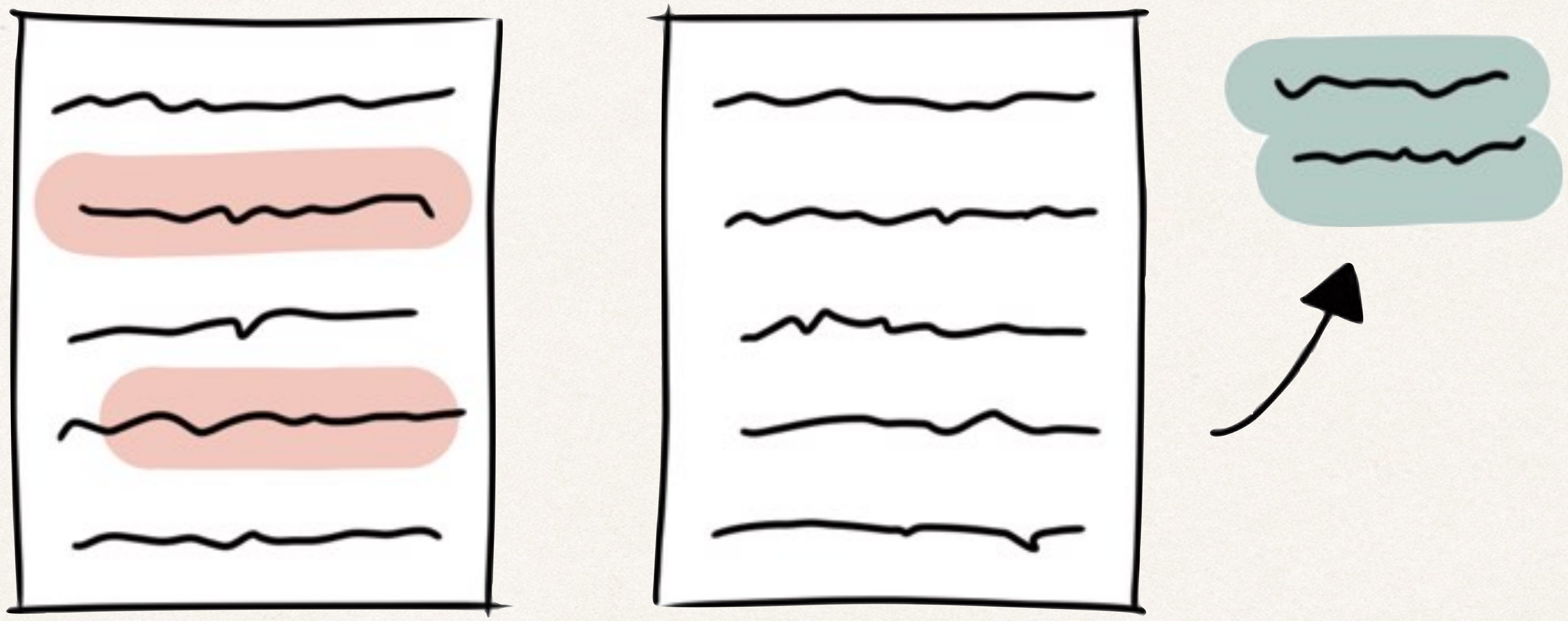
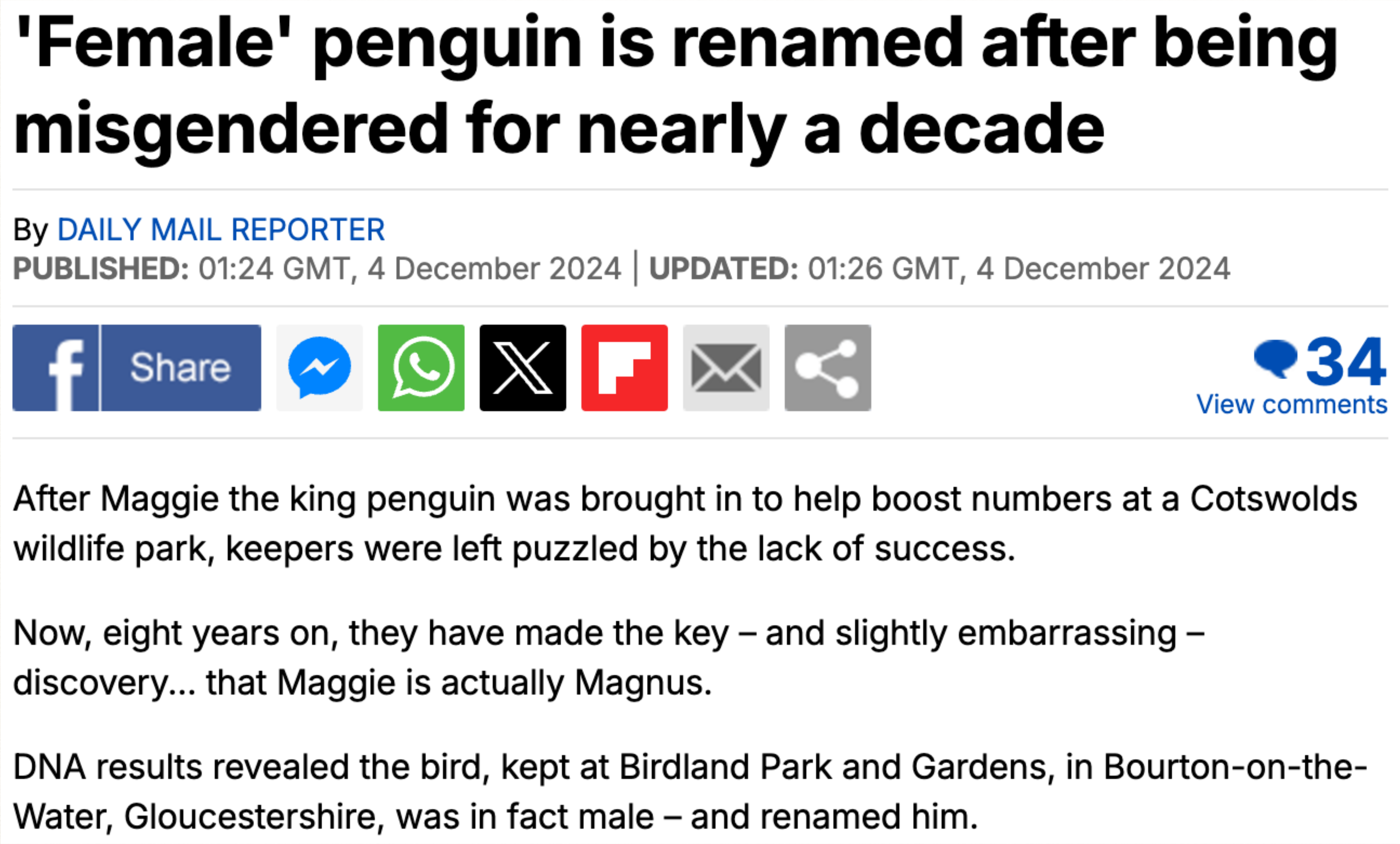
Class 11 - Evaluation

What does a good summary look like?

EXTRACTIVE ABSTRACTIVE





'Female' penguin is renamed after being misgendered for nearly a decade

By DAILY MAIL REPORTER

PUBLISHED: 01:24 GMT, 4 December 2024 | **UPDATED:** 01:26 GMT, 4 December 2024

After Maggie the king penguin was brought in to help boost numbers at a Cotswolds wildlife park, keepers were left puzzled by the lack of success.

Now, eight years on, they have made the key – and slightly embarrassing – discovery... that Maggie is actually Magnus.

DNA results revealed the bird, kept at Birdland Park and Gardens, in Bourton-on-the-Water, Gloucestershire, was in fact male – and renamed him.

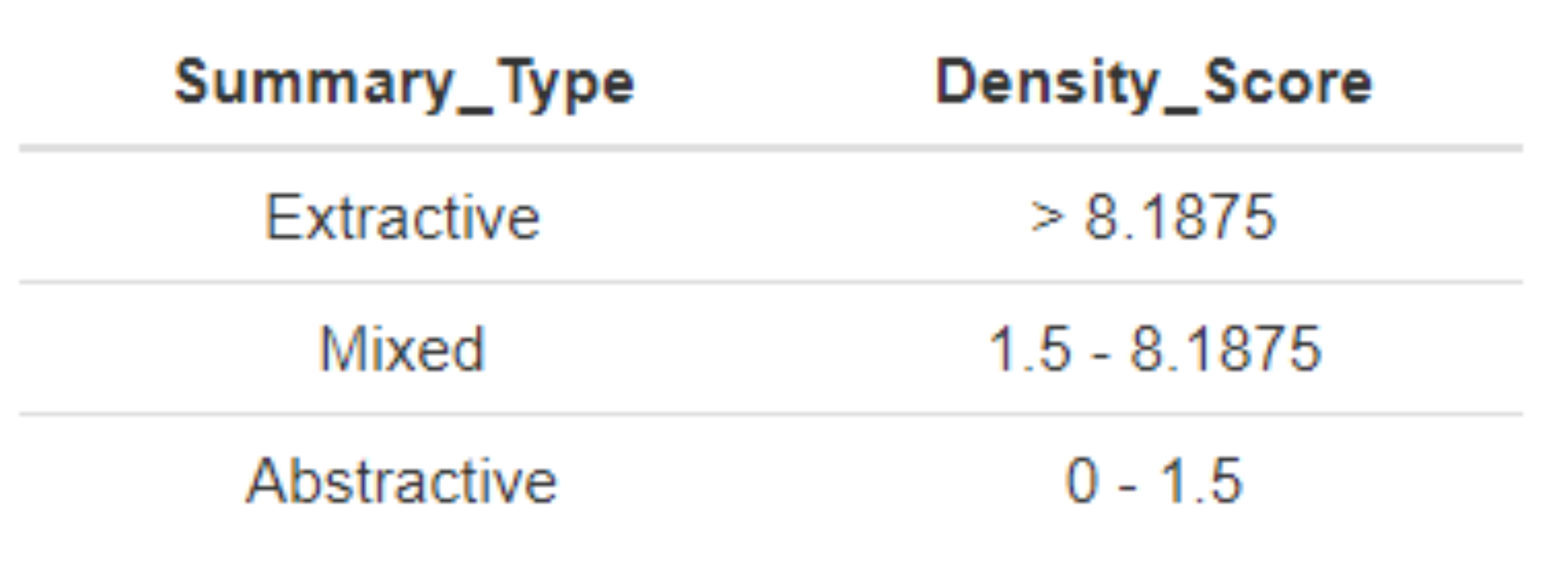
Density

- Fragments
  - Extracted n-gram overlaps
- Density
  - The extent to which the sequences of fragments covers the summary itself
    - Length of fragments is squared resulting in higher values for summaries with long extractive fragments

$$\mathbf{Density}(\boldsymbol{A},\boldsymbol{S}) = \frac{\mathbf{1}}{|\boldsymbol{S}|} \sum_{\boldsymbol{f} \in \boldsymbol{F}(\boldsymbol{A},\boldsymbol{S})} |\boldsymbol{f}|^2$$

Density

- Fragments
  - Extracted n-gram overlaps
- Density
  - The extent to which the sequences of fragments covers the summary itself
    - Length of fragments is squared resulting in higher values for summaries with long extractive fragments



Summary_Type	Density_Score
Extractive	> 8.1875
Mixed	1.5 - 8.1875
Abstractive	0 - 1.5

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

- Amount of lexical overlap between the generated and the reference summary
  - R-N: Overlap of N-grams
    - R-1: Overlap of unigrams
    - R-2: Overlap of bigrams
  - R-L: Longest Common Sequence (LCS)
    - In-sequence matches

- **Recall**

$$\frac{\textbf{number of common n} - \textbf{grams}}{\textbf{number of n} - \textbf{grams in the reference summary}}$$

  - How much of the reference summary does the generated summary cover?

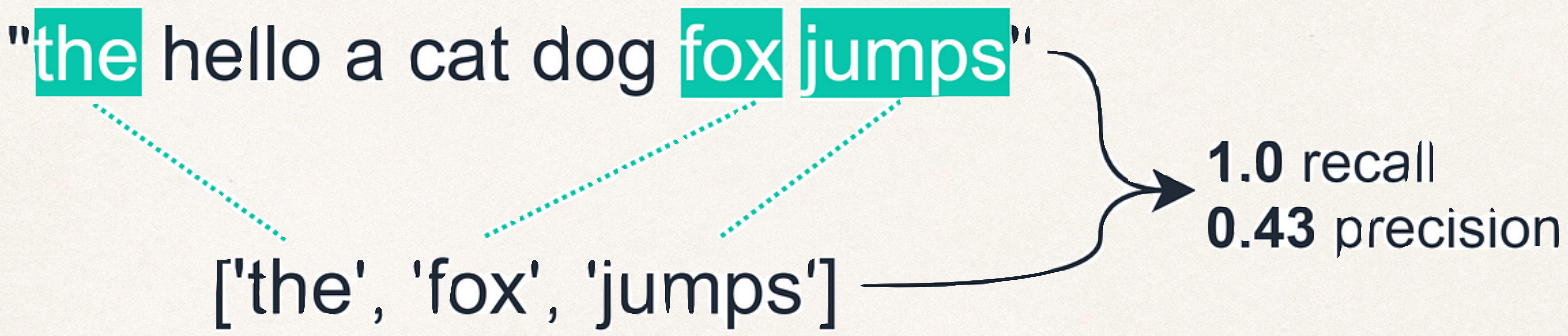
- **Precision**

$$\frac{\textbf{number of common n} - \textbf{grams}}{\textbf{number of n} - \textbf{grams in the generated summary}}$$

  - How many of the words in the generated summary are relevant?

- **F1**

$$2\,\frac{\textbf{precision} \cdot \textbf{recall}}{\textbf{precision} + \textbf{recall}}$$



"the hello a cat dog fox jumps"

['the', 'fox', 'jumps']

1.0 recall
0.43 precision

$$2 * \frac{0.43 * 1.0}{0.43 + 1.0} = 0.6$$

60% f1 score

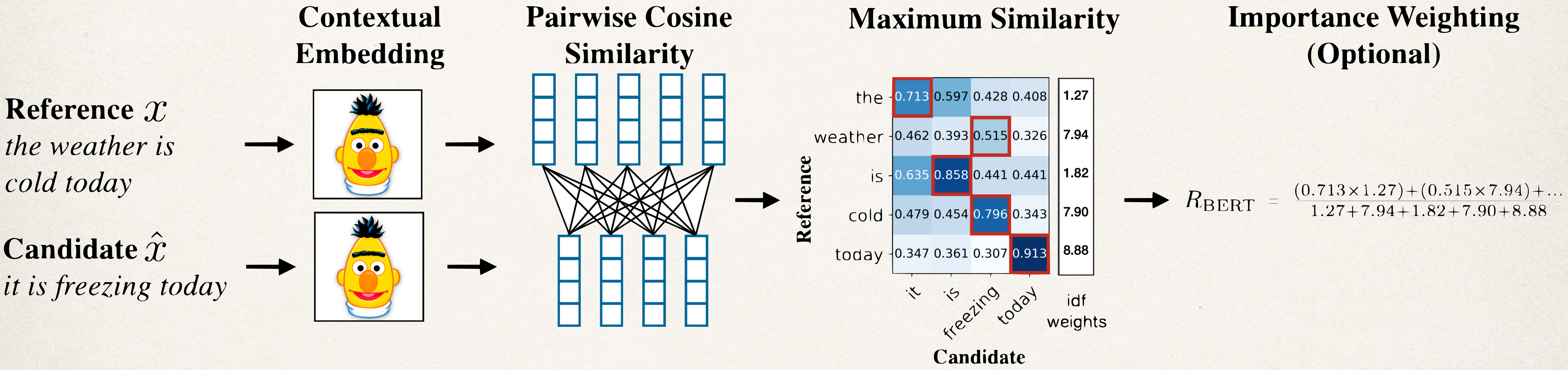
The quick brown fox jumped over the lazy dog.

The fast wood-coloured fox hopped over the lethargic dog.

LOWER ROUGE L F score: 55

Semantically Accurate

BERTScore



Reference x
the weather is cold today

Candidate $\hat{x}$
it is freezing today

Contextual Embedding

Pairwise Cosine Similarity

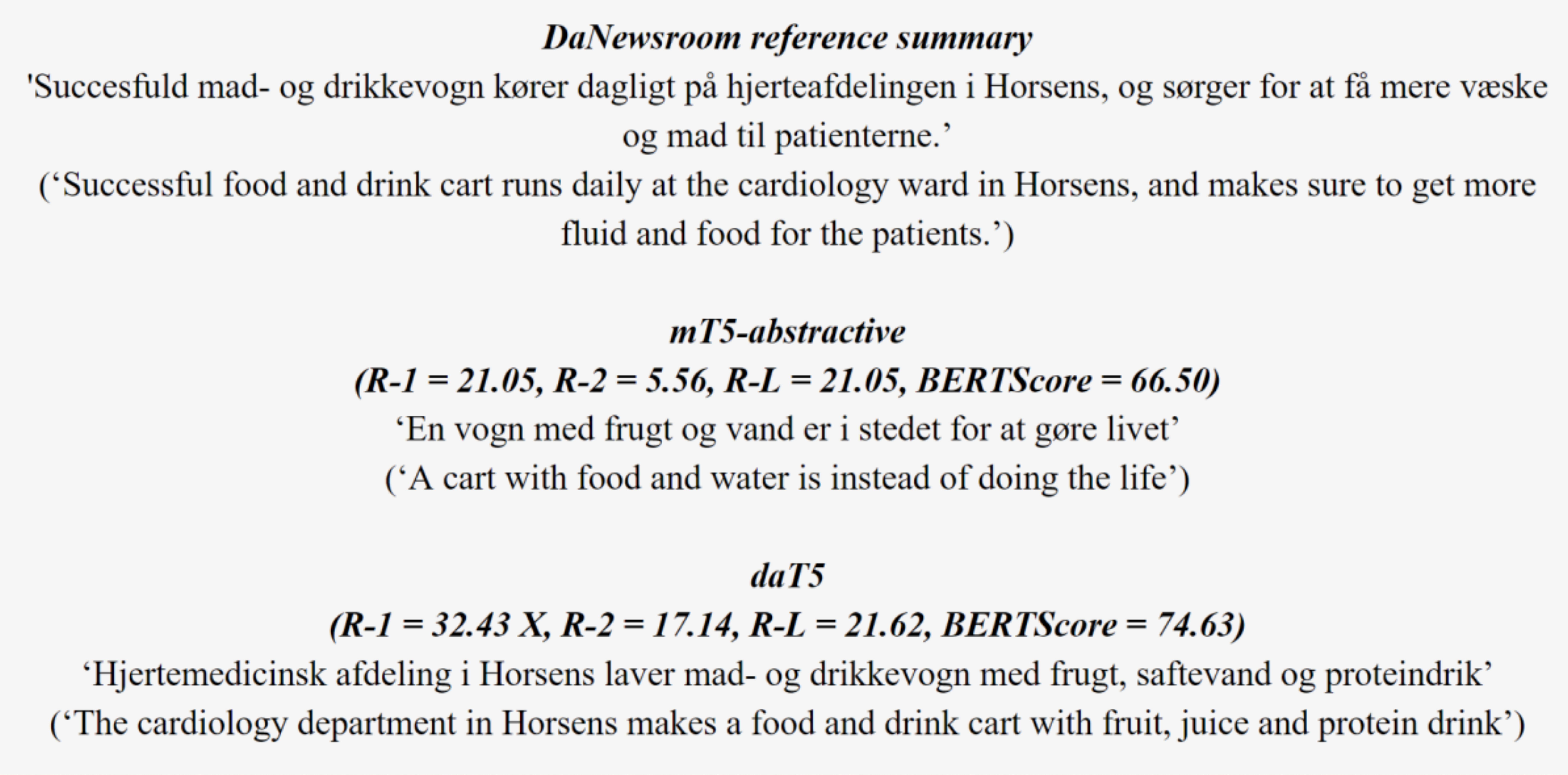
Maximum Similarity

Reference	it	is	freezing	today	idf weights
the	0.713	0.597	0.428	0.408	1.27
weather	0.462	0.393	0.515	0.326	7.94
is	0.635	0.858	0.441	0.441	1.82
cold	0.479	0.454	0.796	0.343	7.90
today	0.347	0.361	0.307	0.913	8.88

Candidate

Importance Weighting (Optional)

$$R_{\mathrm{BERT}} = \frac{(0.713\times1.27)+(0.515\times7.94)+\ldots}{1.27+7.94+1.82+7.90+8.88}$$



DaNewsroom reference summary

'Succesfuld mad- og drikkevogn kører dagligt på hjerteafdelingen i Horsens, og sørger for at få mere væske og mad til patienterne.’

(‘Successful food and drink cart runs daily at the cardiology ward in Horsens, and makes sure to get more fluid and food for the patients.’)

mT5-abstractive

(R-1 = 21.05, R-2 = 5.56, R-L = 21.05, BERTScore = 66.50)

‘En vogn med frugt og vand er i stedet for at gøre livet’

(‘A cart with food and water is instead of doing the life’)

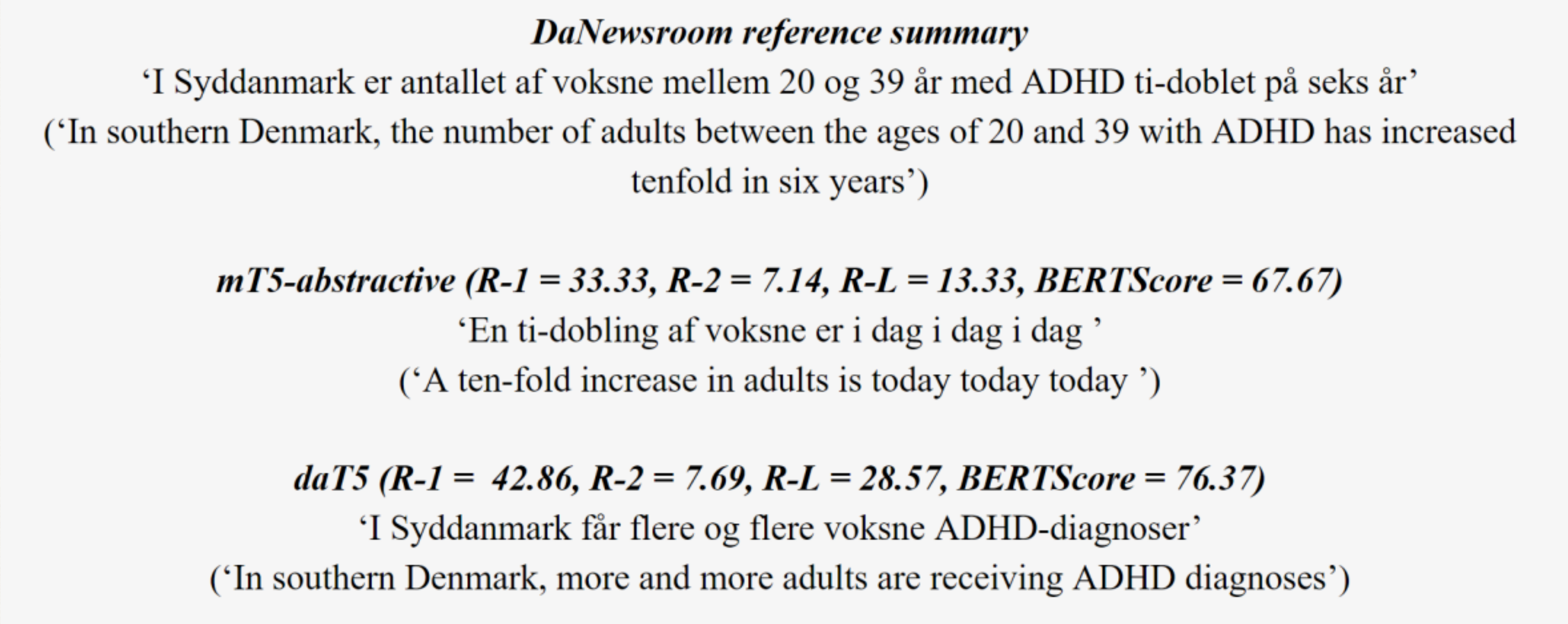
daT5

(R-1 = 32.43 X, R-2 = 17.14, R-L = 21.62, BERTScore = 74.63)

‘Hjertemedicinsk afdeling i Horsens laver mad- og drikkevogn med frugt, saftevand og proteindrik’

(‘The cardiology department in Horsens makes a food and drink cart with fruit, juice and protein drink’)

Kolding, S. *et al*. (2023) ‘DanSumT5: Automatic Abstractive Summarization for Danish’, in *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*. *NoDaLiDa 2023*, Tórshavn, Faroe Islands: University of Tartu Library, pp. 248–264. Available at: https://aclanthology.org/2023.nodalida-1.25 (Accessed: 19 July 2023).



DaNewsroom reference summary

‘I Syddanmark er antallet af voksne mellem 20 og 39 år med ADHD ti-doblet på seks år’

(‘In southern Denmark, the number of adults between the ages of 20 and 39 with ADHD has increased tenfold in six years’)

mT5-abstractive (R-1 = 33.33, R-2 = 7.14, R-L = 13.33, BERTScore = 67.67)

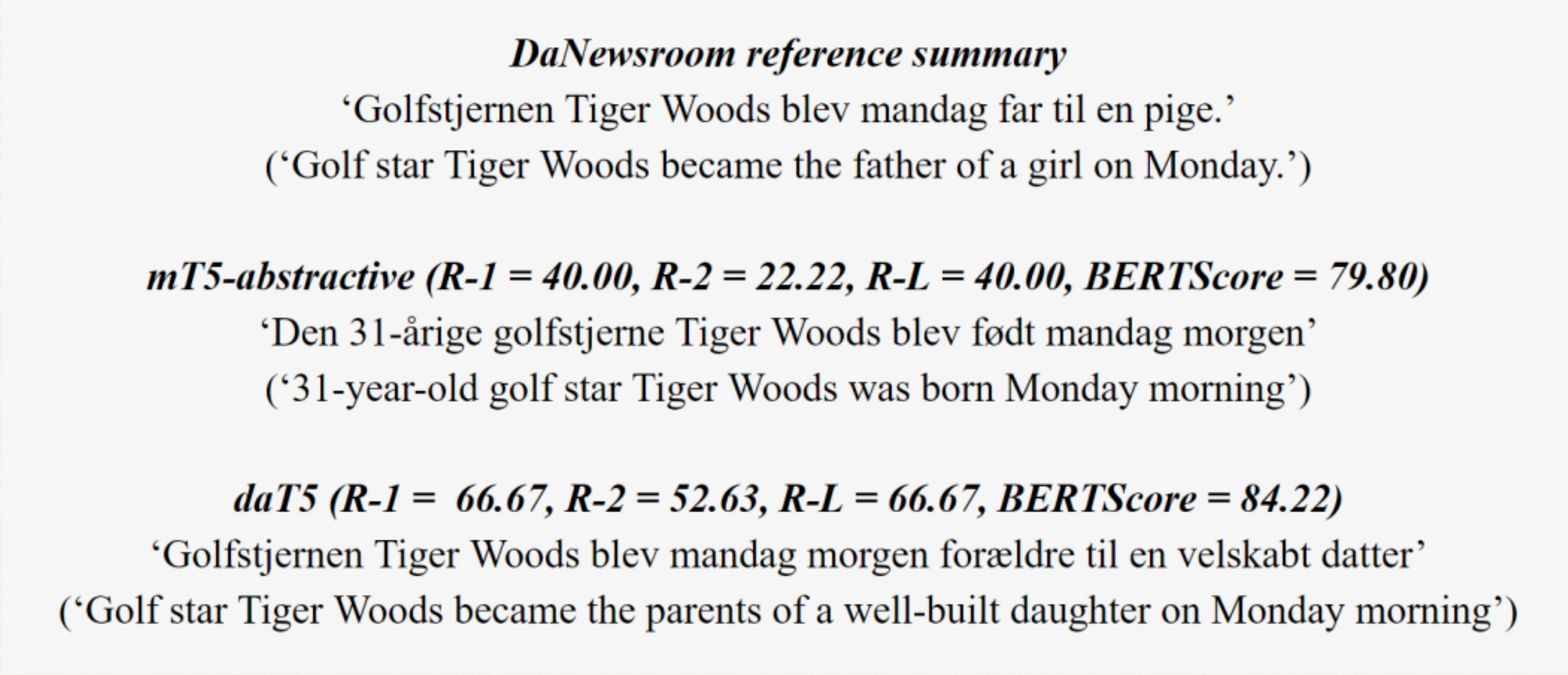
‘En ti-dobling af voksne er i dag i dag i dag ’

(‘A ten-fold increase in adults is today today today ’)

daT5 (R-1 = 42.86, R-2 = 7.69, R-L = 28.57, BERTScore = 76.37)

‘I Syddanmark får flere og flere voksne ADHD-diagnoser’

(‘In southern Denmark, more and more adults are receiving ADHD diagnoses’)



DaNewsroom reference summary

'Golfstjernen Tiger Woods blev mandag far til en pige.'

('Golf star Tiger Woods became the father of a girl on Monday.')

mT5-abstractive (R-1 = 40.00, R-2 = 22.22, R-L = 40.00, BERTScore = 79.80)

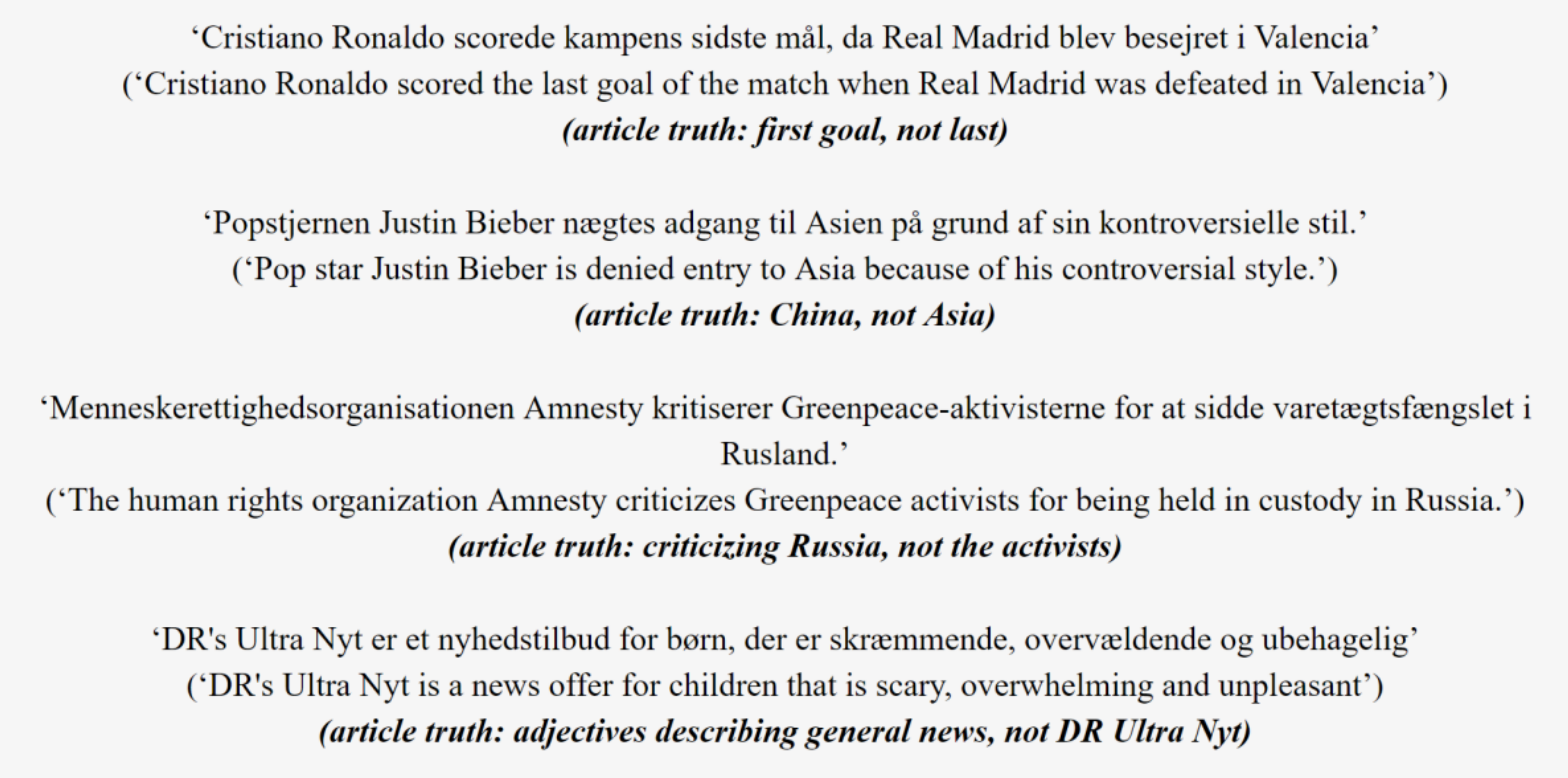
'Den 31-årige golfstjerne Tiger Woods blev født mandag morgen'

('31-year-old golf star Tiger Woods was born Monday morning')

daT5 (R-1 = 66.67, R-2 = 52.63, R-L = 66.67, BERTScore = 84.22)

'Golfstjernen Tiger Woods blev mandag morgen forældre til en velskabt datter'

('Golf star Tiger Woods became the parents of a well-built daughter on Monday morning')



‘Cristiano Ronaldo scorede kampens sidste mål, da Real Madrid blev besejret i Valencia’
(‘Cristiano Ronaldo scored the last goal of the match when Real Madrid was defeated in Valencia’)
(article truth: first goal, not last)

‘Popstjernen Justin Bieber nægtes adgang til Asien på grund af sin kontroversielle stil.’
(‘Pop star Justin Bieber is denied entry to Asia because of his controversial style.’)
(article truth: China, not Asia)

‘Menneskerettighedsorganisationen Amnesty kritiserer Greenpeace-aktivisterne for at sidde varetægtsfængslet i Rusland.’
(‘The human rights organization Amnesty criticizes Greenpeace activists for being held in custody in Russia.’)
(article truth: criticizing Russia, not the activists)

‘DR's Ultra Nyt er et nyhedstilbud for børn, der er skræmmende, overvældende og ubehagelig’
(‘DR's Ultra Nyt is a news offer for children that is scary, overwhelming and unpleasant’)
(article truth: adjectives describing general news, not DR Ultra Nyt)

“A false response by GPT-4 is sometimes referred to as a ‘hallucination,’.”

Indeed, it has become standard in AI to refer to a response that is not justified by the training data as a hallucination. We find this terminology to be problematic for the following 2 reasons:

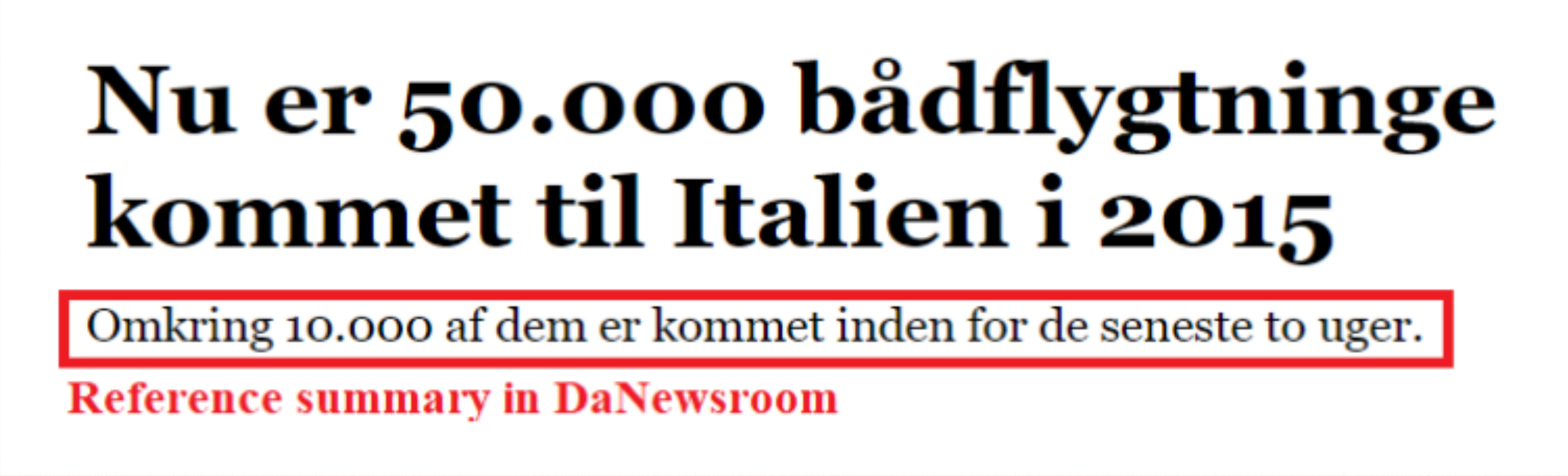
1. **It is an imprecise metaphor**. Hallucination is a medical term used to describe a sensory perception occurring in the absence of an external stimulus. AI models do not have sensory perceptions as such—and when they make errors, it does not occur in the absence of external stimulus. Rather, the data on which AI models are trained can (metaphorically) be considered as external stimuli—as can the prompts eliciting the (occasionally false) responses.

2. More importantly, **it is a highly stigmatizing metaphor**. Hallucinations can accompany many, primarily neurological or mental, illnesses, and represent a hallmark symptom of schizophrenia. Individuals with schizophrenia experience stigma from many sides of society, with inappropriate metaphorical use of the word schizophrenia (with negative connotation) being one of the sources. Metaphorical use of hallucination (also with a clear negative connotation) in AI—a field with clear links to both medicine in general and psychiatry specifically—is, therefore, very unfortunate. Notably, this is occurring at a time when reducing stigma is a top priority for psychiatry at large—in order to improve the lives of those living with mental illness.

– Østergaard, S.D. and Nielbo, K.L. (2023) ‘False Responses From Artificial Intelligence Models Are Not Hallucinations’, Schizophrenia Bulletin, 49(5), pp. 1105–1107. Available at: https://doi.org/10.1093/schbul/sbad068.

Limitations

- The closer to the “gold standard” the better?



Nu er 50.000 bådflygtninge kommet til Italien i 2015

Omkring 10.000 af dem er kommet inden for de seneste to uger.

Reference summary in DaNewsroom

Reference-free metrics

- Generated summary only
  - Evaluate linguistic quality or information content of generated summary, e.g., grammaticality, fluency, etc.
  - LLM judges
- Text-generated summary pairs
  - Use the text to compare with the generated summary, e.g., compression ratio, density, redundancy, etc.
  - BLANC: performance gained by a pre-trained language model with access to a summary while carrying out a language understanding task on the text