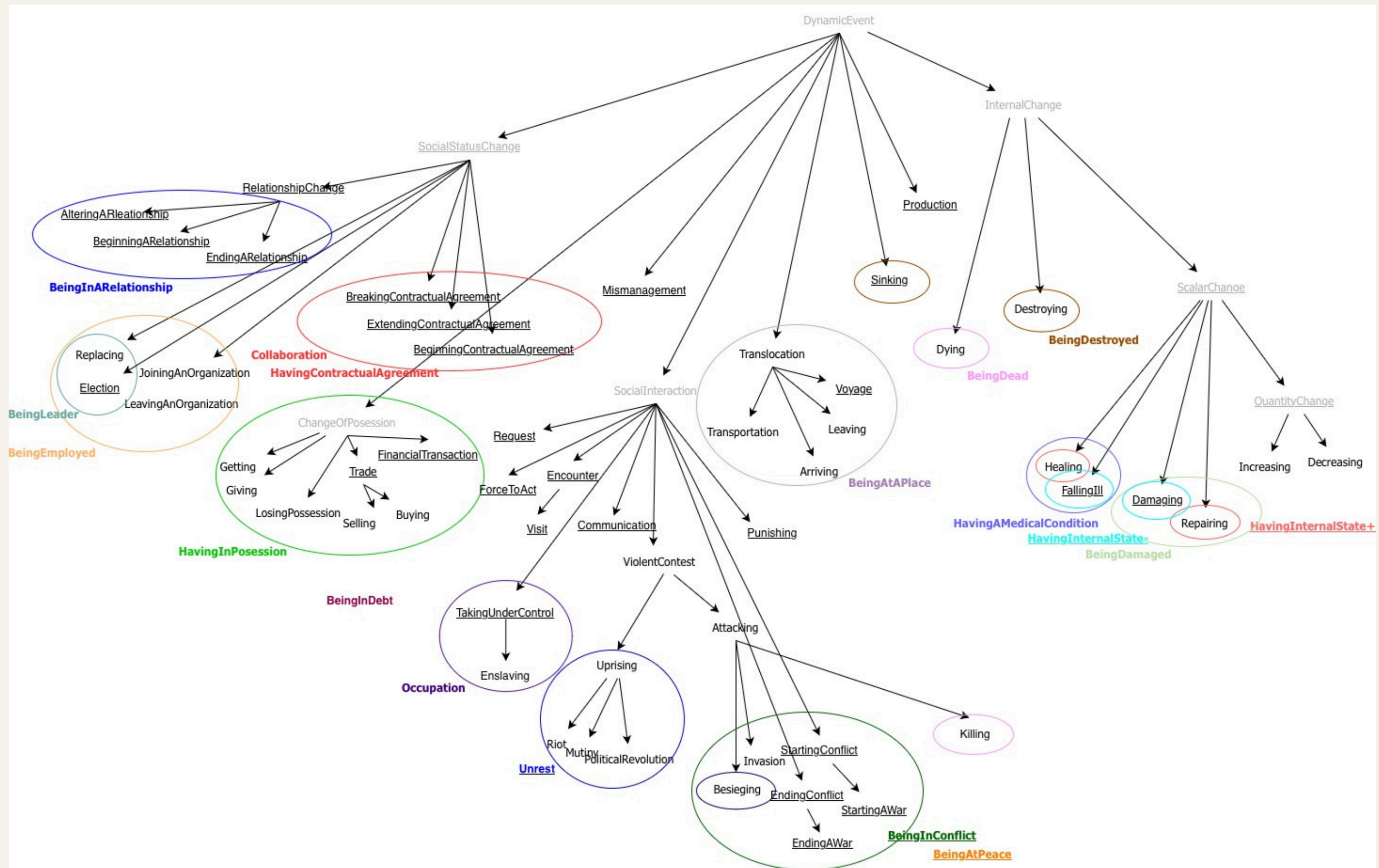
'Event Extraction'

*Since our latest message **left** Batavia with the ships Frederick Hendrik, Hollandia, 't Wapen van Delft 's landts Hollandia, e and De Galias on the 12th of November last year and **came through** the Sunda Strait on the 15th of the same month, did, God bless, on the 17th and the 21th from Taiwan and Japan successively **arrive** de Erasmus and de Vreede, together **loaded with** 810 picul raw Chinese silk and 332 picul copper*

Voorbeeld van 'Event Detection' in het VOC archief

*'Tsedert onsen Jongsten (hiernevens in copie gaande) met de schepen Frederick Hendrik, Hollandia, 't Wapen van Delft, 's landts Hollandia, ende de Galias den 12en November passato in compagnie van Batavia **gescheyden** ende den 15en ditto door de Strate Sunda **geraeckt**, sijn hier, Godtloff, den 17en ende den 21en van Teyouhan ende Jappan successive wel **aengecomen** 't jacht Erasmus ende 't schip de Vreede, t'samen **geladen** 810 picol rouwe Chineesche syde ende 332 picol coper.'*

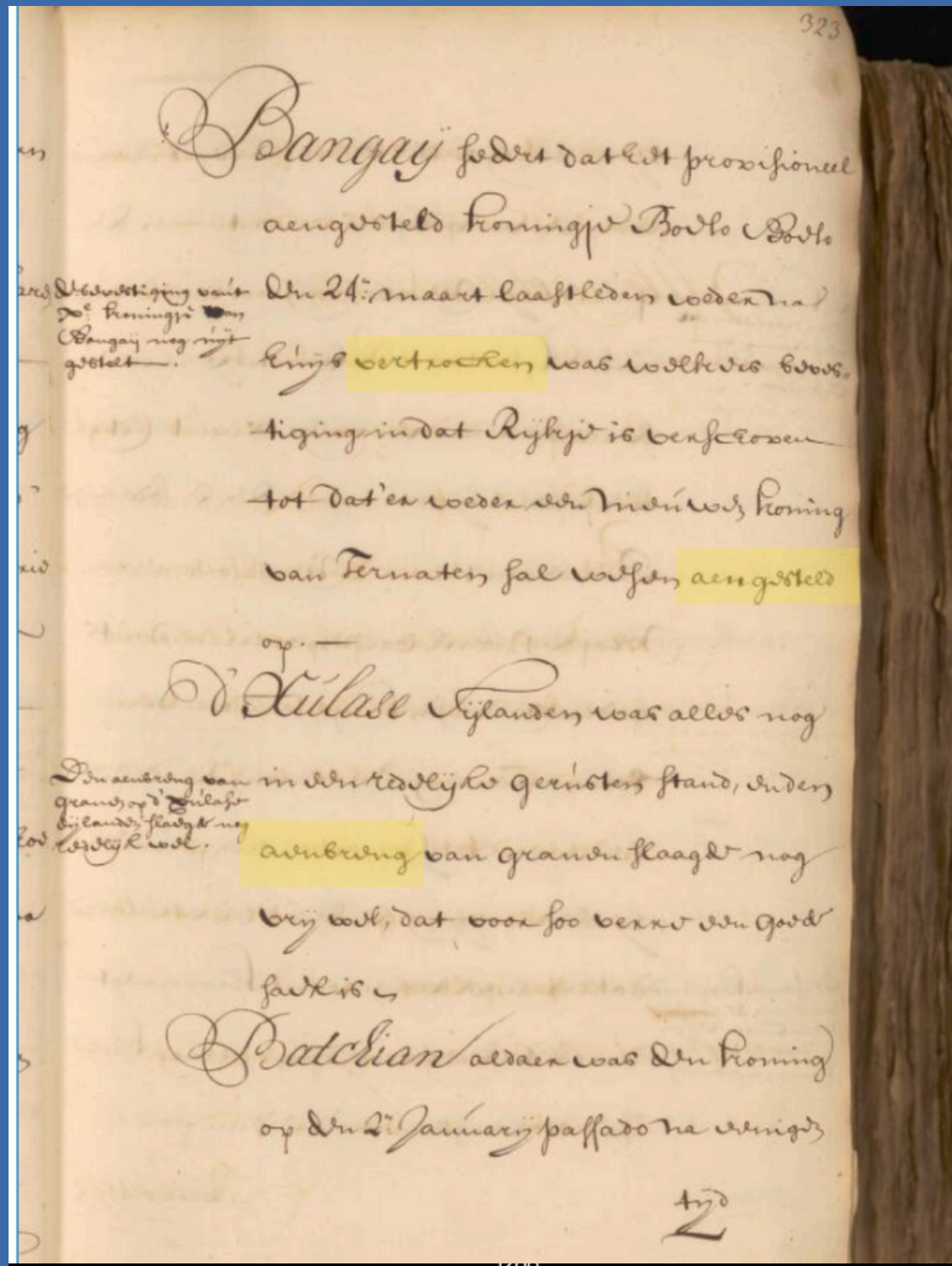
- 'gescheyden' = Leaving
- 'geraeckt' = Translocation
- 'aengecomen' = Arriving
- 'geladen' = Transportation



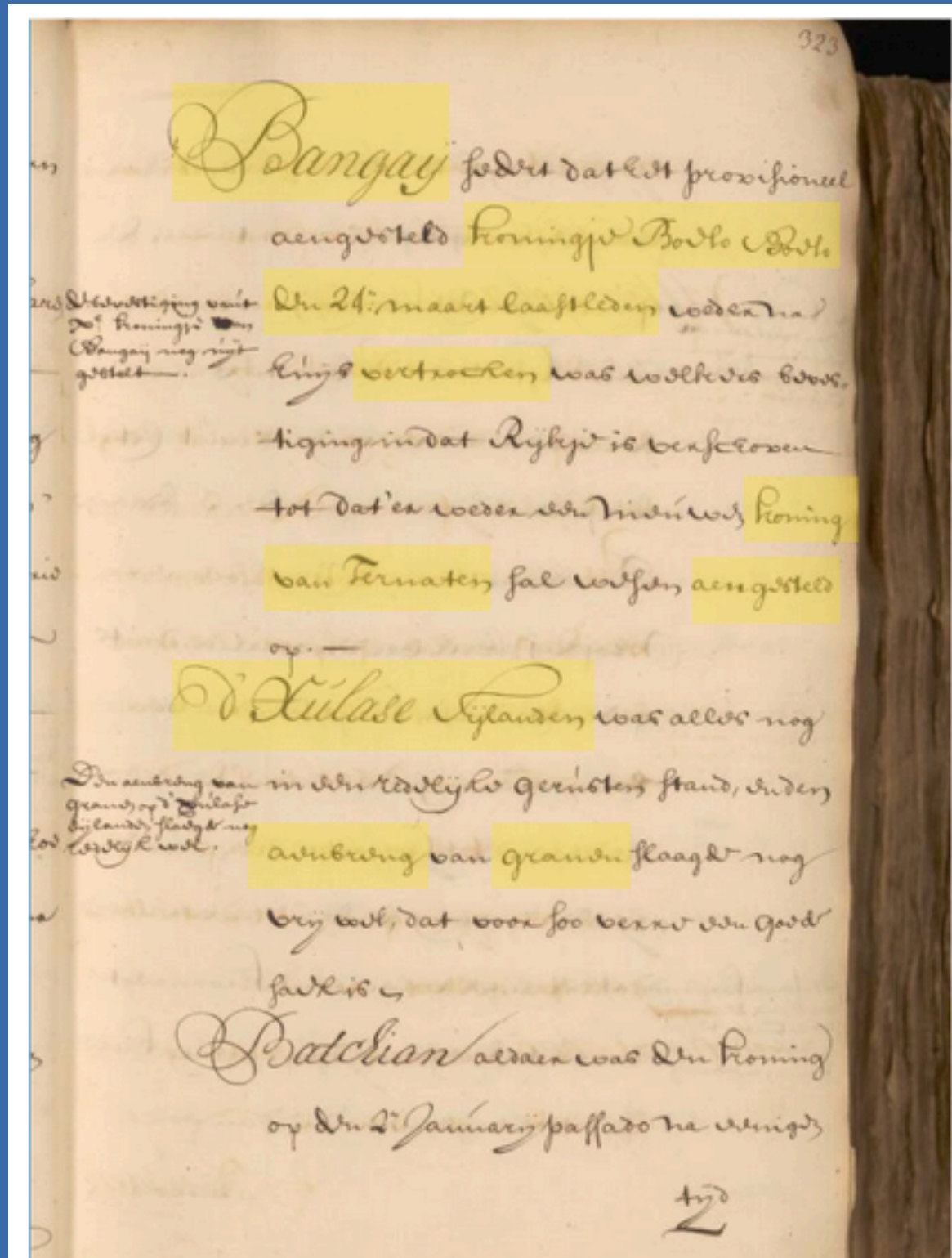
'Event Extraction'

*Since our latest message **left** Batavia with the ships Frederick Hendrik, Hollandia, 't Wapen van Delft 's landts Hollandia, and De Galias on the 12th of November last year and **came through** the Sunda Strait on the 15th of the same month, did, God bless, on the 17th and the 21th from Taiwan and Japan successively **arrive** de Erasmus and de Vreede, together **loaded with** 810 picul raw Chinese silk and 332 picul copper*

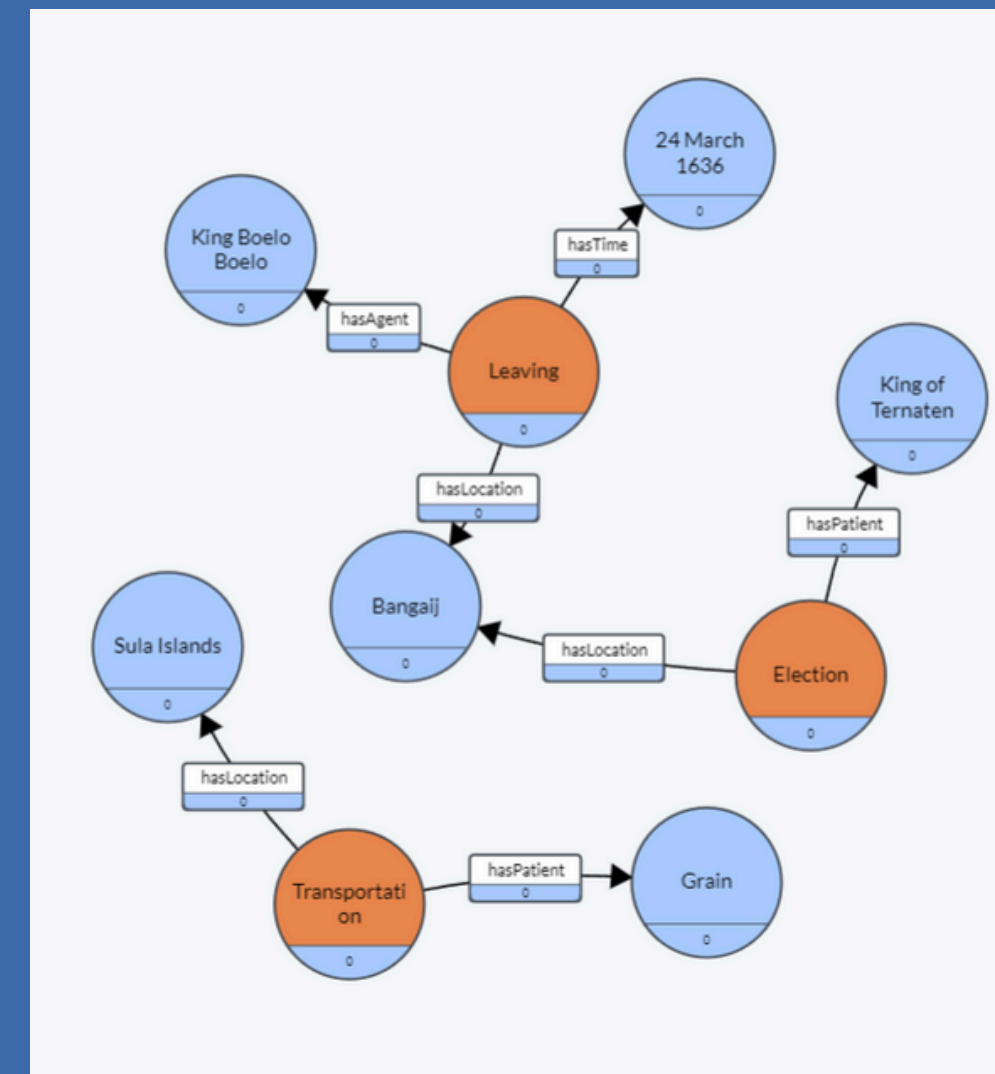
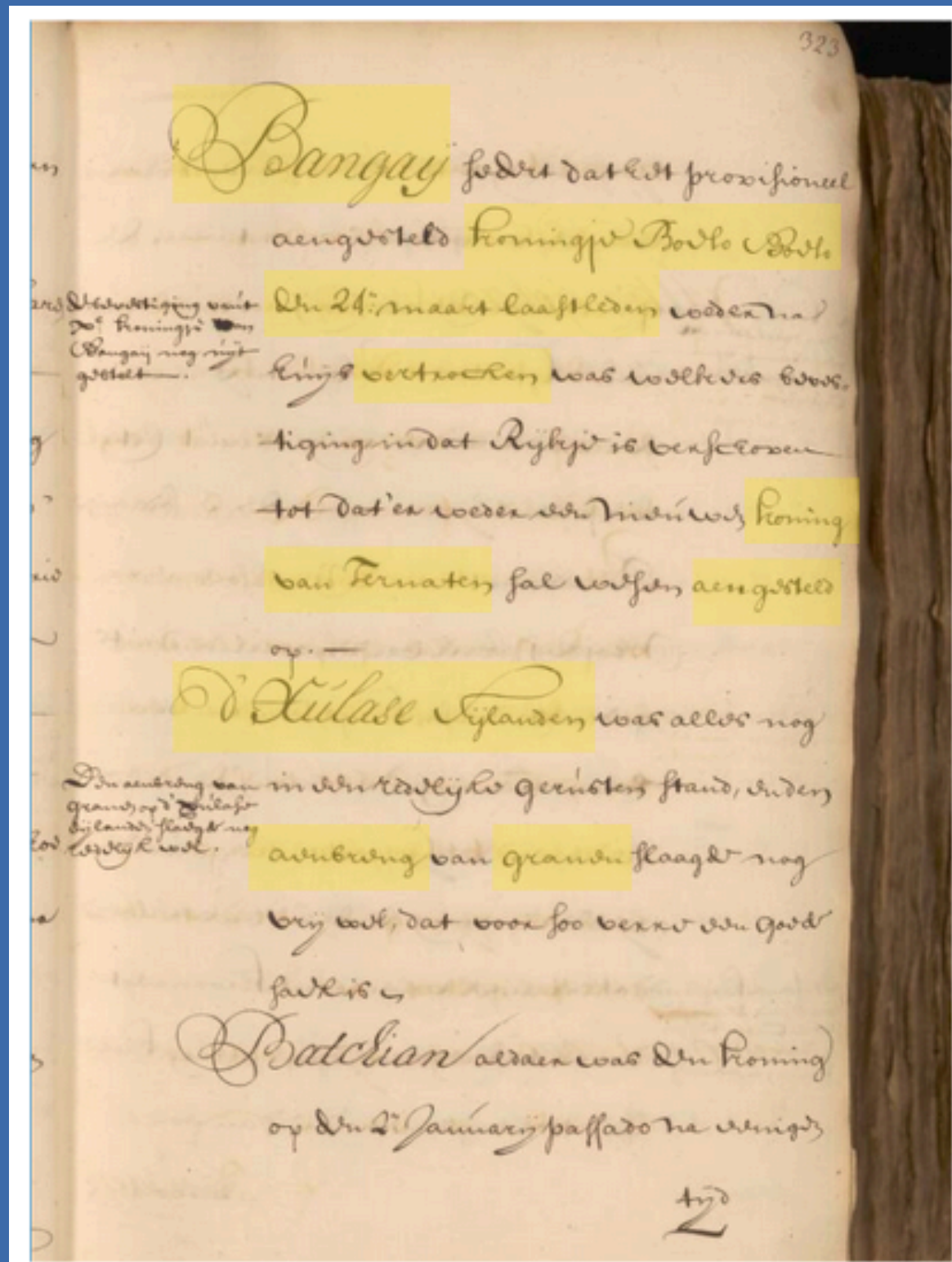
Voorbeeld van 'Event Detection' in het VOC archief



Voorbeeld van 'Event Detection' + 'Entity Recognition' in het VOC archief



Voorbeeld van Machine Learning + Knowledge Representation



Part II

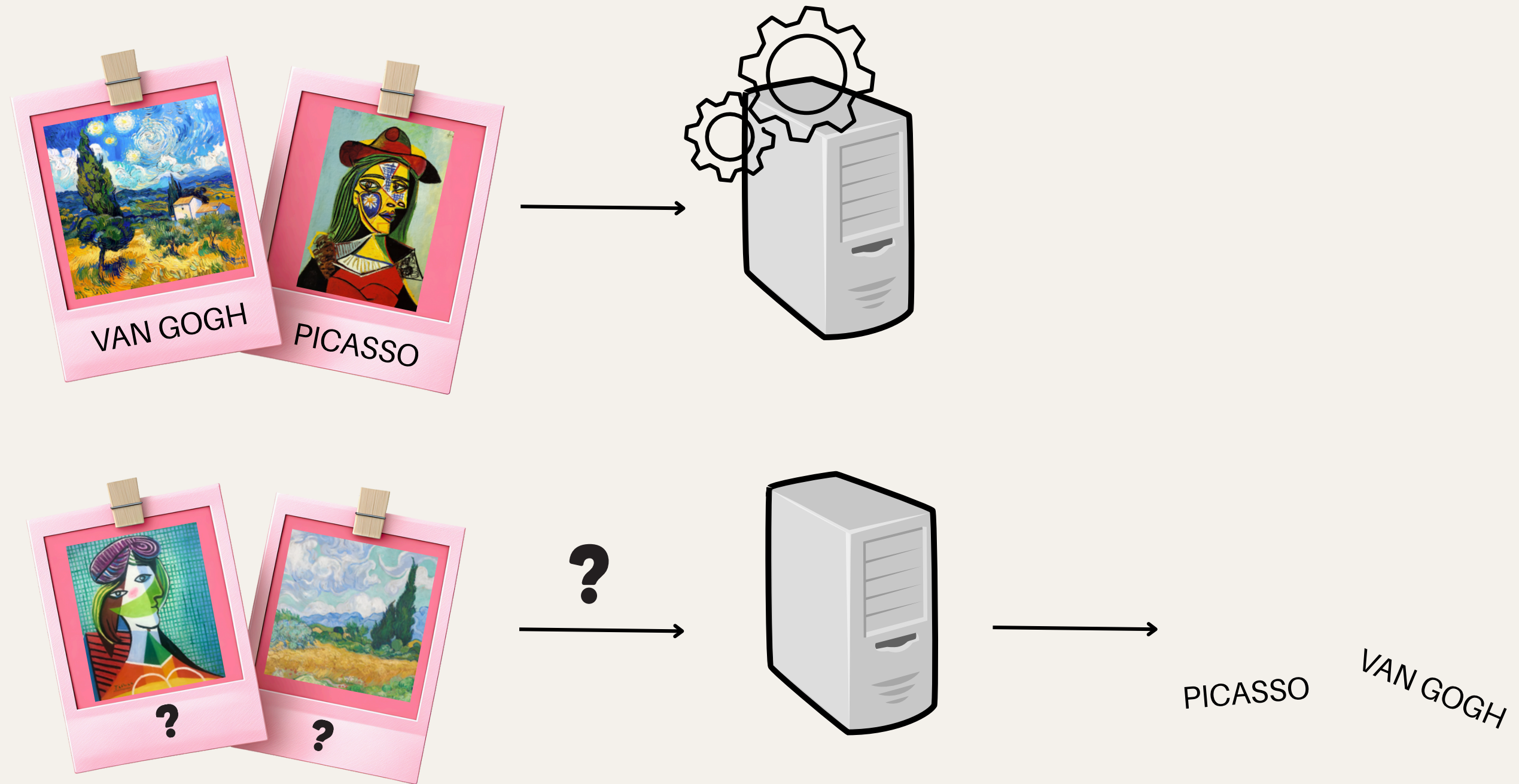
Language Modelling: encoders vs. decoders

- Both: Distributional semantics
- Encoders: Designed to be used for predictive modelling tasks such as classification
- Decoders: Designed for generating text
- Both: You don't need labelled examples to train

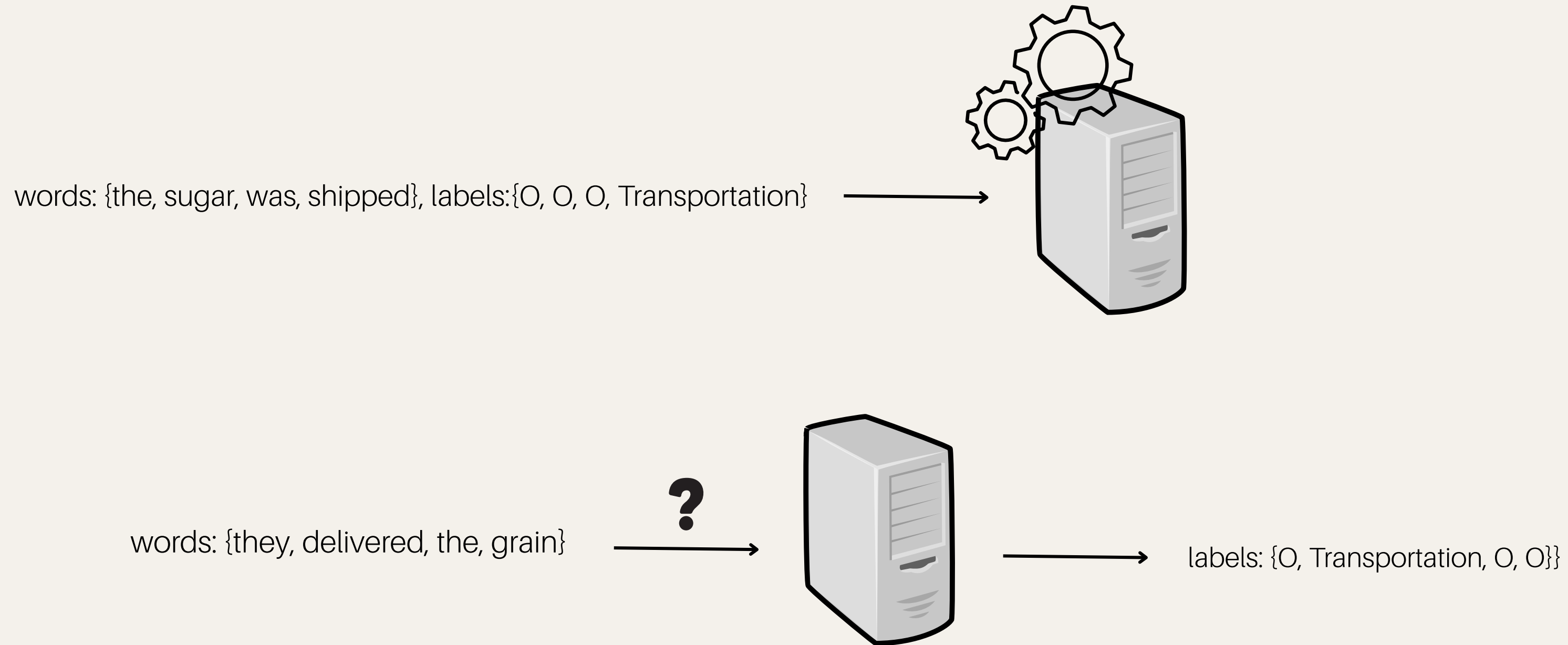
Language Modelling: encoders vs. decoders

- Both: Distributional semantics
- Encoders: Designed to capture semantic representations of words
- Decoders: Designed for generating words
- **Both: You don't need labelled examples to train**

Supervised learning

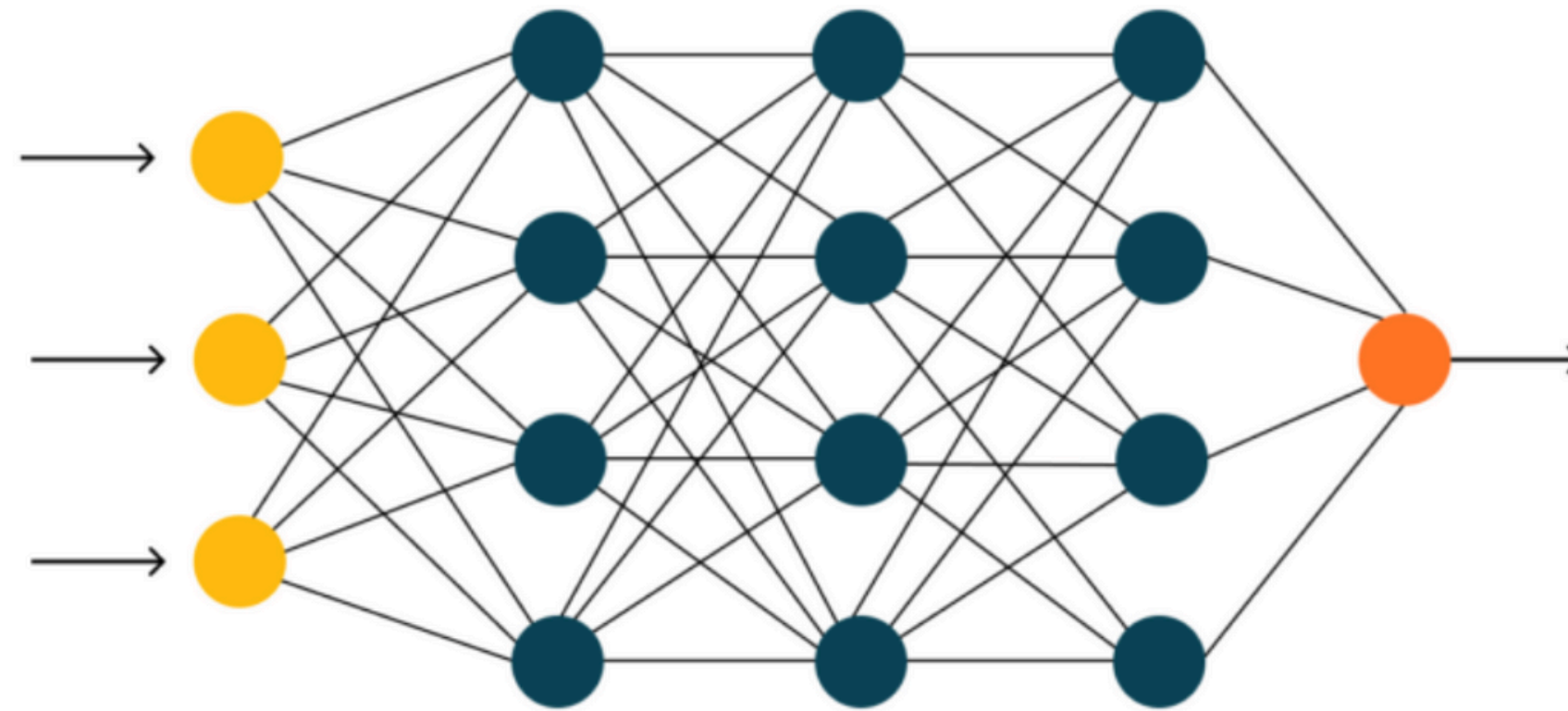


Supervised learning



BERT: an **encoder**

"In the [MASK] people like
to drink hot chocolate"

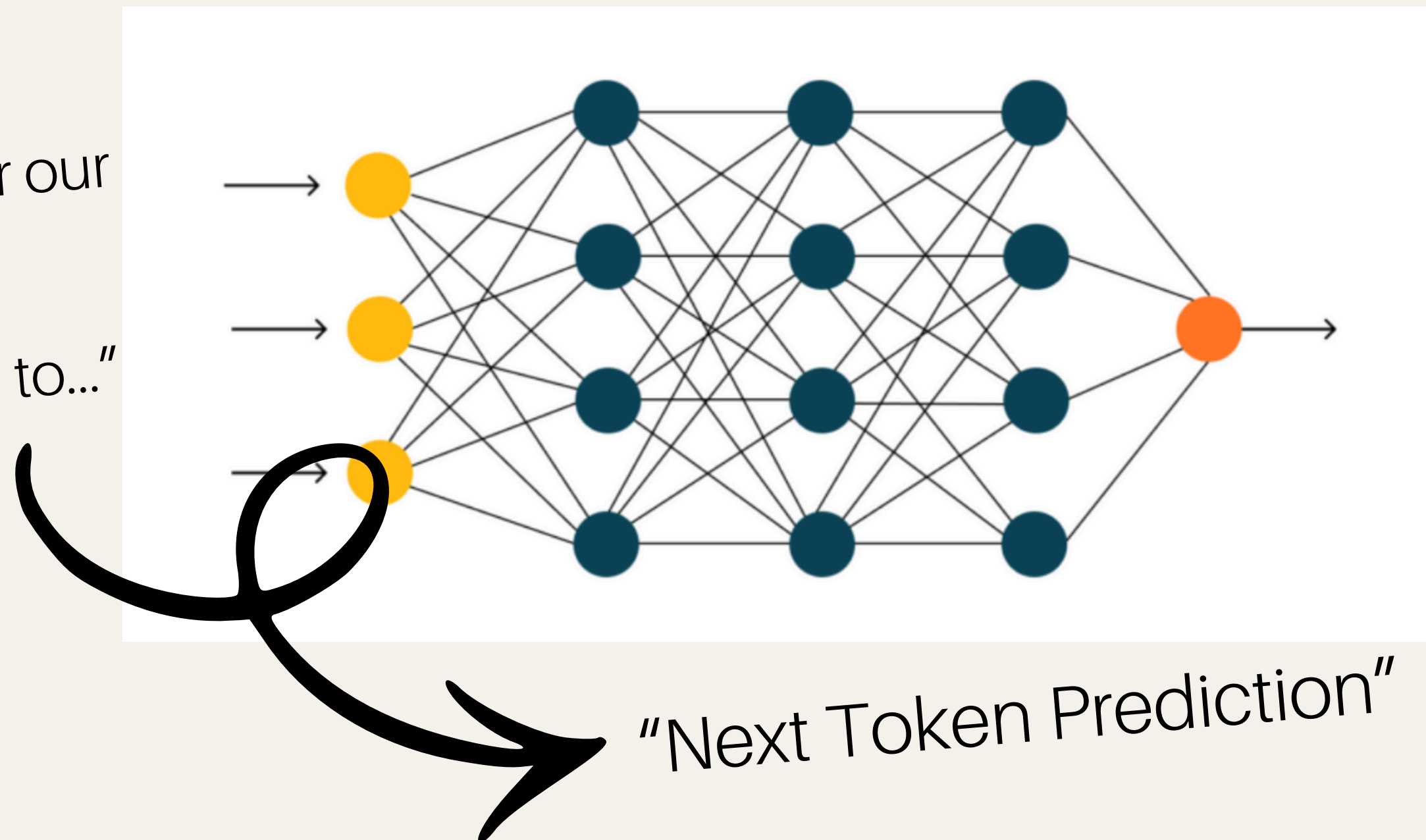


winter

"Masked Language Modelling" (MLM)

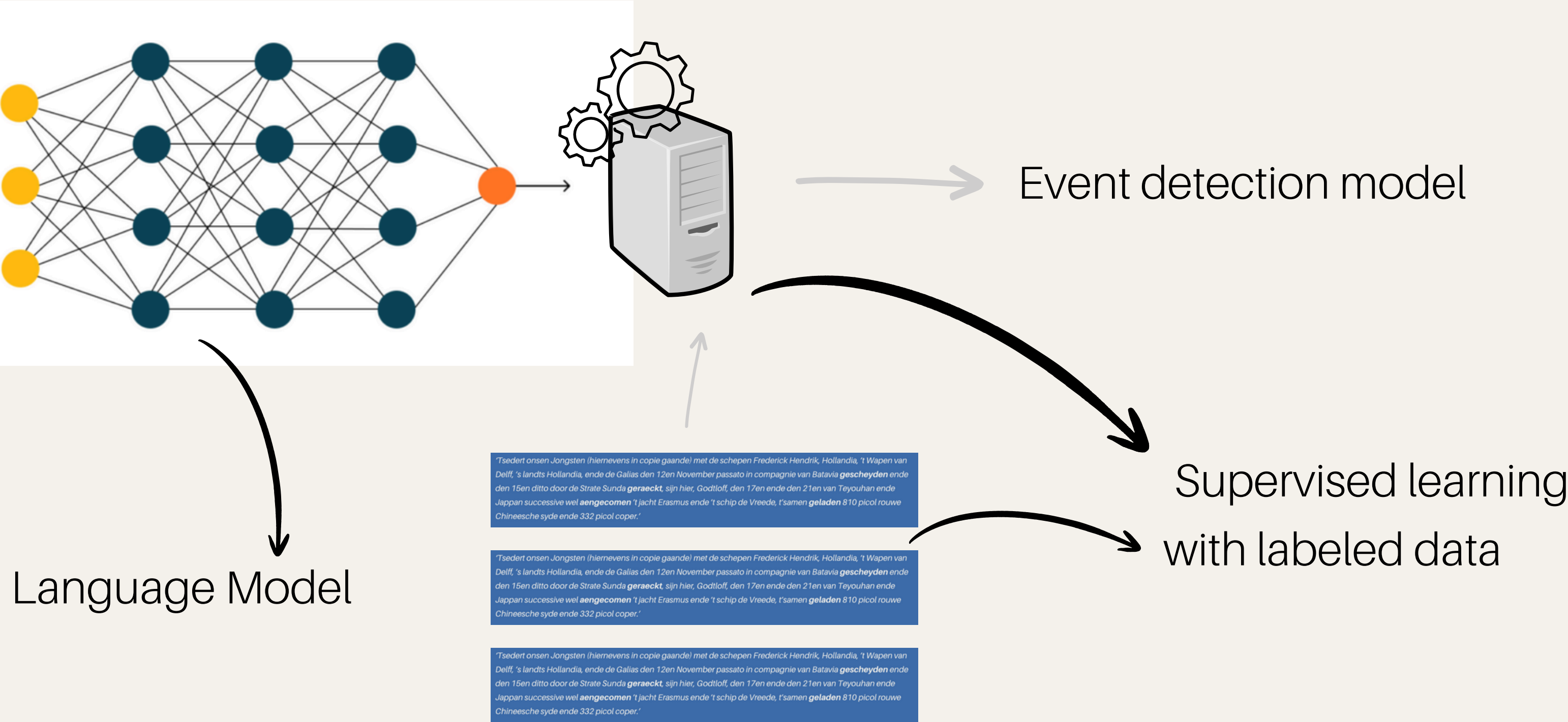
GPT: a decoder

"Dear Vanessa, after our
short conversation
yesterday I wanted to..."



?

(encoder) LLMs are a great basis for more specific NLP tech



INCEpTION

ProjectsDashboard

HelpAdministrationAPIStella VerkijkLog out29 min

Active Learning

Currently not available

Read-onlyTeam KruidnagelGLOBALISE Annotation 2023

NL-HaNA_1.04.02_1812_0803-0808 - 1711 - Originele missive door Willem Backer Jacobsz...en den Gamronse raad aan den Gouverneur Generaal en...1-20 / 23 sentences [doc 4 / 11]

FolderDownloadPreviousNextUndoRedoFirstPreviousNextLastRefreshRepeatSettings

1

EyePreviousNext

RecyclePreviousNext

Settings

1 debvoir kennise te laten toecomen, Even als wij bij die van den 20,,e xb: haer nopende de zijde

Leverantie voor dit 9 Jaer hebben gedaen, en g'ordonneerd, die des mogelijk in de moolen te helpen,

ook gequalificeerd, soo dat niet Lucken wilde; en mier sahit inhamet halsterrig en oneanderlijk op

de zijde Leverantie aendrong die te ontfangen, sampt bij die van den 29,,e daer aen volgende en 15,,e Jan:

haer indagtigd, want dat was al meede vergecten, niet tegenstaende haer op den 15,,e maij daer over gebruikte

breeder resonnementen, miet sahit inhamed op ontfangst van dat „schrijvens, dato 29,,e xb: te laten

versoecken, dat hij ten Eersten zomee conconerenordre geliefde te stellen, aen wie dat

SpanClear

←→

LayerEvent

Textontfangen

No links or relations connect to this annotation.

arguments

Agent: SemArgmier sahit inhamet

Patient: SemArgde zijde Leverantie

Enter role name...Add

categoryGetting

relation type

evokes ⓘ

isOfType ⓘ

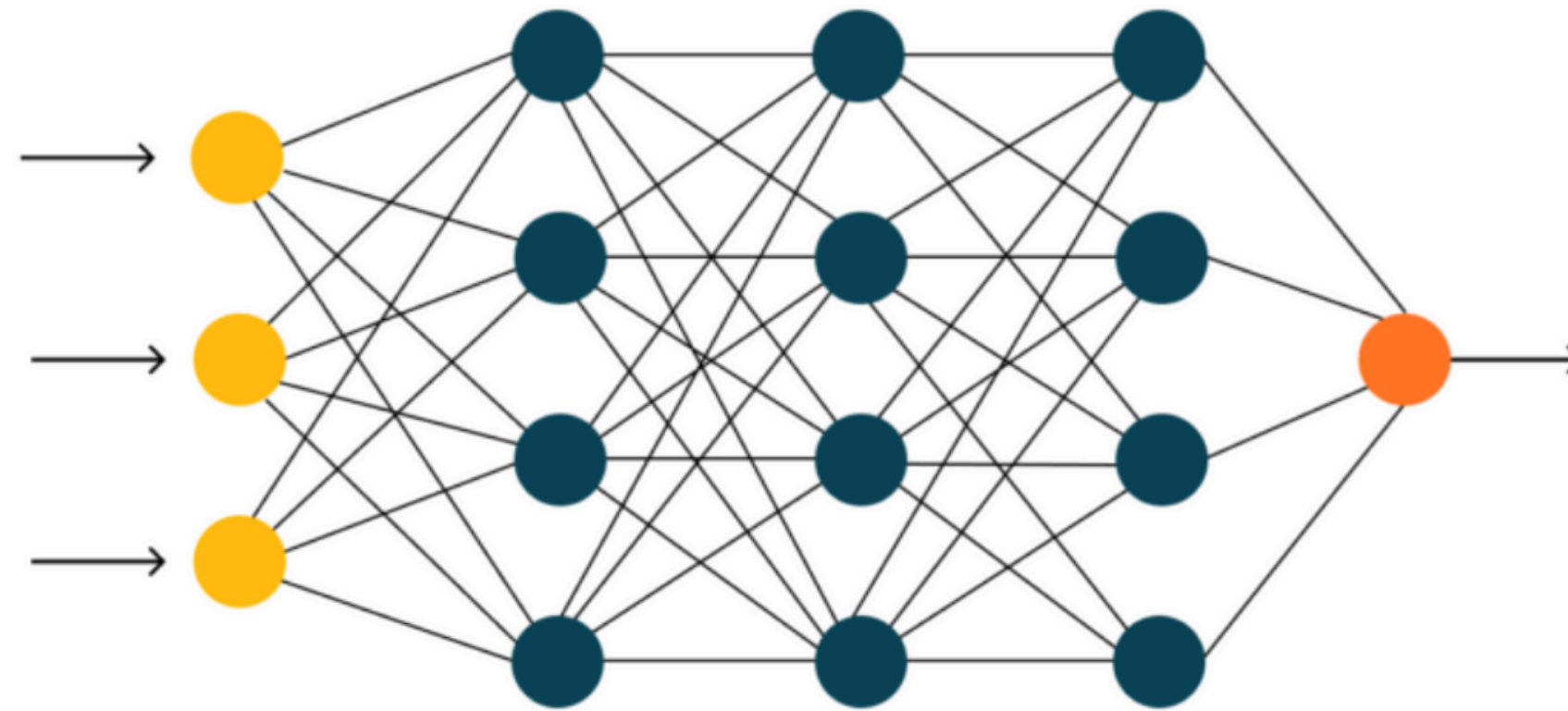
There are many types of encoders

- Multilingual / Monolingual
- BERT / RoBERTa
- Amount of parameters
- Amount of pre-training data
- Domain of pre-training data

The power of domain-specific pre-training

~~"In the [MASK] people
like to drink hot
chocolate"~~

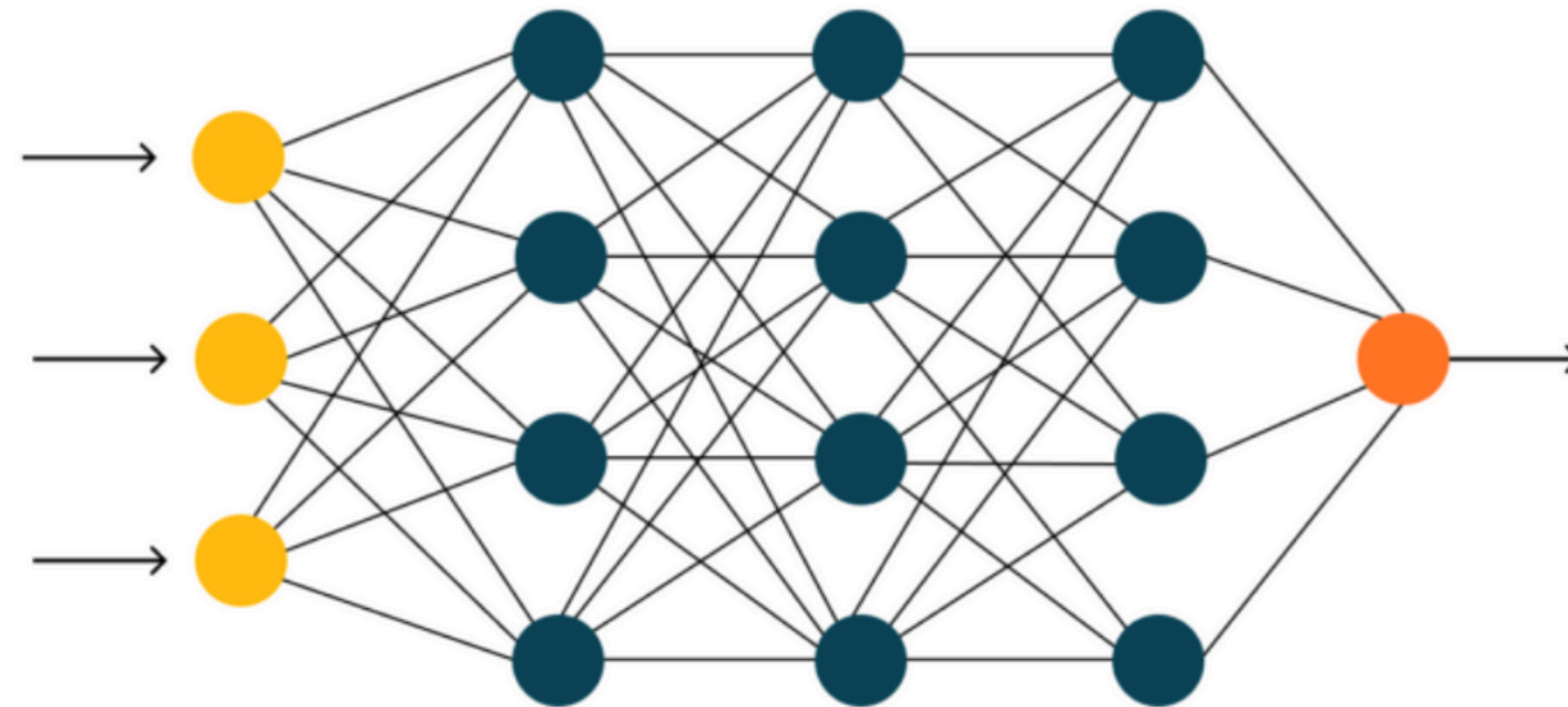
"In de [MASK] drinken
mensen graag warme
chocolademelk"



winter

The power of domain-specific pre-training

"Dato 5 XI zijn vertrokken
acht [MASK] zijnde 't
Waepen van Brabant, 't
Jagt Susanna....."



Scheepen

	Data	Param	tok/byt
GysBERT	H	110M	7.1B /
GysBERT-v2	HV	110M	8.3B /
BERTje	C	109M	2.4B / 12GB
RobBERT	C	117M	6.6B / 39GB
mBERT (base)	M	179M	/
XLNet (base)	M	279M	/ 2.5TB
GloBERTise	V	117M	/ 6GB

Table 1: Models with types of language present in and volume of pre-training data (H = historical; V = VOC; C = contemporary; M = Multilingual). All info missing in this table could not be found in the relevant papers.

[Verkijk et al., 2025](#)

model	vocabulary (k)	layers	dimension	parameters (M)		
				embeddings	attention	total
gloBERTise	50	12	768	40	85	125
dist-mBERT	120	6	768	92	43	135
mBERT	120	12	768	92	85	177
XLNet-base	250	12	768	192	85	277
XLNet-large	250	24	1024	257	302	558

Table 2: Pretrained-model parameters and sizes

[Arnoult et al., 2025](#)

	GysBERT	GysBERT-v2	XLNet	BERTje	RobBERT	mBERT	GloBERTise	CRF	Lexicon
P-event	0.69	<u>0.71</u>	0.62	0.64	0.63	0.63	0.69	0.57	0.83
R-event	0.40	0.39	0.39	0.36	0.37	<u>0.43</u>	0.50	0.30	0.22
F1-event	0.49	0.48	0.46	0.45	0.45	<u>0.50</u>	0.56	0.39	0.34
mention acc.	0.55	0.53	0.52	0.47	0.50	<u>0.57</u>	0.64	0.40	0.34

Table 2: Scores on detecting events averaged over data folds and seeds

Verkijk et al., 2025

model	micro F1			macro F1			weighted F1		
	epoch	F1	std	epoch	F1	std	epoch	F1	std
gloBERTise	15	81.8	0.3	29	70.7	1.0	12	82.5	0.44
dist-mBERT	40	72.3	0.31	40	65.5	1.33	26	73.4	0.62
mBERT	29	77.6	0.83	29	69.5	3.56	29	78.3	1.07
XLNet-base	40	77.1	0.44	40	69.8	0.78	34	78.4	1.19
XLNet-large	36	81.1	1.37	22	72.1	1.84	39	81.7	1.56

Table 3: Validation scores for models trained on the train-A set, averaged over 3 seeds. The best average metric scores are reported along with the epoch at which they were measured and with standard deviation.

Arnoult et al., 2025

Verkijk, 2025

Why?

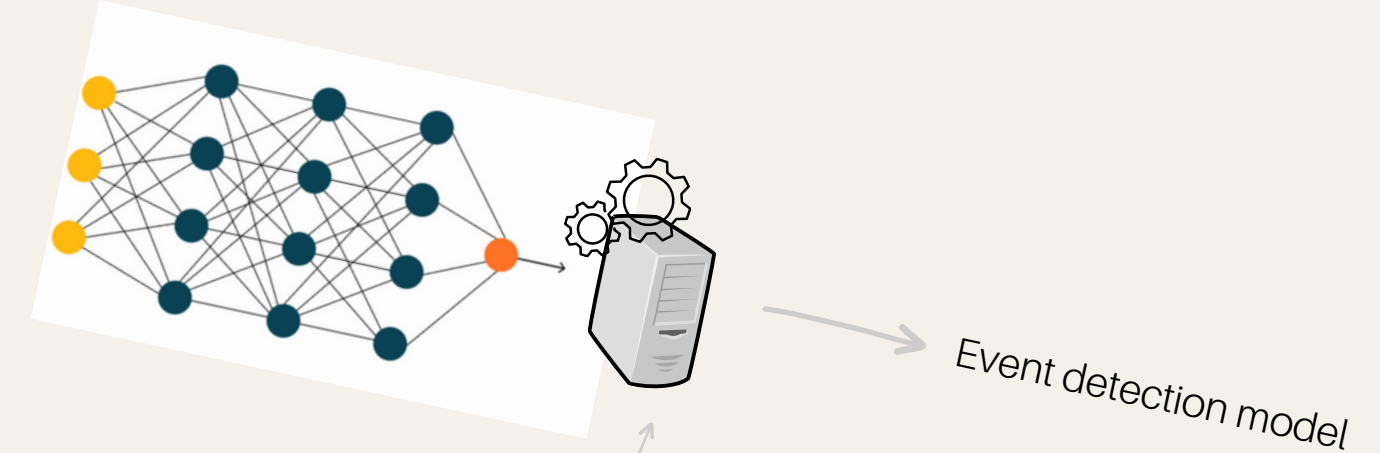
→ *Humanities become a use-case to study the inner workings of LLMs*

What are the implications?

→ *More room for independent creation of LLMs*

→ *Less room for unknown, untraceable biases*

What I would use Generative LMs for



Generative models can be relatively easily be applied to generate new examples of labeled data → **Synthetic data**

*In den jaere des Heeren 1703, werd een scheepslading van pijlaken, zijnde grove wollen stoffen, aan boord van het schip "De Goede Hoop" [**geladen**] en onder groot waakzaamheid en goede orde naer de Oost-Indische havens [**verzonden**] om aldaar ten behoeve van de Compagnie teen gescherpe prijs verhandeld te worden.*

Prompt template with structured data injection

See the following example of an annotated sentence: **{example}**. Create a sentence in Early Modern Dutch in which **{label}** is transported. This is the definition of **{label}**: **{definition}**. Write a sentence that could be found in the Archives of the Dutch East India Company (1602-1799). Make sure to adjust your style to the Dutch East India Company. Put the word that refers to the transportation in square brackets like in the example

Prompt templates with structured data injections

prompt	dataset	example	event
See the following example of an annotated sentence: {example}. Create a sentence in Early Modern Dutch in which {label} is transported. This is the definition of {label}: {definition}. Write a sentence that could be found in the Archives of the Dutch East India Company (1602-1799). Make sure to adjust your style to the Dutch East India Company. Put the word that refers to the transportation in square brackets like in the example.	pp_scheme_goederen.ttl	Yes	Transportation
See the following example of an annotated sentence: {example}. Create a sentence in Early Modern Dutch in which {label} is transported. This is the definition of {label}: {definition}. Write a sentence that could be found in the Archives of the Dutch East India Company (1602-1799). Make sure to adjust your style to the Dutch East India Company. Put the word that refers to the transportation in square brackets like in the example. Do not repeat the definition as given in the prompt, only use it for your own context.	pp_scheme_goederen.ttl	Yes	Transportation
See the following example of an annotated sentence: {example} Create a sentence in Early Modern Dutch in which a translocation is taking place with a {label}. This is the definition of {label}: {definition}. Write a sentence that could be found in the Archives of the Dutch East India Company (1602-1799). Make sure to adjust your style to the Dutch East India Company. Put the word that refers to the translocation in square brackets like in the example. Make sure to vary in syntax.	pp_scheme_vaartuigen.ttl	Yes	Translocation
See the following example of an annotated sentence: {example} Create a sentence in Early Modern Dutch in which a translocation is taking place with a {label}. This is the definition of {label}: {definition}. Write a sentence that could be found in the Archives of the Dutch East India Company (1602-1799). Make sure to adjust your style to the Dutch East India Company. Put the word that refers to the translocation in square brackets like in the example. A translocatiion could for example be an arrival, a departure, a voyage or a transport. Make sure to vary in syntax.	pp_scheme_vaartuigen.ttl	Yes	Translocation
Create a sentence in Early Modern Dutch in which a translocation is taking place with a {label}. This is the definition of {label}: {definition}. Write a sentence that could be found in the Archives of the Dutch East India Company (1602-1799). Make sure to adjust your style to the Dutch East India Company. Put the word that refers to the translocation in square brackets. Do not start your sentence with a time indication.	pp_scheme_vaartuigen.ttl	No	Translocation
See the following example of an annotated sentence: {example}. Create a similar sentence, that differs in syntax and vocabulary, in Early Modern Dutch in which a translocation is taking place. A translocation could for example be an arrival, a departure, a voyage or a transport. Write a sentence that could be found in the Archives of the Dutch East India Company (1602-1799). Make sure to adjust your style to the Dutch East India Company. Put the word that refers to the translocation in square brackets like in the example.	None	Yes	Translocation
Create a sentence in Early Modern Dutch in which {name} is travelling to or from {location}. The way {name} is travelling can be inspired by their profession, which is {function}. Write a sentence that could be found in the Archives of the Dutch East India Company (1602-1799). Make sure to adjust your style to the Dutch East India Company. Put the word that refers to the translocation in square brackets.	naam_functie_plaats.csv	No	Translocation
Create a sentence in Early Modern Dutch in which {name} is travelling to or from {location}. The way {name} is travelling can be inspired by their profession, which is {function}. Write a sentence that could be found in the Archives of the Dutch East India Company (1602-1799). Make sure to adjust your style to the Dutch East India Company. Put the word that refers to the translocation in square brackets. Do not start the sentence with {name}.	naam_functie_plaats.csv	No	Translocation

Let’s say...

- I have 8 different templates
- I make 5 different prompts for each template
- I ask the model to generate 20 examples for each prompt



800 new training examples

There are many chat model options!

- Claude (<https://claude.ai/login?returnTo=%2F%3F>) 😊
- Gemini (<https://gemini.google.com/app>)
- DeepSeek (https://chat.deepseek.com/sign_in)
- Copilot (<https://copilot.microsoft.com>)
- ChatGPT (<https://chatgpt.com>) 😡

GLOBALISE

Website: <https://globalise.huygens.knaw.nl>

Publicaties: <https://globalise.huygens.knaw.nl/output/>

Blog: <https://globalise.huygens.knaw.nl/yes-okay-but-what-were-they-doing/>

Transcriptions viewer: <https://transcriptions.globalise.huygens.knaw.nl>

STELLA VERKIJK
PHD KANDIDAAT

GLOBALISE PROJECT

VRIJE UNIVERSITEIT AMSTERDAM
HUYGENS INSTITUUT
AMSTERDAM AI

S.VERKIJK@VU.NL