

# Assessing scientific (dis)agreement: a cross-disciplinary approach to rhetorical citation classification using language model transfer





Anne-Sophie Foussat, Vincent Guigue, Nicolas Sauvion, Robert Bossy, Claire Nédellec



INRAE

université  
PARIS-SACLAY

# Introduction : Why classify citation rhetorical functions ?

- Profusion of scientific information 
- An evolving body of knowledge  
→ built on debates and consensus 
- Citations of a discovery = an indicator of its reliability 
- Citation frequency *is not* a relevant indicator of scientific agreement
- The reason for the citation, its **role in the argumentation** is critical,  
= *rhetorical function* 

# Introduction : Why classify citation rhetorical functions ?

“... it is the preferred host plant of *C. pruni* (Lauterer 1999) and accordingly, we found *C. pruni* on every *P. spinosa* tested sometimes at high population densities. This is supported by data from Carraro et al. (2002), Yvon et al. (2004) or Maier et al. (2013)...”

Jarausch et al. (2019)

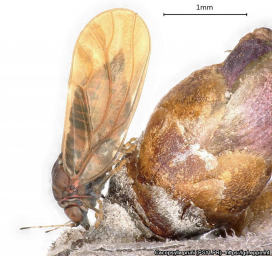
→ citation passage :  
*context* - citance (reference) - *context*

→ class = support

P. Spinosa



C. pruni



Source EPPO

# Introduction: Ecological context & scientific challenges

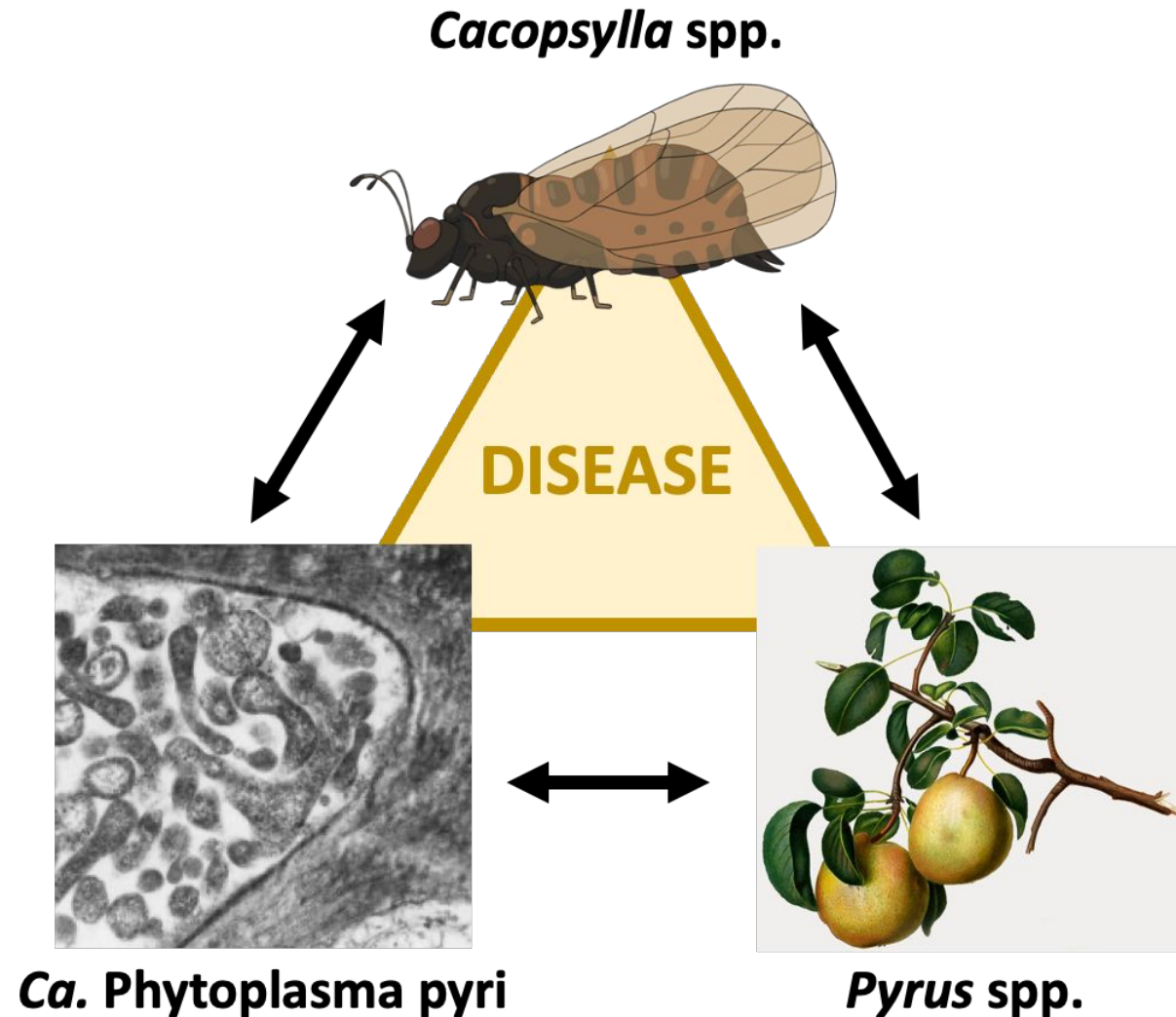
- **Reliability** of findings in complex biological interactions.
- **Fast knowledge evolution**
- **Critical for**
  - **Early detection**
  - **Management** of human, animal, and plant health

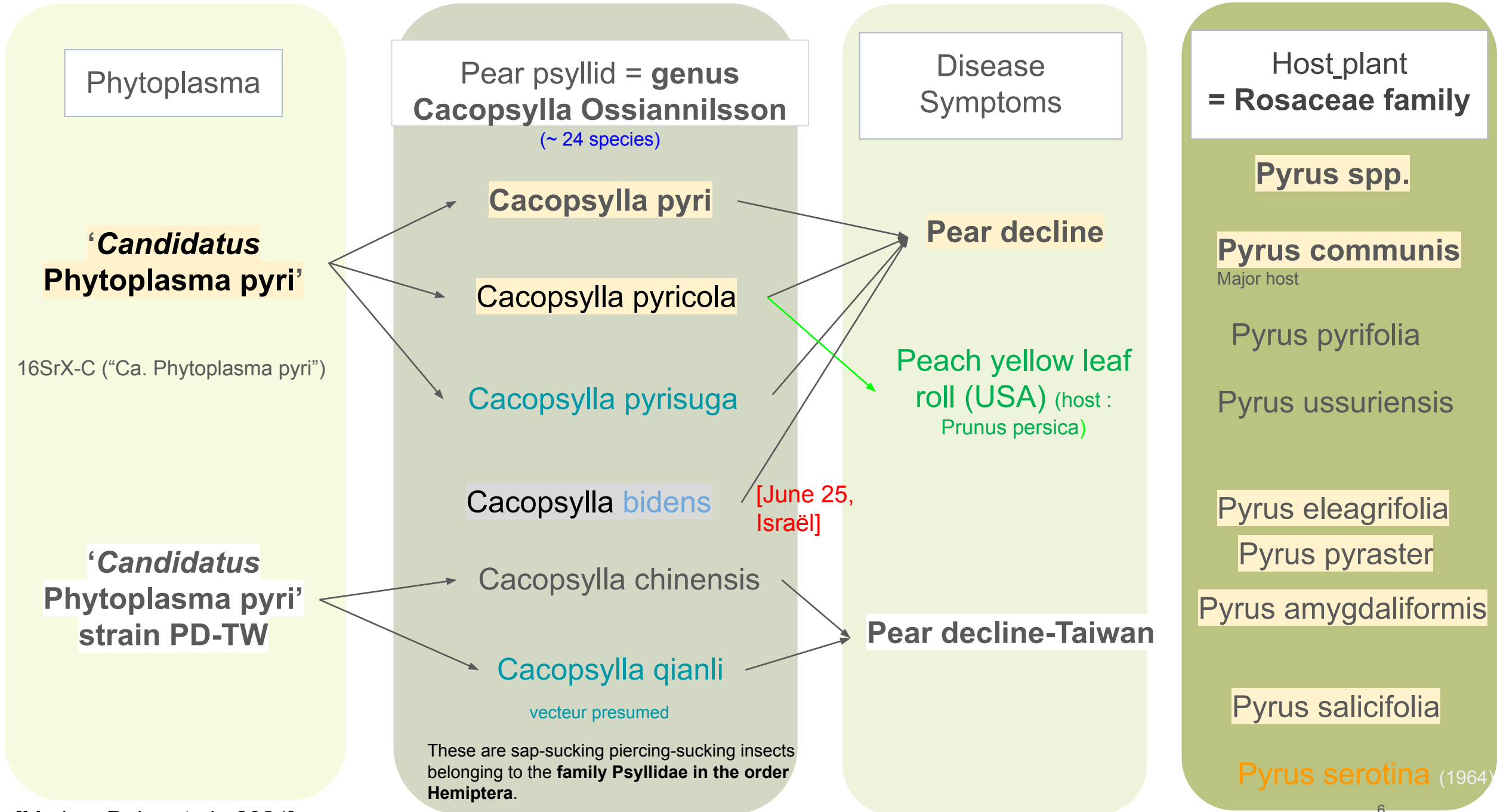


## **Challenges in establishing a classification framework**

- Rely on **subjective** citation analysis of the citing author intend
- Need for the semantic citation **context** to classify
- **Disciplinary differences** in citation practices

# *Pear decline* example: an evolving body of knowledge involving complex interactions





[Myriam Dulor et al., 2024]

# Introduction : Rhetorical citation classification transfer

- Previous work on citation classification in NLP:
  - Rule-based model (Garzone & Mercer, 2000)
  - Supervised machine learning classification based on linguistic features (Teufel et al., 2006)
  - Fine-tuning of SciBERT (Zhang et al., 2022; Jiang & Chen, 2023)
  - GPT-3 prompting (Kunnath et al., 2023)

Available annotated datasets in NLP: Jiang2021 (on demand)

- Our corpus on *Pear Decline* disease  
(interactions vectors - plants - pathogens)



# Research questions



- 1) Is the **rhetorical citation function typology** developed in computational linguistics adequate and sufficient **for analyzing citation discourse** in studies of **biological interactions**?



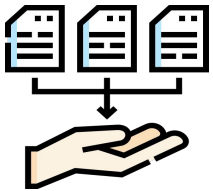
- 2) Can **language models** trained on computational linguistics datasets be effectively transferred to **classify the rhetorical function** of citations in ecological literature?



# Method: Datasets

**Corpus PD100cit** : 100 in-context citations manually annotated according to our guidelines

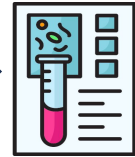
Document collection on  
*Pear Decline*



Filter on research  
articles



Document  
conversion into XML



Article “cleaning” to  
target vectors



Citation passage  
extraction using regular  
expressions



Rhetorical and biological  
class annotation



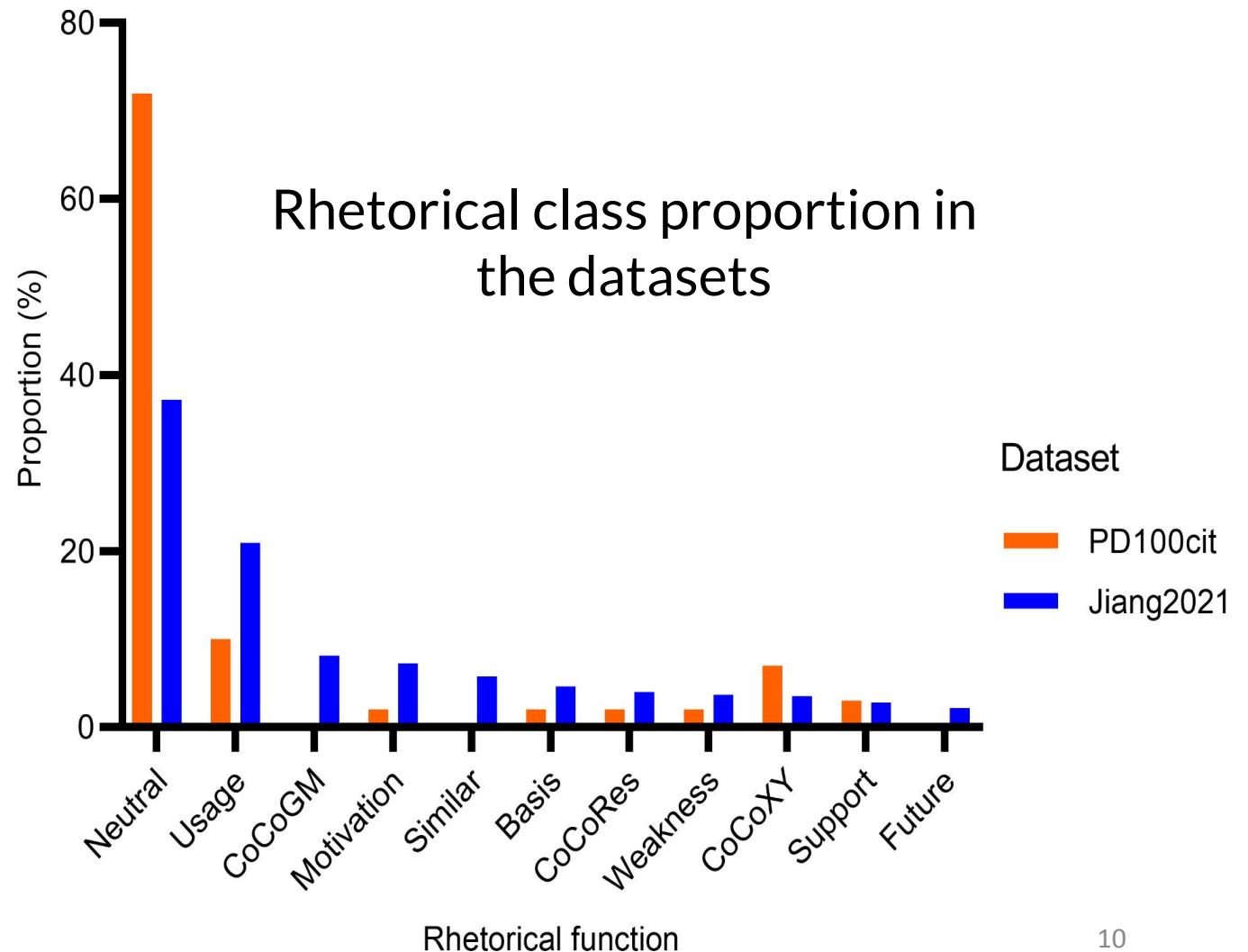
Random selection  
of 100 citations

Myriam Dolor internship

[https://github.com/AnneSophie148/PhD/tree/main/citation\\_classification](https://github.com/AnneSophie148/PhD/tree/main/citation_classification)

# Method: the rhetorical classes

- Typology from previous work by Jiang and Chen (2023), based on Teufel (2009)
- Includes classes such as *Comparison, Contrast* (CoCo) and *Weakness*  
→ useful for analyzing reliability
- Designed for the computational linguistics domain



# Method: the rhetorical classes

Neutral	background information without stance
Usage	tools, data, or techniques
CoCoGM	compares or contrasts goals or methods ( <i>citing work vs cited work</i> )
Motivation	justify the current research
Similar	highlights a similarity
Basis	intellectual foundations of the current work
CoCoRes	compares or contrasts results ( <i>citing work vs cited work</i> )
Weakness	criticizes the cited work
CoCoXY	compares or contrast two cited articles
Support	supports the cited work
Future	mentions future work

# Reminder: citance, reference, rhetorical class, methods

Jarausch et al. (2019)

*“... it is the preferred host plant of *C. pruni* (Lauterer 1999) and accordingly, we found *C. pruni* on every *P. spinosa* tested sometimes at high population densities. This is supported by data from Carraro et al. (2002), Yvon et al. (2004) or Maier et al. (2013)...”*

→ citation passage :

*context* - citance (reference) - *context*

→ rhetorical class = support

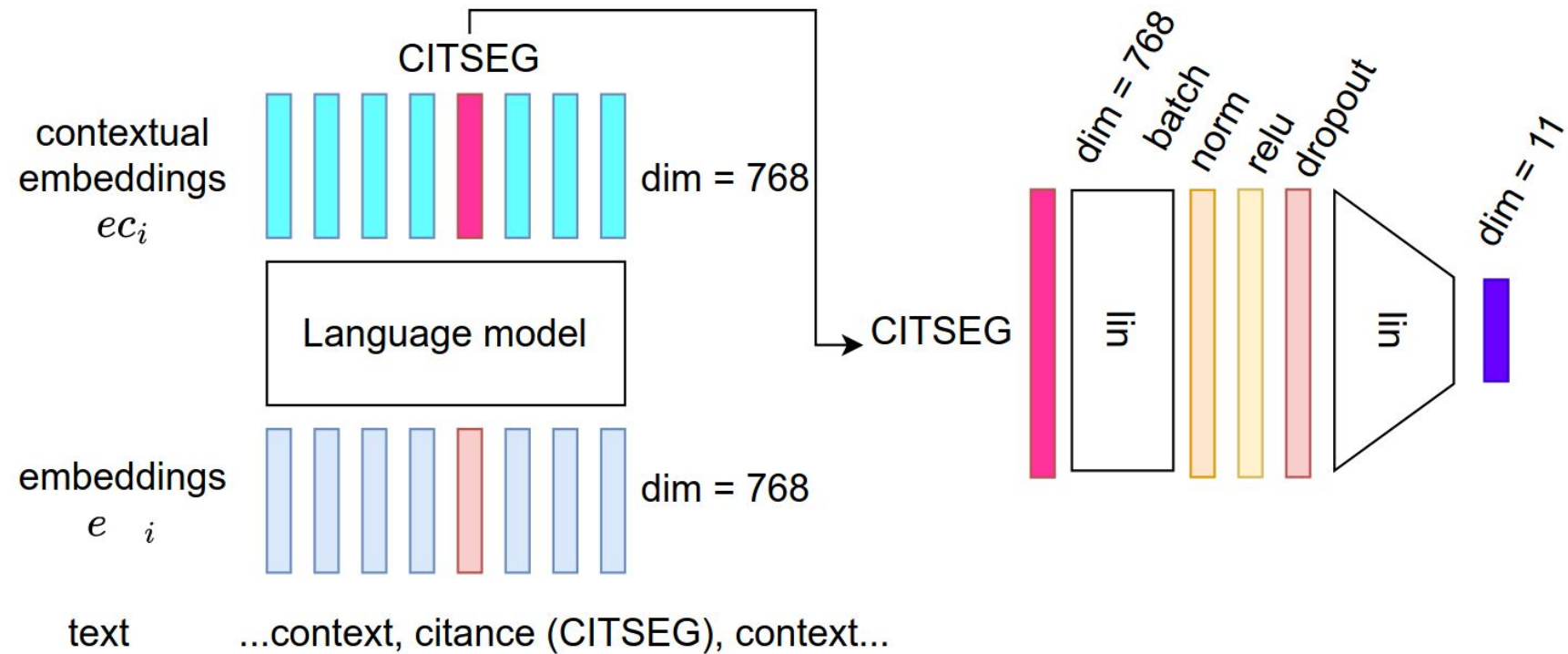
## Classification problem

- Predict the rhetorical class of the citation given the citation passage

## Two methods

- Transfer learning with language models
- Hard prompting a generative model

# Method : Transfer of LM fine-tuned in NLP to biology



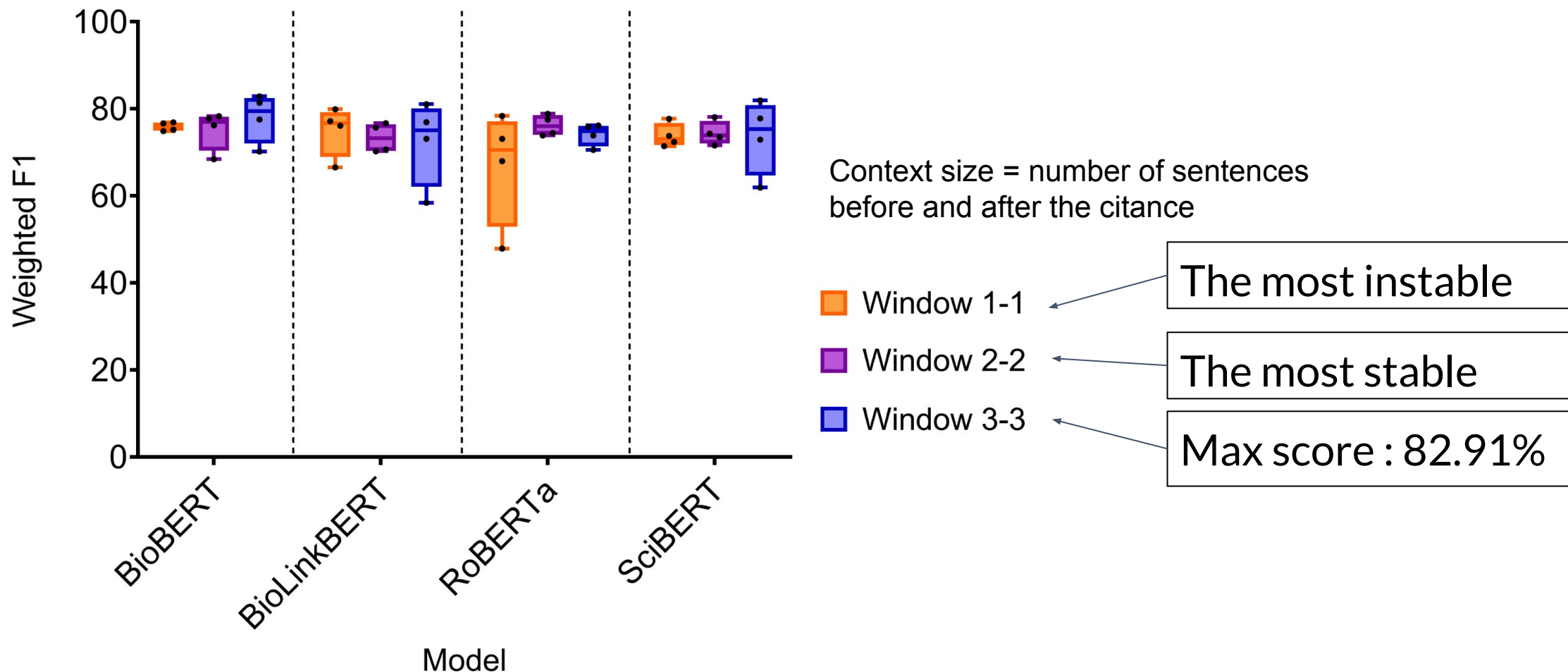
**Our architecture**

# Method: Transfer of LM fine-tuned in NLP to biology

- Configuration
  - Train : Jiang2021 (3356 citations)  
80% train ; 20% validation
  - Test : PD100cit (100 citations)
- Language models :
  - SciBERT
  - RoBERTa
  - BioBERT
  - BioLinkBERT
- Context (number of sentences):
  - 3-3
  - 2-2
  - 1-1

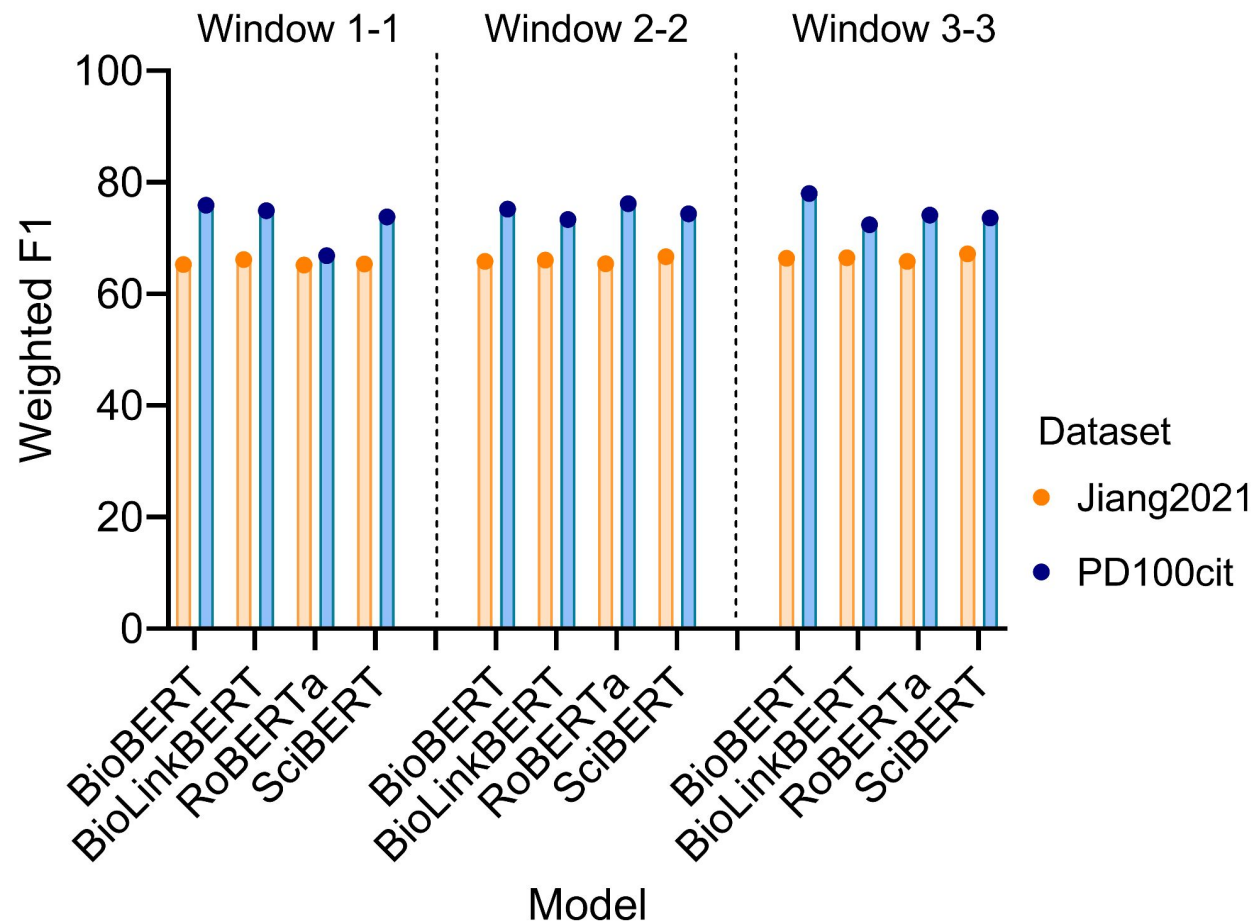


# Results : Transfer of LM fine-tuned in NLP to biology



Weighted F1 scores on PD100cit by context window size and model across four seeds

# Results : Transfer of LM fine-tuned in NLP to biology



Best average F1 weighted :

- PD100cit : 78.08%
- Jiang2021 : 67.18%

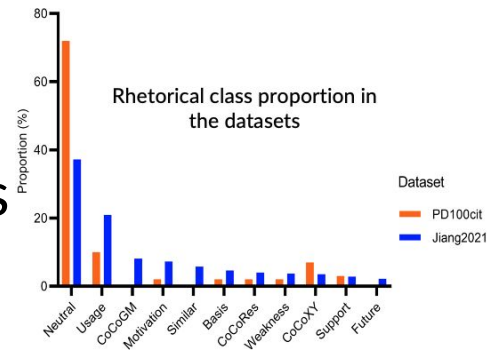
How to explain those differences ?

Average weighted F1 scores by model and context window on PD100cit and Jiang2021

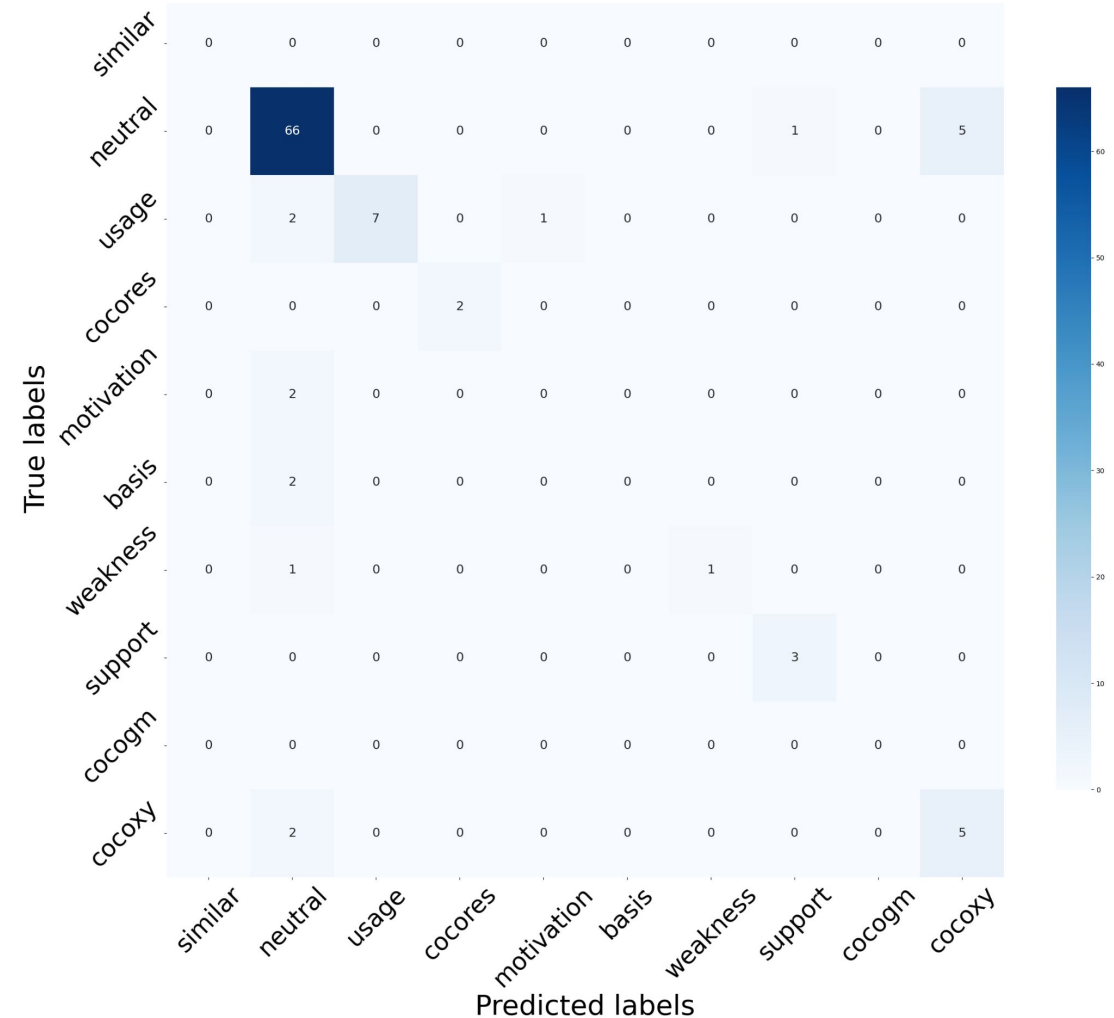


# Results : Transfer of LM fine-tuned in NLP to biology

- Unbalanced classes



- Linguistic cues for some classes (“usage”, “support”) easier to predict
- Ambiguities between “neutral” and “weakness” (direct criticism or reference to criticism)



# Method: Prompting GPT-4 for the rhetorical citation classification

## Task

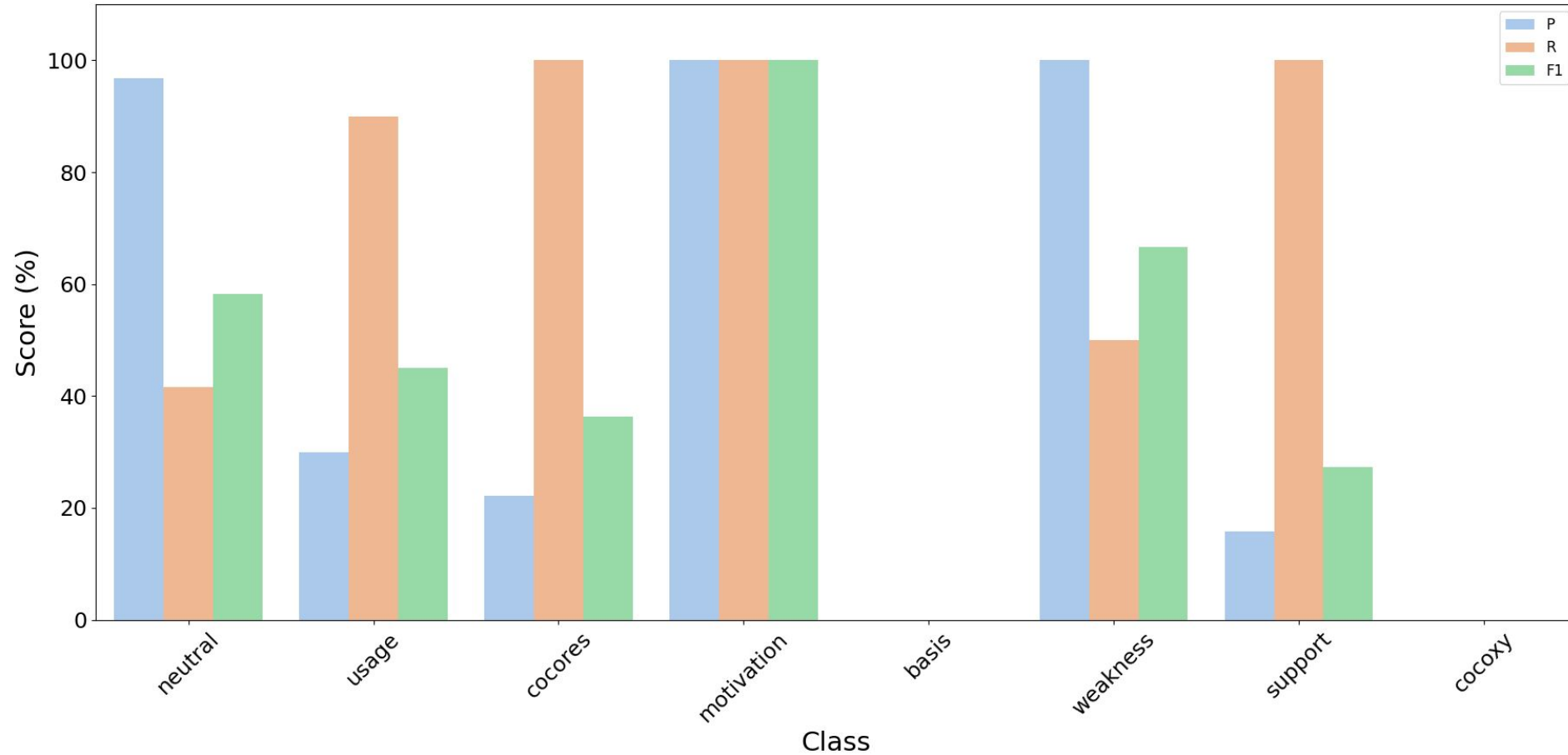
*“you are a classifier that assigns the citations in a scientific article passage to a rhetorical class. The rhetorical class represents the argumentation role played by the cited work in the context of the citing text. Each citation is denoted using the placeholder (CITSEG). Your output must be the single most appropriate class label from the predefined list below and nothing else.”*

- Class definition  
[..]
- One example for each class
- [..]

# Results : Prompting GPT-4 for the rhetorical citation classification

- Little variation across the different context windows

Highest score with 1-1 window: 51.32%



P, R and F1 per class prompting GPT-4 on window 1-1

# Conclusion

- PD100cit dataset, annotation guidelines, and code on GitHub
- Best scores obtained with the **fine-tuned BERT-like models** and 3-3 window size despite instabilities, with little variation across models
- **Fine-tuned LMs outperform GPT-4** (without fine-tuning)
- **Good transfer** results in ecology of fine-tuned LMs trained in NLP
- **Rhetorical classes to be revised** to focus on reliability

# Future work

- Confirm the results on a larger dataset
- A given paper may be cited for different (sub)findings) of irrelevant reasons. Also assign a biological class to each citation it.  
Ex. vector-location<> vector-transmit-pathogen-location
- Estimate reliability through the analysis of the temporal dynamics of citations with their rhetorical class
- Leverage article metadata to assess quality
- Deepen classification with LLMs: use open models (e.g. Qwen, Deepseek, Llama)

Thank you for your attention!

Questions?

# Index

# Définition et exemples des classes

Classe	Définition	Exemple
<b>Basis</b>	La citation reconnaît les fondements intellectuels du travail actuel.	Based on the information obtained in this and previous studies (14,15; A. H. Purcell, unpublished data), pear growers are now aware that pears are the primary reservoir for PYLR in northern California.
<b>CoCoGM</b>	Contraste ou comparaison des objectifs ou méthodes entre l'article citant et l'article cité.	The PCR amplification conditions were the same as proposed by Ghanim et al. [43]
<b>CoCoRes</b>	Comparaison ou contraste des résultats entre l'article citant et l'article cité.	Comparing the detection of grapevine yellows phytoplasma in planthoppers, only 66% of the PCR positives were also positive by enzyme-linked immunosorbent assay (35).
<b>CoCoXY</b>	Comparaison ou contraste entre deux articles cités.	Ullman and Mclean (1986) and Garzo et al. (2012) also observed the same number of teeth on the mandibles of the psyllids <i>C. pyricola</i> and <i>Diaphorina citri</i> respectively, whereas Pollard (1970) found 8 teeth in adults (7 teeth in nymphs) on the mandibles of <i>C. mali</i> .
<b>Future</b>	Mentionne des perspectives ou des travaux futurs.	(Non observée dans PD100cit)
<b>Motivation</b>	Justifie la recherche actuelle par des résultats antérieurs prometteurs ou des enjeux scientifiques.	Regarding the capability of <i>B. nigricornis</i> to transmit CaLsol, previous field work has shown that <i>B. nigricornis</i> can become naturally infected with CaLsol haplotype E (Teresani et al. 2014; 2015), so further research to assess the vector efficiency of this psyllid species was needed.
<b>Neutral</b>	Fournit des informations de fond sans positionnement.	The number of protrusions varies among different species, which may be related to the hardness of the leaves of the host plant (Forbes, 1977; Rosell et al., 1995; Zhao et al.; Garzo et al., 2012).
<b>Similar</b>	Met en évidence une similarité.	(Non observée dans PD100cit)
<b>Support</b>	Apporte un soutien ou une confirmation.	Sequence similarity values within taxa and divergence between taxa largely confirm the results of previous work (Seemüller et al., 1994; Seemüller et al., 1998b).
<b>Usage</b>	Mentionne une méthode, un outil ou des données utilisés dans le travail du citant.	In direct PCR or the first amplification of semi-nested PCR, the universal phytoplasma primers P1/P7 (Deng and Hiruki, 1991; Schneider et al., 1995) were used.
<b>Weakness</b>	Critique ou souligne les limites du travail cité.	To compare our results with those papers, we preferred to use the same methodology despite some potential limitations, such as a poor fit.

TABLE 1 – Définitions et exemples des classes rhétoriques



# Résultats : prompt de GPT-4

Label	Citation	Context 3-3			Context 2-2			Context 1-1		
	nb	P	R	F1	P	R	F1	P	R	F1
Neutral	72	100.00	37.50	54.55	96.43	37.50	54.00	96.77	41.67	58.25
Usage	10	28.13	90.00	42.86	31.03	90.00	46.15	30.00	90.00	45.00
CoCores	2	18.18	100.00	30.77	12.50	50.00	20.00	22.22	100.00	36.36
Motivation	2	100.00	100.00	100.00	66.67	100.00	80.00	100.00	100.00	100.00
Basis	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Weakness	2	50.00	50.00	50.00	50.00	50.00	50.00	100.00	50.00	66.67
Support	3	17.65	100.00	30.00	15.79	100.00	27.27	15.79	100.00	27.27
CoCoXY	7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4: Precision (P), Recall (R), and F1-score (F1) per label and context window size for GPT-4 prompt classification

# F1 moyens et écart-type sur la validation

Modele	Contexte 1-1		Contexte 2-2		Contexte 3-3	
	F1_macro	F1_pondérée	F1_macro	F1_pondérée	F1_macro	F1_pondérée
BioBERT	50.69 (3.55)	65.31 (2.34)	51.92 (3.91)	65.87 (2.74)	51.64 (3.21)	66.43 (2.00)
BioLinkBERT	51.57 (3.11)	66.17 (2.13)	51.66 (2.39)	66.10 (2.17)	<b>52.34</b> (2.38)	66.52 (2.25)
RoBERTa	50.45 (3.85)	65.21 (2.66)	51.12 (1.82)	65.41 (2.31)	51.15 (3.06)	65.85 (2.09)
SciBERT	49.89 (2.74)	65.38 (1.99)	52.17 (2.98)	66.70 (2.22)	51.99 (3.48)	<b>67.18</b> (2.45)

TABLE 2 – Valeurs F1 macro et F1 pondérée moyennes sur les quatre graines et écart-type entre parenthèse par fenêtre de contexte et par modèle sur l'ensemble de validation corpus Jiang2021

# Meilleurs F1 sur la validation

Modèle	Graine	Contexte 1-1		Contexte 2-2		Contexte 3-3	
		Best_F1m	Best_F1w	Best_F1m	Best_F1w	Best_F1m	Best_F1w
BioBERT	1965	47.66	63.59	47.93	64.99	46.63	64.81
	42	46.84	63.99	49.77	64.83	51.65	65.11
	5171	55.31	69.33	<b>58.29</b>	<b>70.46</b>	55.50	69.82
	798	52.97	64.32	51.69	63.20	52.78	65.98
BioLinkBERT	1965	49.68	65.15	50.48	65.56	50.85	65.24
	42	48.04	64.00	48.57	63.03	51.80	65.12
	5171	56.29	69.68	54.98	69.03	<b>56.36</b>	<b>70.41</b>
	798	52.27	65.87	52.60	66.78	50.36	65.30
RoBERTa	1965	47.11	62.21	49.83	64.57	50.30	65.18
	42	46.16	63.56	50.60	64.26	48.19	64.29
	5171	54.78	69.23	54.22	69.35	<b>56.28</b>	<b>69.43</b>
	798	53.77	65.86	49.82	63.47	49.84	64.49
SciBERT	1965	48.11	64.24	49.67	65.54	48.10	65.07
	42	46.43	63.16	49.40	65.25	49.52	65.56
	5171	53.24	68.46	56.78	70.55	<b>57.03</b>	<b>71.29</b>
	798	51.78	65.65	52.84	65.47	53.30	66.82

TABLE 3 – Meilleurs scores F1 macro (Best\_F1m) et F1 pondérée (Best\_F1w) par modèle, graine et longueur de fenêtre sur l'ensemble de validation corpus Jiang2021

# P, R et F1 par classe sur PD100cit et Jiang2021

Classe	Nb citations	P	R	F1
Neutral	72	88.60	83.68	85.59
Usage	10	82.64	67.5	74.09
CoCores	2	75	62.5	66.67
Motivation	2	39.58	37.5	35.0
Basis	2	0.00	0.00	0.00
Weakness	2	83.33	62.5	68.34
Support	3	64.58	100.00	75.59
CoCoXY	7	43.61	60.71	47.52

P, R, et F1 moyens (en %) et nombre de citations par classe sur PD100cit en test avec BioBERT affiné en fenêtre 3-3

Classe	P	R	F1
Neutral	74.20	68.62	70.75
Usage	73.52	79.76	76.32
CoCores	65.08	61.87	63.10
Motivation	48.86	58.47	51.96
Basis	55.40	53.99	54.12
Weakness	47.48	46.48	45.54
Support	39.49	27.70	35.46
CoCoXY	27.09	18.97	21.29
CoCoGM	60.66	66.36	63.17
Similar	61.05	62.76	61.77
Future	68.83	60.78	64.41

P, R, et F1 moyens (en %) par classe sur l'ensemble de validation corpus Jiang2021 avec BioBERT affiné en fenêtre 3-3

# Résultats : Transfert de LM affinés en TAL vers la biologie

