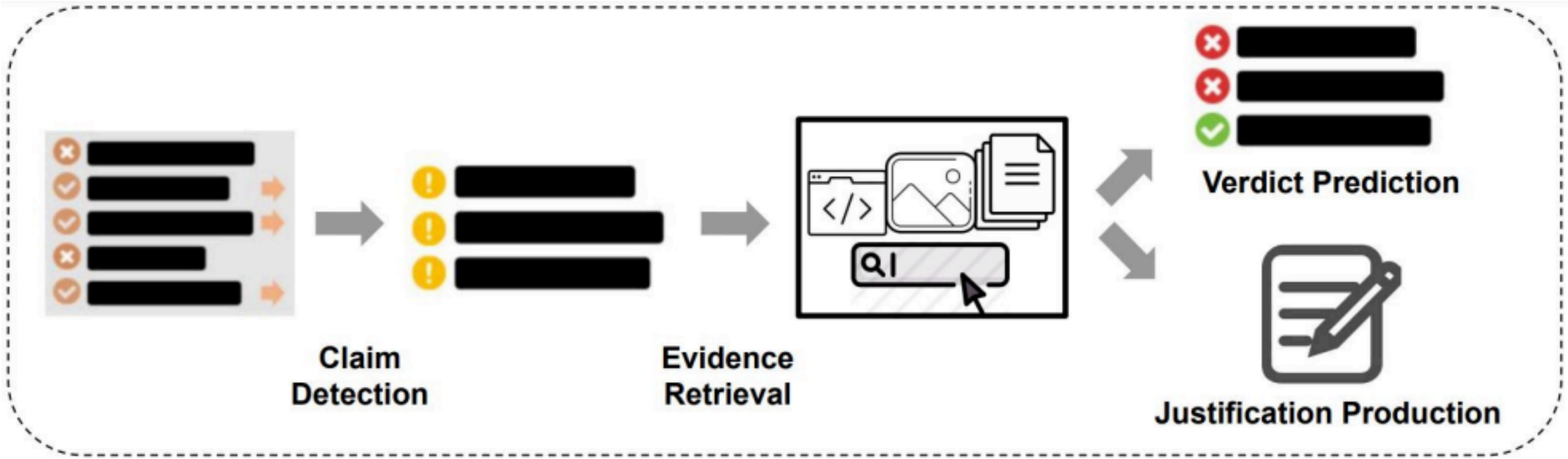
CHEF: A Pilot Chinese Dataset for Evidence-Based Fact-Checking

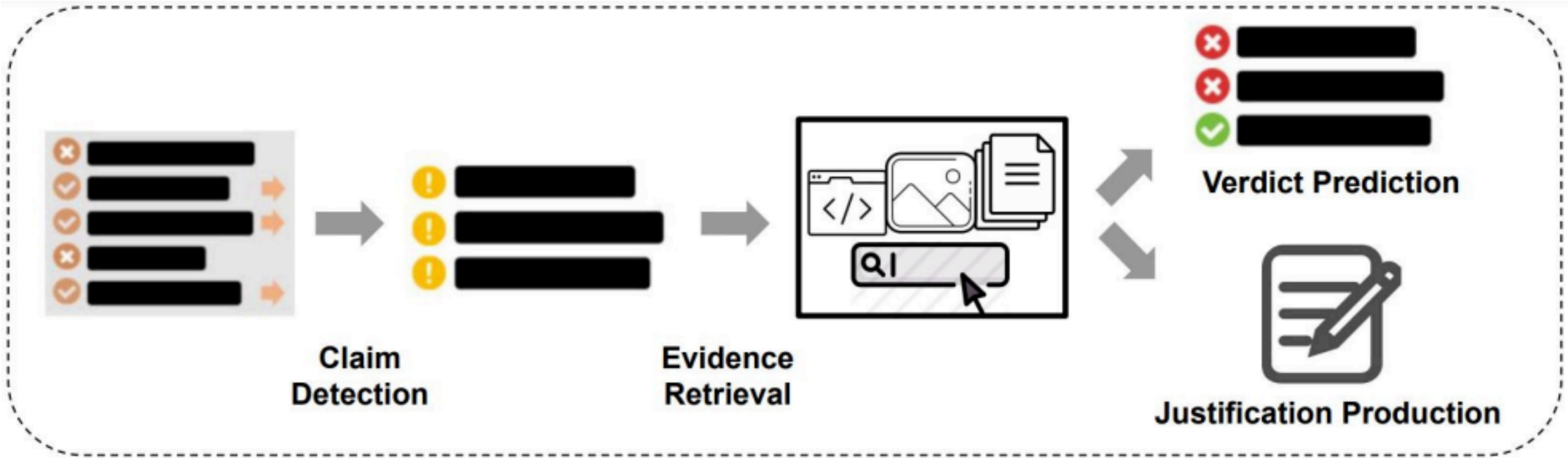
Xuming Hu[1*], Zhijiang Guo[2*], Guanyu Wu[1], Aiwei Liu[1], Lijie Wen[1] ,Philip S. Yu[1,3]

[1] Tsinghua University
[2] University of Cambridge
[3] University of Illinois at Chicago

Automated Fact Checking

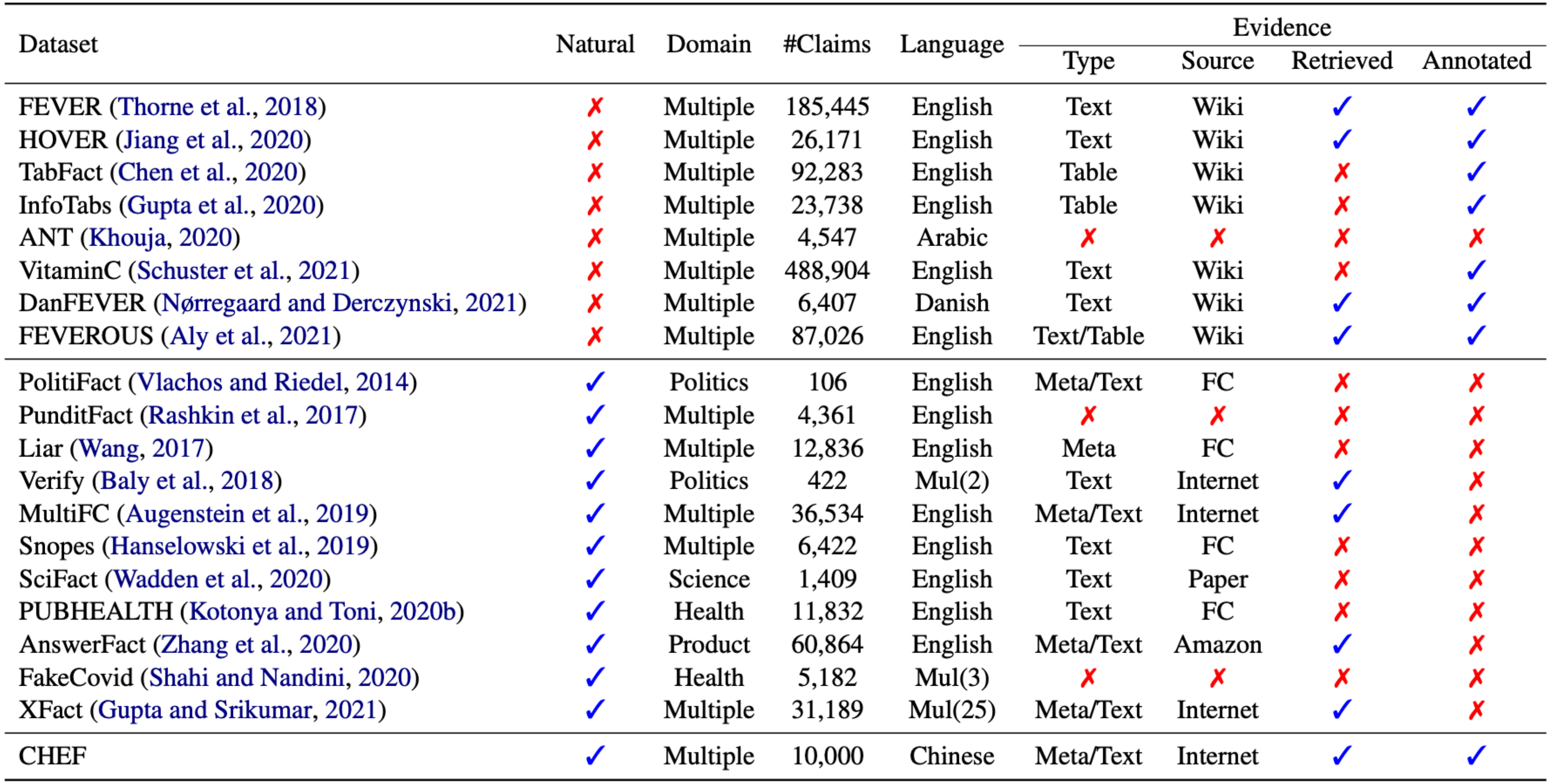


Automated Fact Checking



- A handful of non-English Datasets.
- Claims are created by non-English articles.

Dataset Comparisons

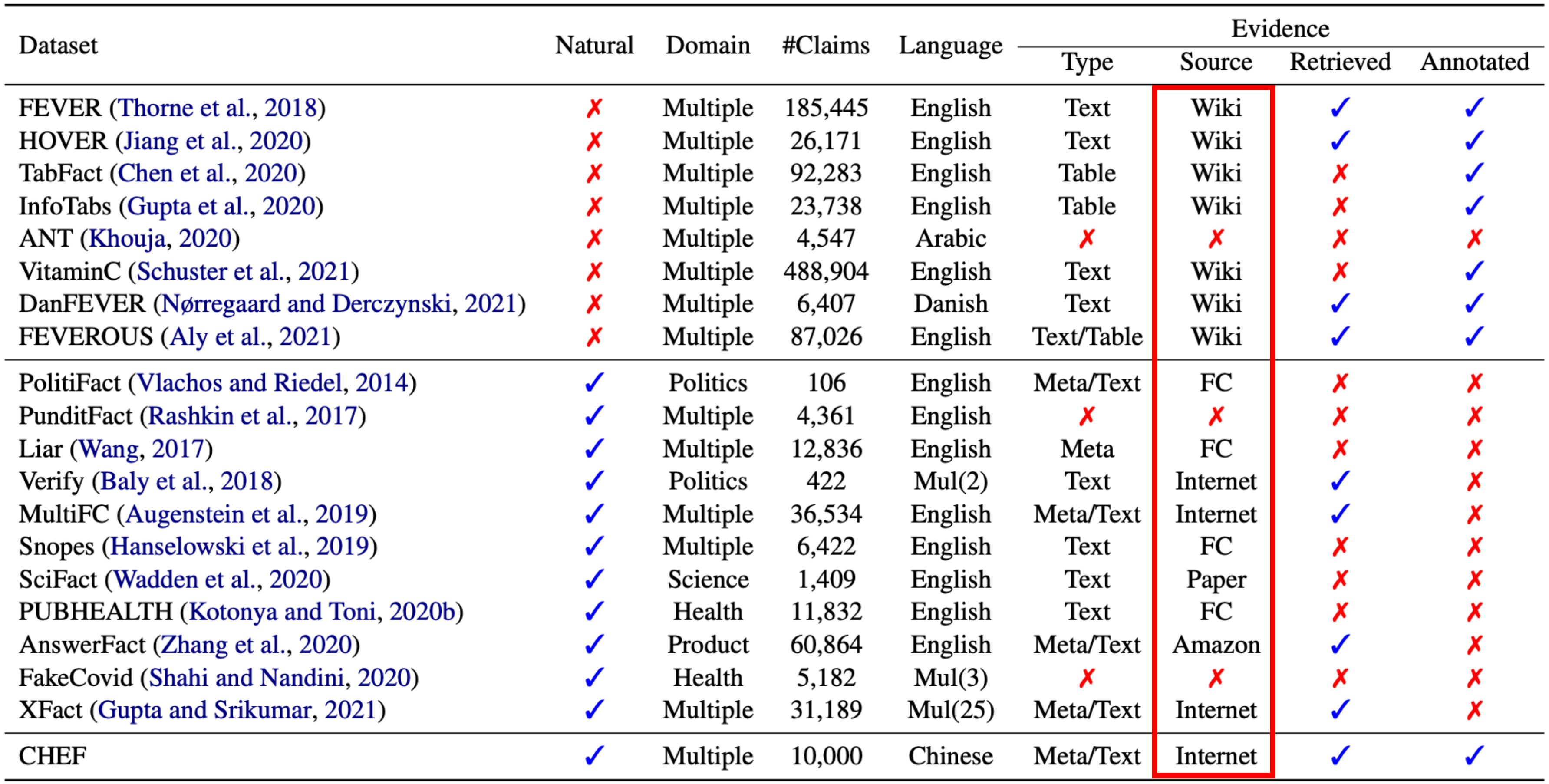


Dataset	Natural	Domain	#Claims	Language	Evidence			
					Type	Source	Retrieved	Annotated
FEVER (Thorne et al., 2018)	✗	Multiple	185,445	English	Text	Wiki	✓	✓
HOVER (Jiang et al., 2020)	✗	Multiple	26,171	English	Text	Wiki	✓	✓
TabFact (Chen et al., 2020)	✗	Multiple	92,283	English	Table	Wiki	✗	✓
InfoTabs (Gupta et al., 2020)	✗	Multiple	23,738	English	Table	Wiki	✗	✓
ANT (Khouja, 2020)	✗	Multiple	4,547	Arabic	✗	✗	✗	✗
VitaminC (Schuster et al., 2021)	✗	Multiple	488,904	English	Text	Wiki	✗	✓
DanFEVER (Nørregaard and Derczynski, 2021)	✗	Multiple	6,407	Danish	Text	Wiki	✓	✓
FEVEROUS (Aly et al., 2021)	✗	Multiple	87,026	English	Text/Table	Wiki	✓	✓
PolitiFact (Vlachos and Riedel, 2014)	✓	Politics	106	English	Meta/Text	FC	✗	✗
PunditFact (Rashkin et al., 2017)	✓	Multiple	4,361	English	✗	✗	✗	✗
Liar (Wang, 2017)	✓	Multiple	12,836	English	Meta	FC	✗	✗
Verify (Baly et al., 2018)	✓	Politics	422	Mul(2)	Text	Internet	✓	✗
MultiFC (Augenstein et al., 2019)	✓	Multiple	36,534	English	Meta/Text	Internet	✓	✗
Snopes (Hanselowski et al., 2019)	✓	Multiple	6,422	English	Text	FC	✗	✗
SciFact (Wadden et al., 2020)	✓	Science	1,409	English	Text	Paper	✗	✗
PUBHEALTH (Kotonya and Toni, 2020b)	✓	Health	11,832	English	Text	FC	✗	✗
AnswerFact (Zhang et al., 2020)	✓	Product	60,864	English	Meta/Text	Amazon	✓	✗
FakeCovid (Shahi and Nandini, 2020)	✓	Health	5,182	Mul(3)	✗	✗	✗	✗
XFact (Gupta and Srikumar, 2021)	✓	Multiple	31,189	Mul(25)	Meta/Text	Internet	✓	✗
CHEF	✓	Multiple	10,000	Chinese	Meta/Text	Internet	✓	✓

Table: Comparisons of fact-checking datasets.

- Natural
- Synthetic

Dataset Comparisons



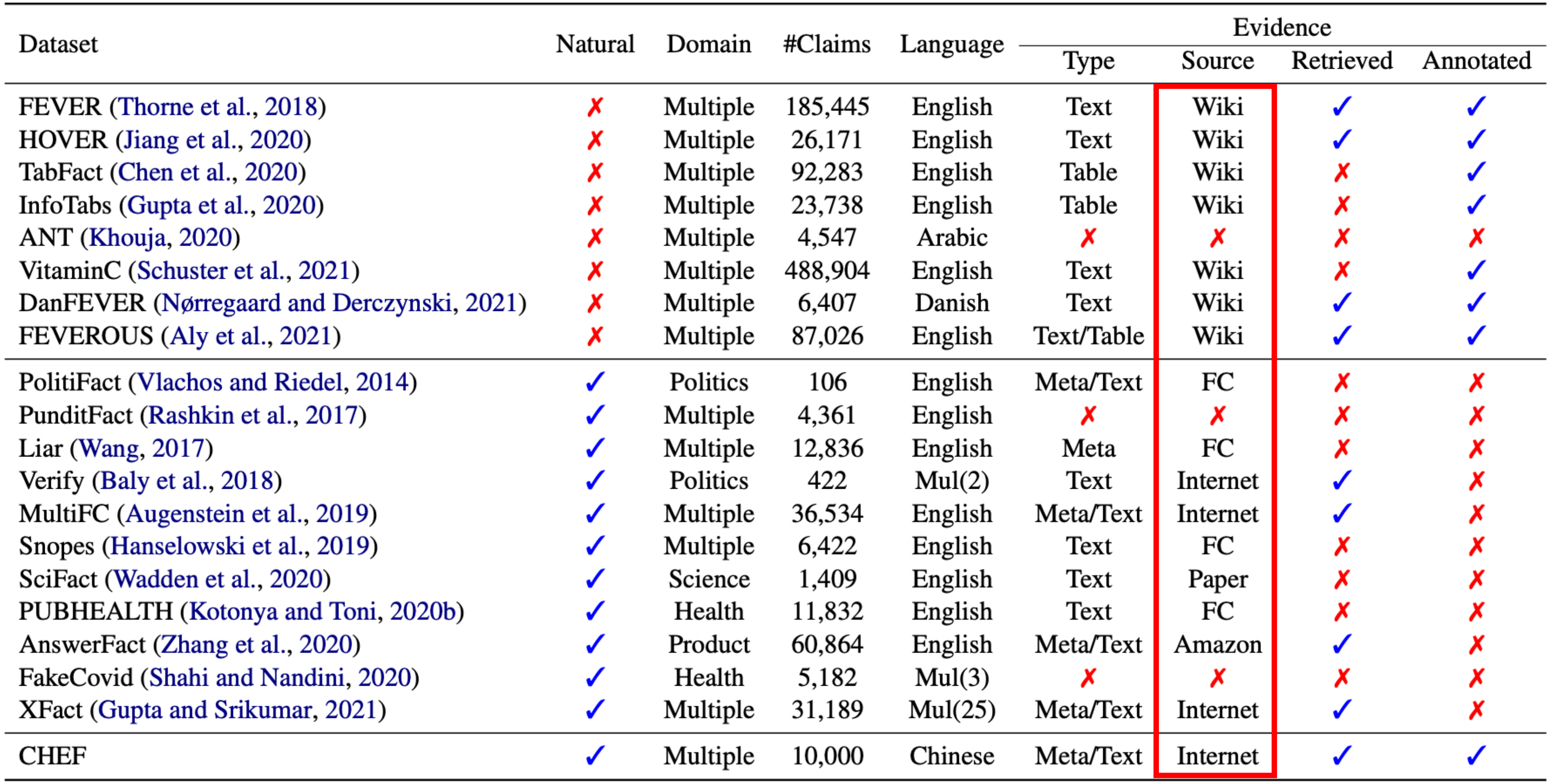
Dataset	Natural	Domain	#Claims	Language	Evidence			
					Type	Source	Retrieved	Annotated
FEVER (Thorne et al., 2018)	✗	Multiple	185,445	English	Text	Wiki	✓	✓
HOVER (Jiang et al., 2020)	✗	Multiple	26,171	English	Text	Wiki	✓	✓
TabFact (Chen et al., 2020)	✗	Multiple	92,283	English	Table	Wiki	✗	✓
InfoTabs (Gupta et al., 2020)	✗	Multiple	23,738	English	Table	Wiki	✗	✓
ANT (Khouja, 2020)	✗	Multiple	4,547	Arabic	✗	✗	✗	✗
VitaminC (Schuster et al., 2021)	✗	Multiple	488,904	English	Text	Wiki	✗	✓
DanFEVER (Nørregaard and Derczynski, 2021)	✗	Multiple	6,407	Danish	Text	Wiki	✓	✓
FEVEROUS (Aly et al., 2021)	✗	Multiple	87,026	English	Text/Table	Wiki	✓	✓
PolitiFact (Vlachos and Riedel, 2014)	✓	Politics	106	English	Meta/Text	FC	✗	✗
PunditFact (Rashkin et al., 2017)	✓	Multiple	4,361	English	✗	✗	✗	✗
Liar (Wang, 2017)	✓	Multiple	12,836	English	Meta	FC	✗	✗
Verify (Baly et al., 2018)	✓	Politics	422	Mul(2)	Text	Internet	✓	✗
MultiFC (Augenstein et al., 2019)	✓	Multiple	36,534	English	Meta/Text	Internet	✓	✗
Snopes (Hanselowski et al., 2019)	✓	Multiple	6,422	English	Text	FC	✗	✗
SciFact (Wadden et al., 2020)	✓	Science	1,409	English	Text	Paper	✗	✗
PUBHEALTH (Kotonya and Toni, 2020b)	✓	Health	11,832	English	Text	FC	✗	✗
AnswerFact (Zhang et al., 2020)	✓	Product	60,864	English	Meta/Text	Amazon	✓	✗
FakeCovid (Shahi and Nandini, 2020)	✓	Health	5,182	Mul(3)	✗	✗	✗	✗
XFact (Gupta and Srikumar, 2021)	✓	Multiple	31,189	Mul(25)	Meta/Text	Internet	✓	✗
CHEF	✓	Multiple	10,000	Chinese	Meta/Text	Internet	✓	✓

Table: Comparisons of fact-checking datasets.

Synthetic:

- Restricted world knowledge to a single source.
- Claims created artificially by mutating sentences from Wikipedia articles.

Dataset Comparisons



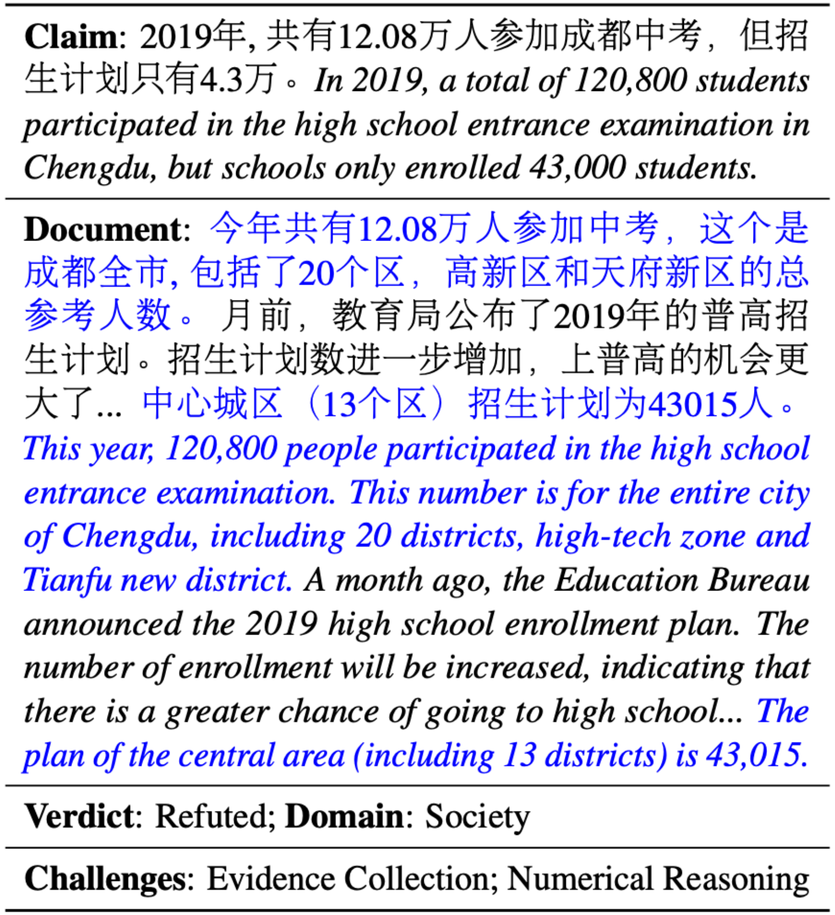
Dataset	Natural	Domain	#Claims	Language	Evidence			
					Type	Source	Retrieved	Annotated
FEVER (Thorne et al., 2018)	✗	Multiple	185,445	English	Text	Wiki	✓	✓
HOVER (Jiang et al., 2020)	✗	Multiple	26,171	English	Text	Wiki	✓	✓
TabFact (Chen et al., 2020)	✗	Multiple	92,283	English	Table	Wiki	✗	✓
InfoTabs (Gupta et al., 2020)	✗	Multiple	23,738	English	Table	Wiki	✗	✓
ANT (Khouja, 2020)	✗	Multiple	4,547	Arabic	✗	✗	✗	✗
VitaminC (Schuster et al., 2021)	✗	Multiple	488,904	English	Text	Wiki	✗	✓
DanFEVER (Nørregaard and Derczynski, 2021)	✗	Multiple	6,407	Danish	Text	Wiki	✓	✓
FEVEROUS (Aly et al., 2021)	✗	Multiple	87,026	English	Text/Table	Wiki	✓	✓
PolitiFact (Vlachos and Riedel, 2014)	✓	Politics	106	English	Meta/Text	FC	✗	✗
PunditFact (Rashkin et al., 2017)	✓	Multiple	4,361	English	✗	✗	✗	✗
Liar (Wang, 2017)	✓	Multiple	12,836	English	Meta	FC	✗	✗
Verify (Baly et al., 2018)	✓	Politics	422	Mul(2)	Text	Internet	✓	✗
MultiFC (Augenstein et al., 2019)	✓	Multiple	36,534	English	Meta/Text	Internet	✓	✗
Snopes (Hanselowski et al., 2019)	✓	Multiple	6,422	English	Text	FC	✗	✗
SciFact (Wadden et al., 2020)	✓	Science	1,409	English	Text	Paper	✗	✗
PUBHEALTH (Kotonya and Toni, 2020b)	✓	Health	11,832	English	Text	FC	✗	✗
AnswerFact (Zhang et al., 2020)	✓	Product	60,864	English	Meta/Text	Amazon	✓	✗
FakeCovid (Shahi and Nandini, 2020)	✓	Health	5,182	Mul(3)	✗	✗	✗	✗
XFact (Gupta and Srikumar, 2021)	✓	Multiple	31,189	Mul(25)	Meta/Text	Internet	✓	✗
CHEF	✓	Multiple	10,000	Chinese	Meta/Text	Internet	✓	✓

Table: Comparisons of fact-checking datasets.

Natural:

- Fact checking websites are small in size.
- Summary snippets do not provide sufficient information.

CHEF



Claim: 2019年, 共有12.08万人参加成都中考，但招生计划只有4.3万。*In 2019, a total of 120,800 students participated in the high school entrance examination in Chengdu, but schools only enrolled 43,000 students.*

Document: 今年共有12.08万人参加中考，这个是成都全市, 包括了20个区，高新区和天府新区的总参考人数。月前，教育局公布了2019年的普高招生计划。招生计划数进一步增加，上普高的机会更大了... 中心城区（13个区）招生计划为43015人。*This year, 120,800 people participated in the high school entrance examination. This number is for the entire city of Chengdu, including 20 districts, high-tech zone and Tianfu new district. A month ago, the Education Bureau announced the 2019 high school enrollment plan. The number of enrollment will be increased, indicating that there is a greater chance of going to high school... The plan of the central area (including 13 districts) is 43,015.*

Verdict: Refuted; **Domain**: Society

Challenges: Evidence Collection; Numerical Reasoning

Table: An example from CHEF.

CHEF: CHinese dataset for Evidence-based Fact-checking

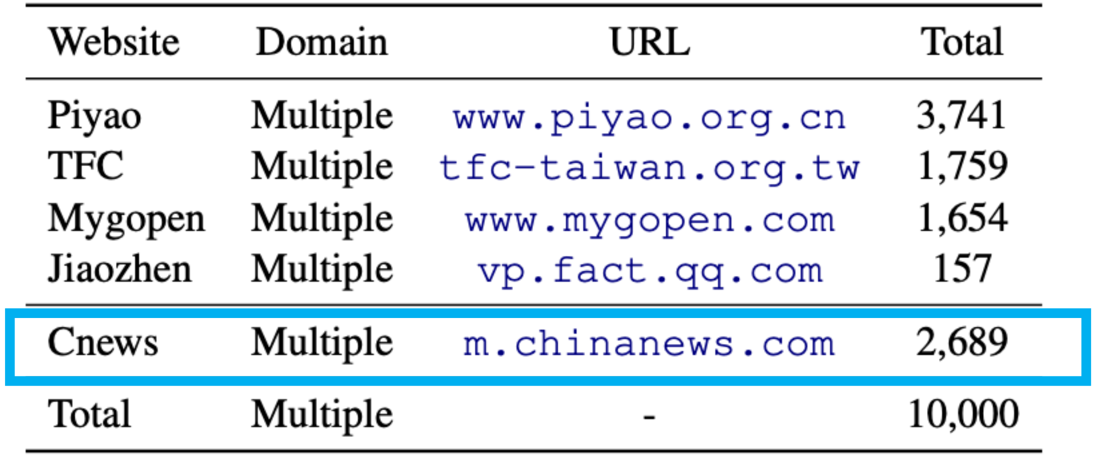
- 10,000 real-world claims
- 6 Chinese fact-checking websites
- Annotated evidence
- Developed suitable guidelines
- Performed data validation

Dataset Construction

- Data collection
- Claim labeling
- Evidence retrieval
- Data validation

Dataset Construction

- Data collection



Website	Domain	URL	Total
Piyao	Multiple	www.piyao.org.cn	3,741
TFC	Multiple	tfc-taiwan.org.tw	1,759
Mygopen	Multiple	www.mygopen.com	1,654
Jiaozhen	Multiple	vp.fact.qq.com	157
Cnews	Multiple	m.chinanews.com	2,689
Total	Multiple	-	10,000

Table: Statistics of data source.

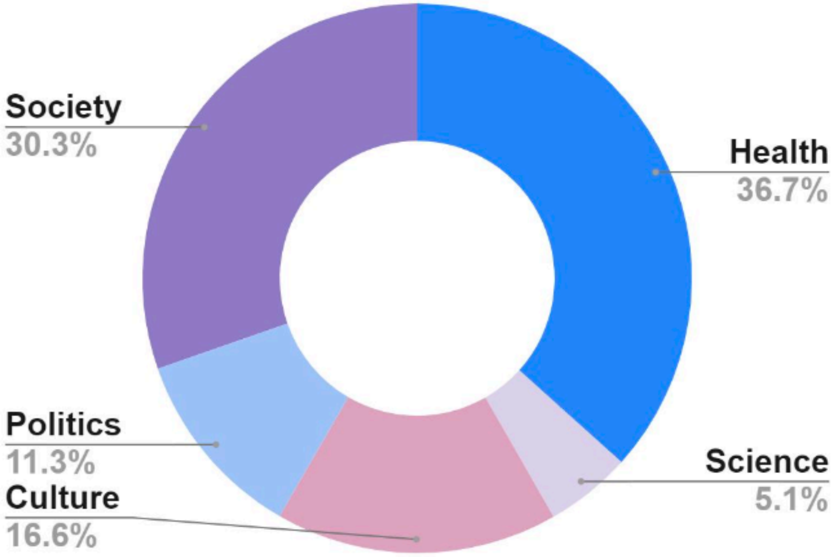
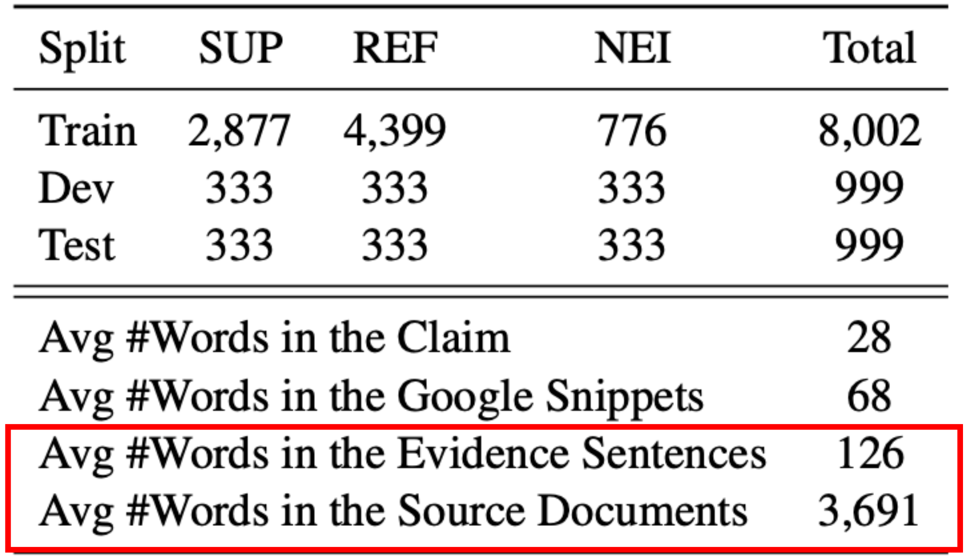


Figure: Distributions of domains.

Dataset Construction

- Claim Labeling



Split	SUP	REF	NEI	Total
Train	2,877	4,399	776	8,002
Dev	333	333	333	999
Test	333	333	333	999
Avg #Words in the Claim				28
Avg #Words in the Google Snippets				68
Avg #Words in the Evidence Sentences				126
Avg #Words in the Source Documents				3,691

Table: Dataset split sizes and statistics for CHEF.

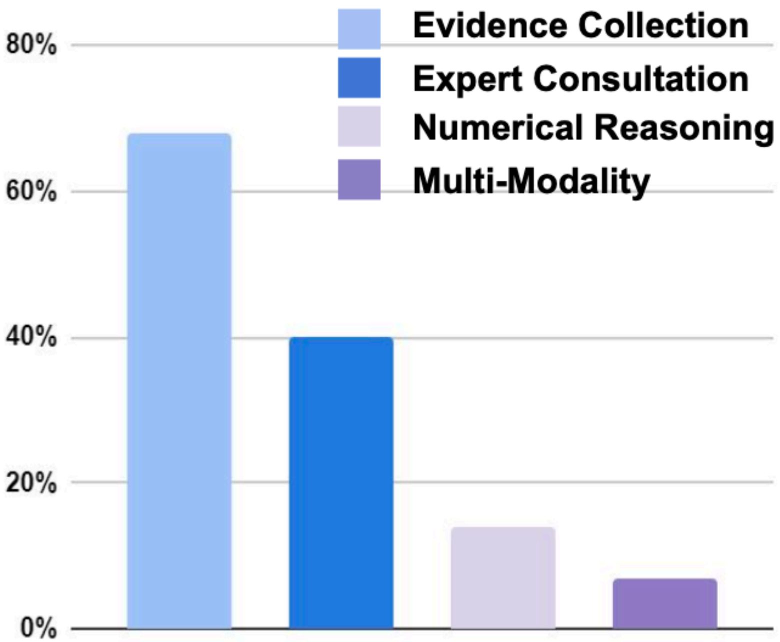
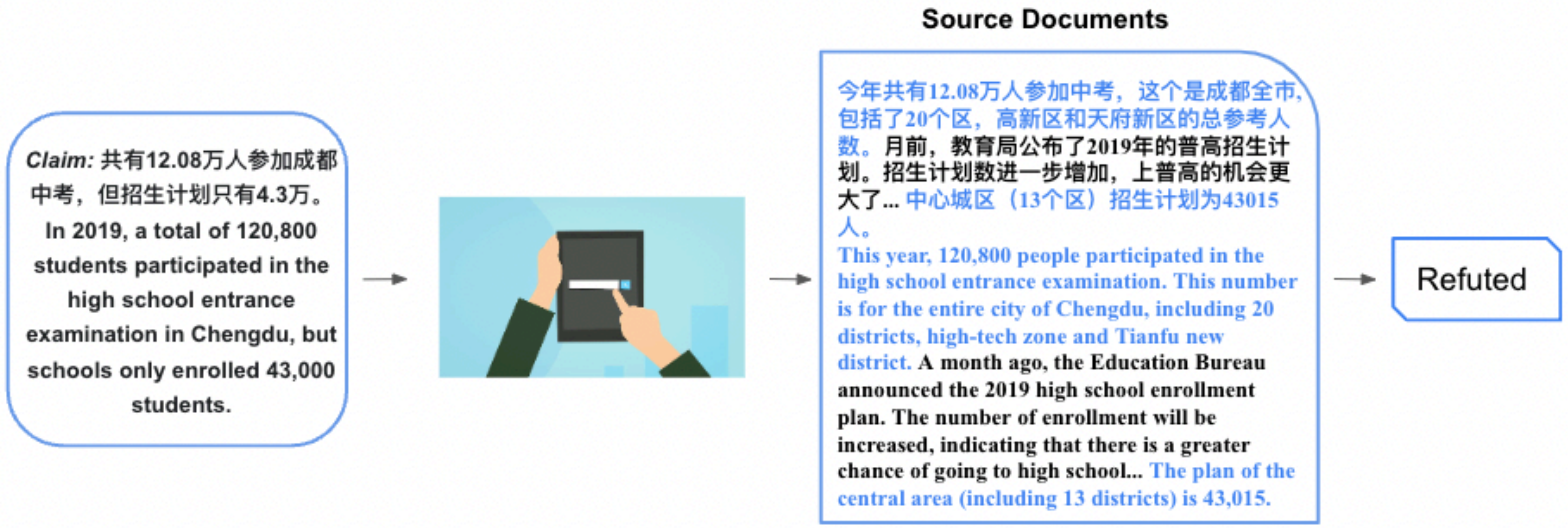


Figure: Distributions of challenges.

Dataset Construction

- Evidence Retrieval



Claim: 共有12.08万人参加成都中考，但招生计划只有4.3万。 In 2019, a total of 120,800 students participated in the high school entrance examination in Chengdu, but schools only enrolled 43,000 students.

Source Documents

今年共有12.08万人参加中考，这个是成都全市,包括了20个区，高新区和天府新区的总参考人数。月前，教育局公布了2019年的普高招生计划。招生计划数进一步增加，上普高的机会更大了... 中心城区（13个区）招生计划为43015人。

This year, 120,800 people participated in the high school entrance examination. This number is for the entire city of Chengdu, including 20 districts, high-tech zone and Tianfu new district. A month ago, the Education Bureau announced the 2019 high school enrollment plan. The number of enrollment will be increased, indicating that there is a greater chance of going to high school... The plan of the central area (including 13 districts) is 43,015.

Refuted

The claim is refuted by the evidence, which are sentences retrieved (highlighted) from the document.

Dataset Construction

- Data Validation

  5-way inter-annotator agreement

  - 310 Claims
  - 5 Annotators

  Fleiss K score = 0.74

  Another 310 Claims

  - 88.7% were labeled correctly
  - 83.6% provided sufficient information

Baseline Systems

- Pipeline Systems

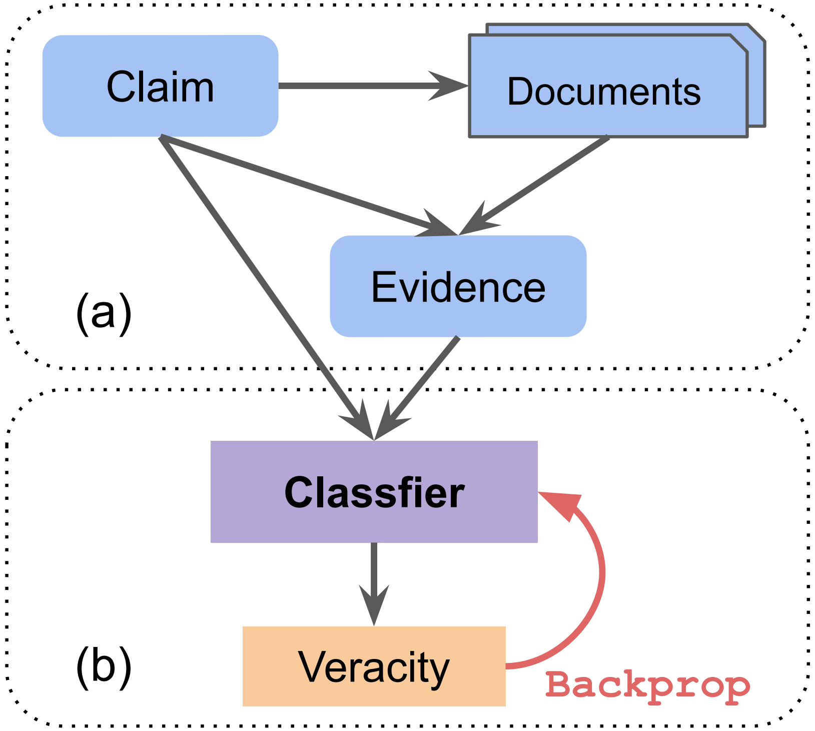


Figure: Pipeline Systems

Baseline Systems

- Pipeline Systems

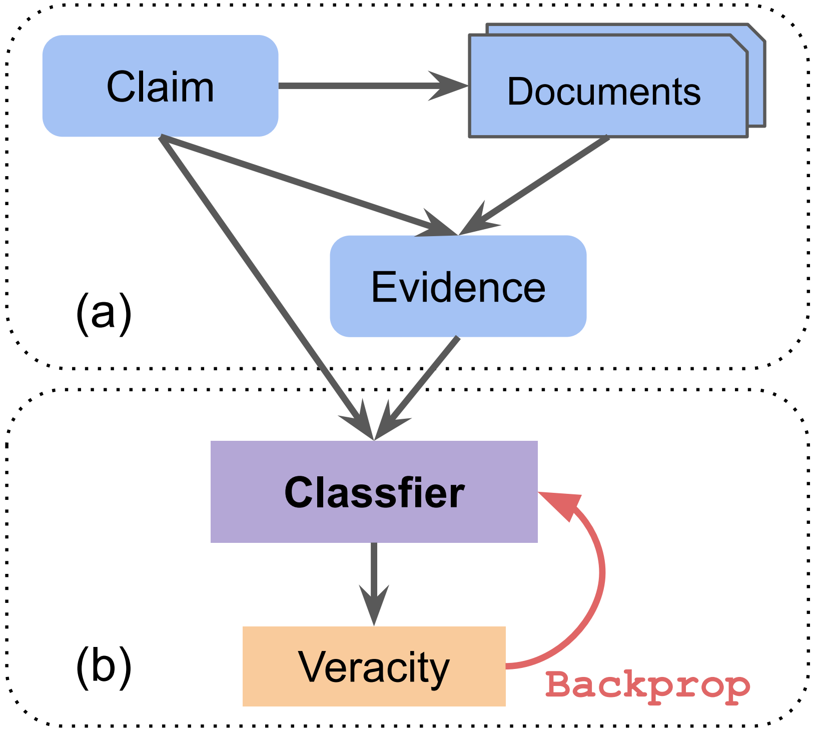


Figure: Pipeline Systems

Evidence Retrieval

- Surface Ranker: TF-IDF
- Semantic Ranker: Cosine similarity
- Hybrid Ranker: RankSVM
- Google Snippets: Google Search Engine

Baseline Systems

- Pipeline Systems

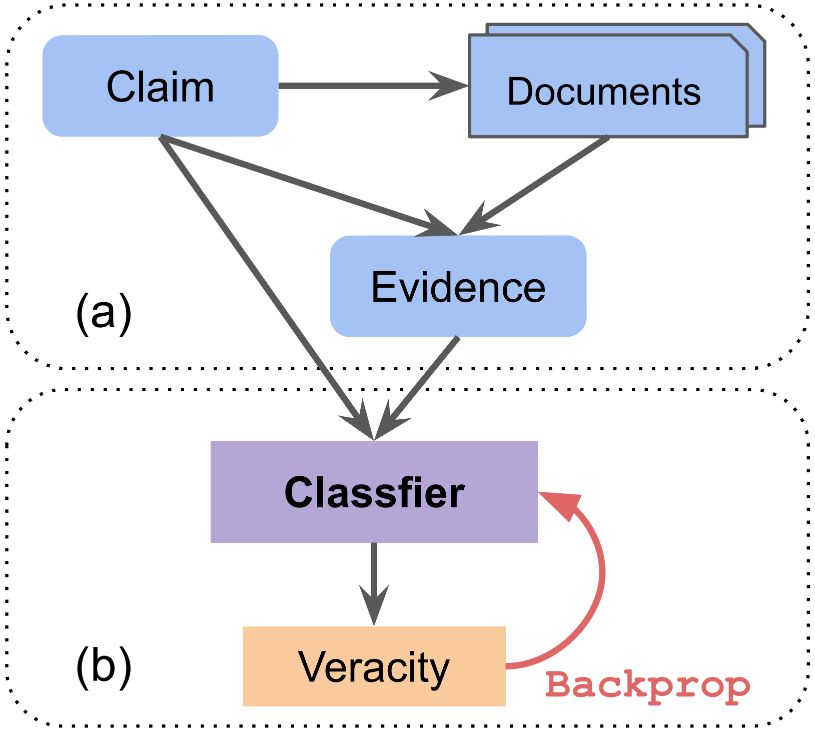


Figure: Pipeline Systems

Veracity Prediction

- BERT-Based Model
- Attention-Based Model
- Graph-Based Model

Baseline Systems

- Joint Systems

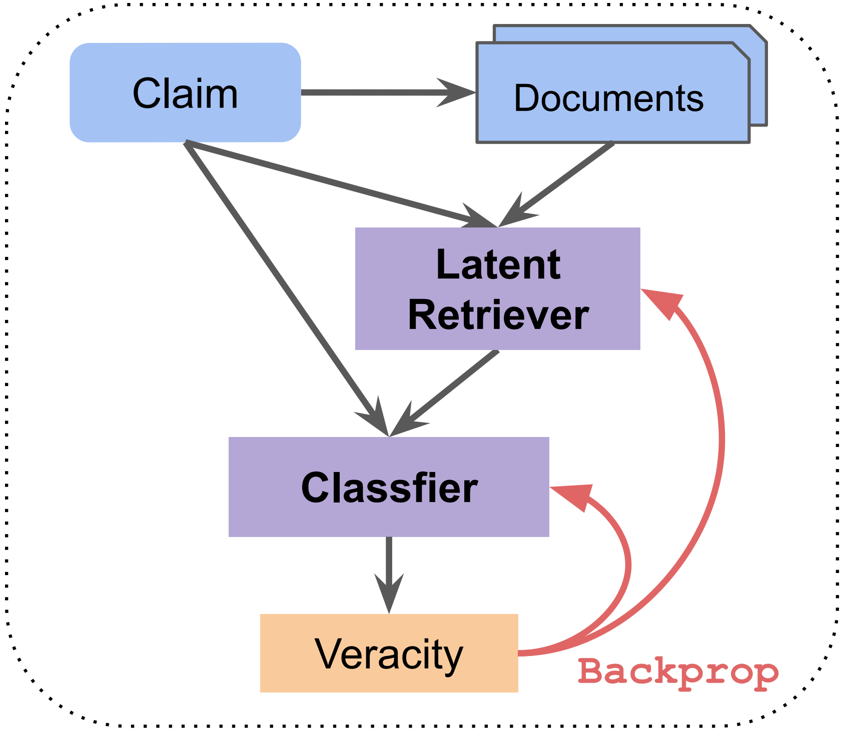


Figure: Joint Systems

Latent Retriever

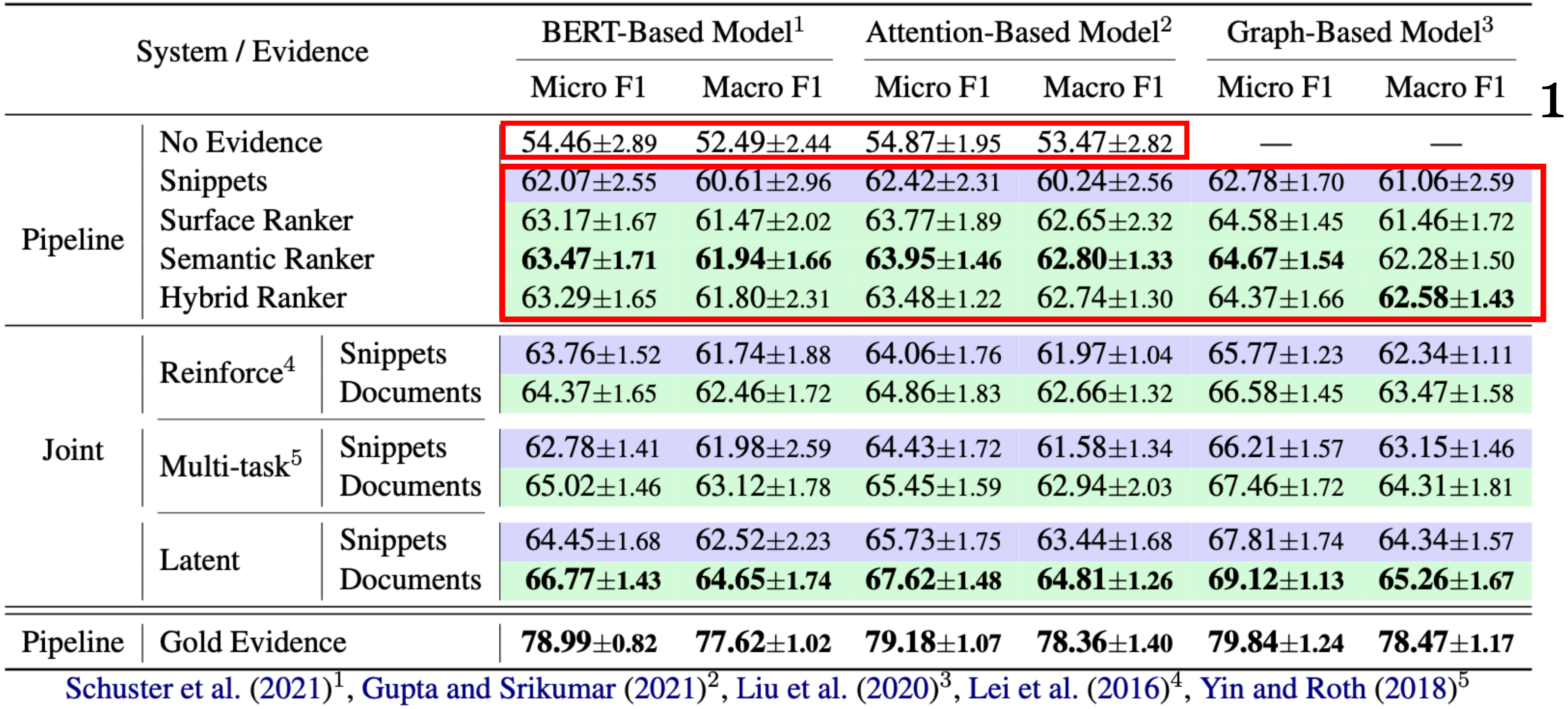
- Hard Kumaraswamy distribution (Bastings et al., 2019)

Baseline Systems

- More Baselines
- Reinforce (Lei et al. 2016)
- Multi-task (Yin and Roth 2018)

Experiments and Analysis

- Main Results



System / Evidence			BERT-Based Model[1]		Attention-Based Model[2]		Graph-Based Model[3]	
			Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1
Pipeline	No Evidence		54.46±2.89	52.49±2.44	54.87±1.95	53.47±2.82	—	—
	Snippets		62.07±2.55	60.61±2.96	62.42±2.31	60.24±2.56	62.78±1.70	61.06±2.59
	Surface Ranker		63.17±1.67	61.47±2.02	63.77±1.89	62.65±2.32	64.58±1.45	61.46±1.72
	Semantic Ranker		**63.47±1.71**	**61.94±1.66**	**63.95±1.46**	**62.80±1.33**	**64.67±1.54**	62.28±1.50
	Hybrid Ranker		63.29±1.65	61.80±2.31	63.48±1.22	62.74±1.30	64.37±1.66	**62.58±1.43**
Joint	Reinforce[4]	Snippets	63.76±1.52	61.74±1.88	64.06±1.76	61.97±1.04	65.77±1.23	62.34±1.11
		Documents	64.37±1.65	62.46±1.72	64.86±1.83	62.66±1.32	66.58±1.45	63.47±1.58
	Multi-task[5]	Snippets	62.78±1.41	61.98±2.59	64.43±1.72	61.58±1.34	66.21±1.57	63.15±1.46
		Documents	65.02±1.46	63.12±1.78	65.45±1.59	62.94±2.03	67.46±1.72	64.31±1.81
	Latent	Snippets	64.45±1.68	62.52±2.23	65.73±1.75	63.44±1.68	67.81±1.74	64.34±1.57
		Documents	**66.77±1.43**	**64.65±1.74**	**67.62±1.48**	**64.81±1.26**	**69.12±1.13**	**65.26±1.67**
Pipeline	Gold Evidence		**78.99**±0.82	**77.62**±1.02	**79.18**±1.07	**78.36**±1.40	**79.84**±1.24	**78.47**±1.17

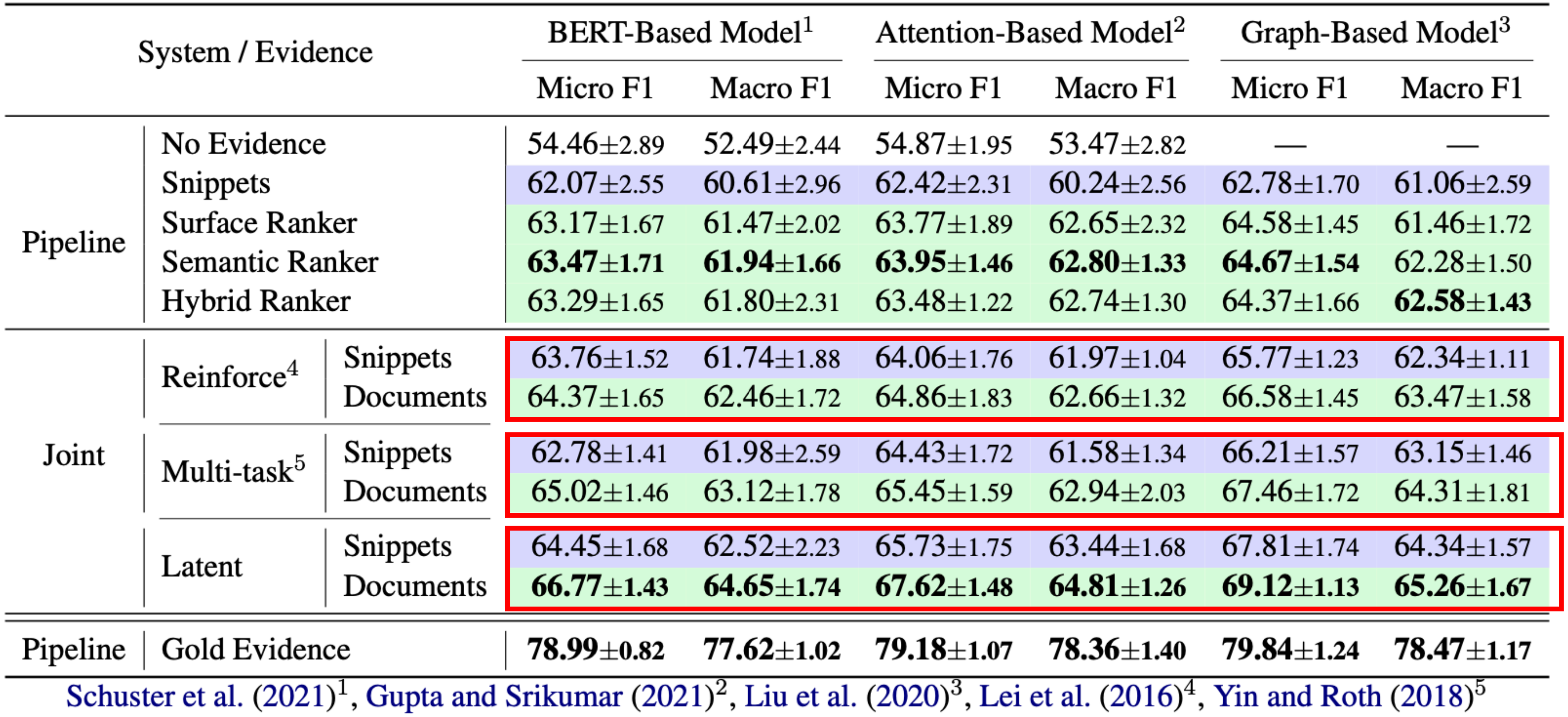
Schuster et al. (2021)[1], Gupta and Srikumar (2021)[2], Liu et al. (2020)[3], Lei et al. (2016)[4], Yin and Roth (2018)[5]

Table: Main results.

1. Evidence plays an important role in verifying real-world claims.

Experiments and Analysis

- Main Results



System / Evidence			BERT-Based Model[1]		Attention-Based Model[2]		Graph-Based Model[3]	
			Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1
Pipeline	No Evidence		54.46±2.89	52.49±2.44	54.87±1.95	53.47±2.82	—	—
	Snippets		62.07±2.55	60.61±2.96	62.42±2.31	60.24±2.56	62.78±1.70	61.06±2.59
	Surface Ranker		63.17±1.67	61.47±2.02	63.77±1.89	62.65±2.32	64.58±1.45	61.46±1.72
	Semantic Ranker		**63.47±1.71**	**61.94±1.66**	**63.95±1.46**	**62.80±1.33**	**64.67±1.54**	62.28±1.50
	Hybrid Ranker		63.29±1.65	61.80±2.31	63.48±1.22	62.74±1.30	64.37±1.66	**62.58±1.43**
Joint	Reinforce[4]	Snippets	63.76±1.52	61.74±1.88	64.06±1.76	61.97±1.04	65.77±1.23	62.34±1.11
		Documents	64.37±1.65	62.46±1.72	64.86±1.83	62.66±1.32	66.58±1.45	63.47±1.58
	Multi-task[5]	Snippets	62.78±1.41	61.98±2.59	64.43±1.72	61.58±1.34	66.21±1.57	63.15±1.46
		Documents	65.02±1.46	63.12±1.78	65.45±1.59	62.94±2.03	67.46±1.72	64.31±1.81
	Latent	Snippets	64.45±1.68	62.52±2.23	65.73±1.75	63.44±1.68	67.81±1.74	64.34±1.57
		Documents	**66.77±1.43**	**64.65±1.74**	**67.62±1.48**	**64.81±1.26**	**69.12±1.13**	**65.26±1.67**
Pipeline	Gold Evidence		**78.99**±0.82	**77.62**±1.02	**79.18**±1.07	**78.36**±1.40	**79.84**±1.24	**78.47**±1.17

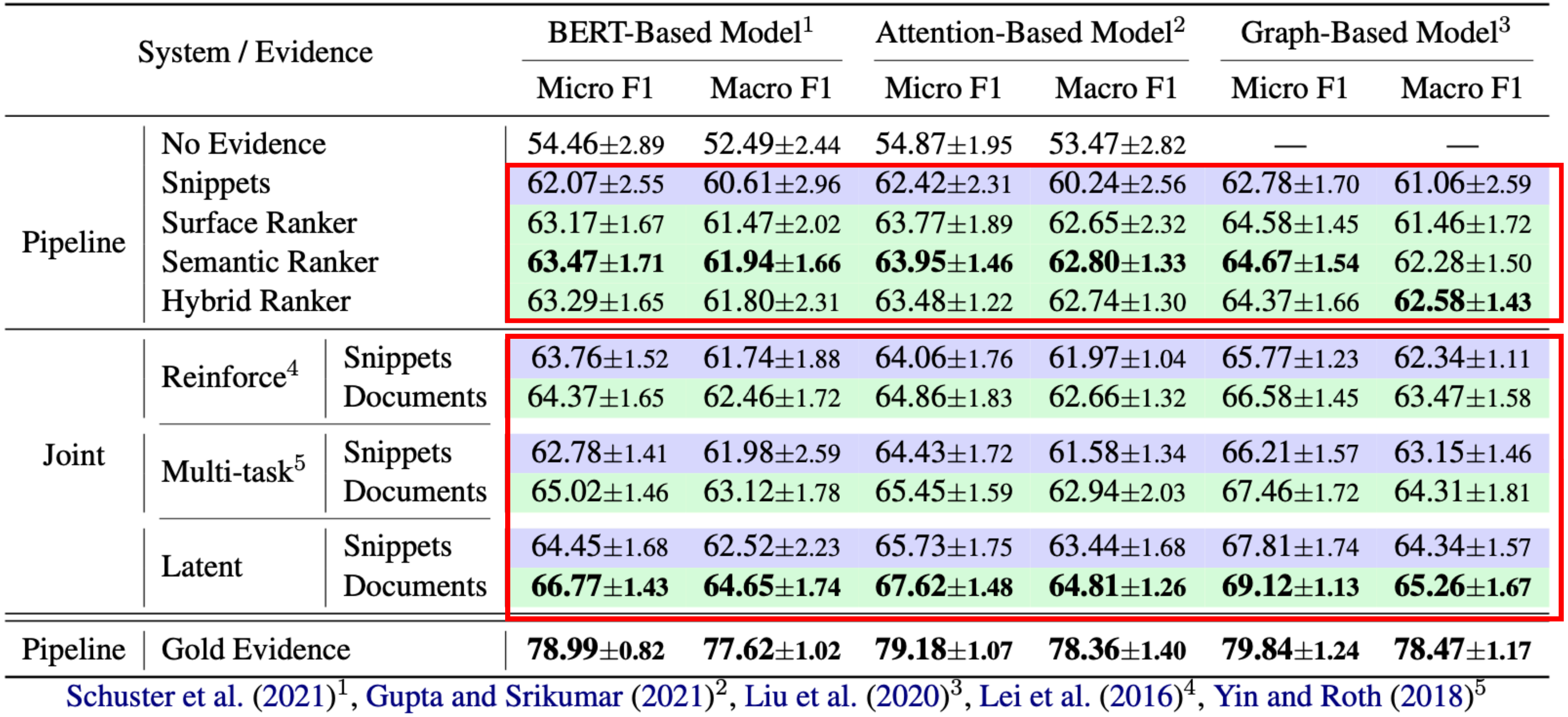
Schuster et al. (2021)[1], Gupta and Srikumar (2021)[2], Liu et al. (2020)[3], Lei et al. (2016)[4], Yin and Roth (2018)[5]

Table: Main results.

1. Evidence plays an important role in verifying real-world claims.
2. Retrieving evidence sentences from documents achieve better F1 scores than directly use the summary snippets.

Experiments and Analysis

- Main Results



System / Evidence			BERT-Based Model[1]		Attention-Based Model[2]		Graph-Based Model[3]	
			Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1
Pipeline	No Evidence		54.46±2.89	52.49±2.44	54.87±1.95	53.47±2.82	—	—
	Snippets		62.07±2.55	60.61±2.96	62.42±2.31	60.24±2.56	62.78±1.70	61.06±2.59
	Surface Ranker		63.17±1.67	61.47±2.02	63.77±1.89	62.65±2.32	64.58±1.45	61.46±1.72
	Semantic Ranker		**63.47±1.71**	**61.94±1.66**	**63.95±1.46**	**62.80±1.33**	**64.67±1.54**	62.28±1.50
	Hybrid Ranker		63.29±1.65	61.80±2.31	63.48±1.22	62.74±1.30	64.37±1.66	**62.58±1.43**
Joint	Reinforce[4]	Snippets	63.76±1.52	61.74±1.88	64.06±1.76	61.97±1.04	65.77±1.23	62.34±1.11
		Documents	64.37±1.65	62.46±1.72	64.86±1.83	62.66±1.32	66.58±1.45	63.47±1.58
	Multi-task[5]	Snippets	62.78±1.41	61.98±2.59	64.43±1.72	61.58±1.34	66.21±1.57	63.15±1.46
		Documents	65.02±1.46	63.12±1.78	65.45±1.59	62.94±2.03	67.46±1.72	64.31±1.81
	Latent	Snippets	64.45±1.68	62.52±2.23	65.73±1.75	63.44±1.68	67.81±1.74	64.34±1.57
		Documents	**66.77±1.43**	**64.65±1.74**	**67.62±1.48**	**64.81±1.26**	**69.12±1.13**	**65.26±1.67**
Pipeline	Gold Evidence		**78.99**±0.82	**77.62**±1.02	**79.18**±1.07	**78.36**±1.40	**79.84**±1.24	**78.47**±1.17

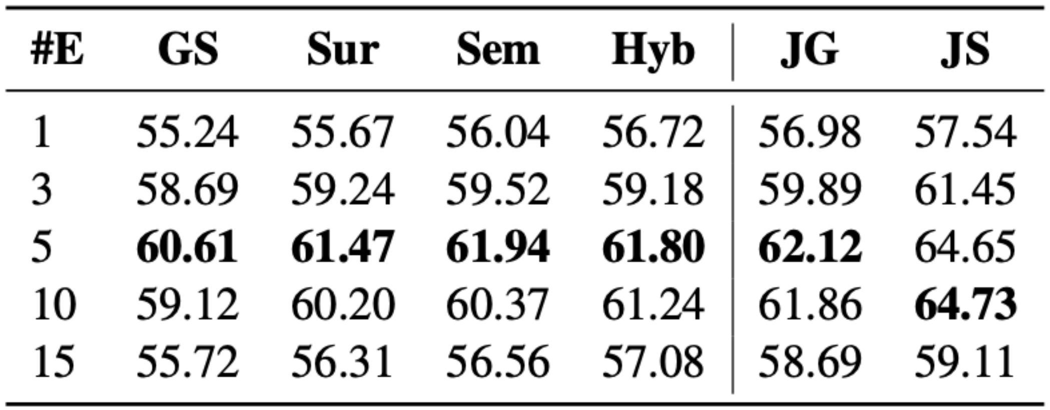
Schuster et al. (2021)[1], Gupta and Srikumar (2021)[2], Liu et al. (2020)[3], Lei et al. (2016)[4], Yin and Roth (2018)[5]

Table: Main results.

1. Evidence plays an important role in verifying real-world claims.
2. Retrieving evidence sentences from documents achieve better F1 scores than directly use the summary snippets.
3. Joint system outperforms pipeline system consistently with both Google snippets and source documents as inputs.

Experiments and Analysis

- Effect of Evidence



#E	GS	Sur	Sem	Hyb	JG	JS
1	55.24	55.67	56.04	56.72	56.98	57.54
3	58.69	59.24	59.52	59.18	59.89	61.45
5	**60.61**	**61.47**	**61.94**	**61.80**	**62.12**	64.65
10	59.12	60.20	60.37	61.24	61.86	**64.73**
15	55.72	56.31	56.56	57.08	58.69	59.11

Table: Effect of Evidence. #E indicates the number of evidence.

The fluctuation results indicate that both **quantity and quality** of retrieved evidence affect the performance.

- Fewer evidence -> incomplete coverage
- More evidence -> irrelevant sentences

Experiments and Analysis

- Performance against Claim Length

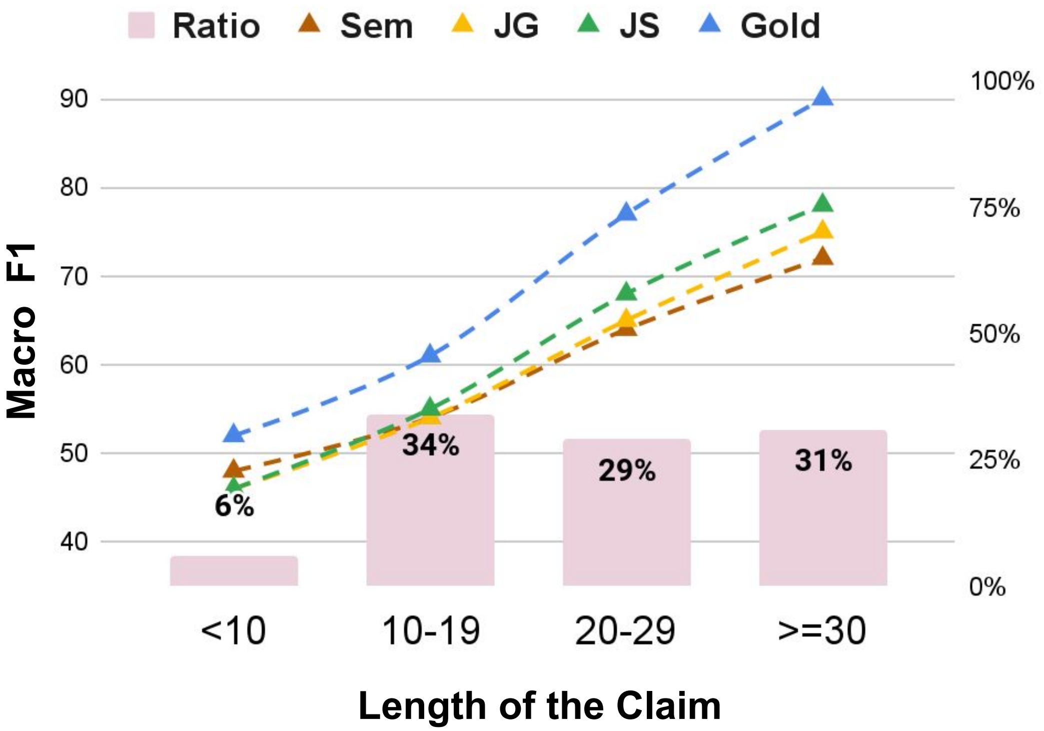


Figure: Comparisons against claim lengths.

1. Most claims are longer than **10** words.
2. Performance of the systems on short claims is lower than other.

Experiments and Analysis

- Performance against Classes and Domains

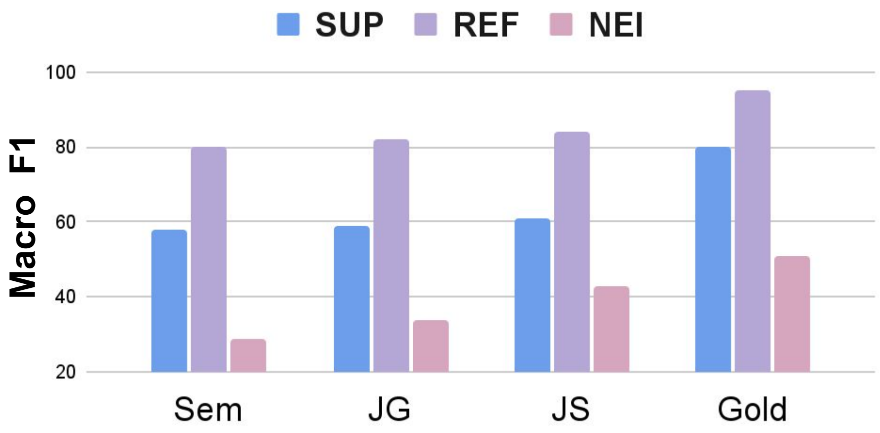


Figure: Per-class results.

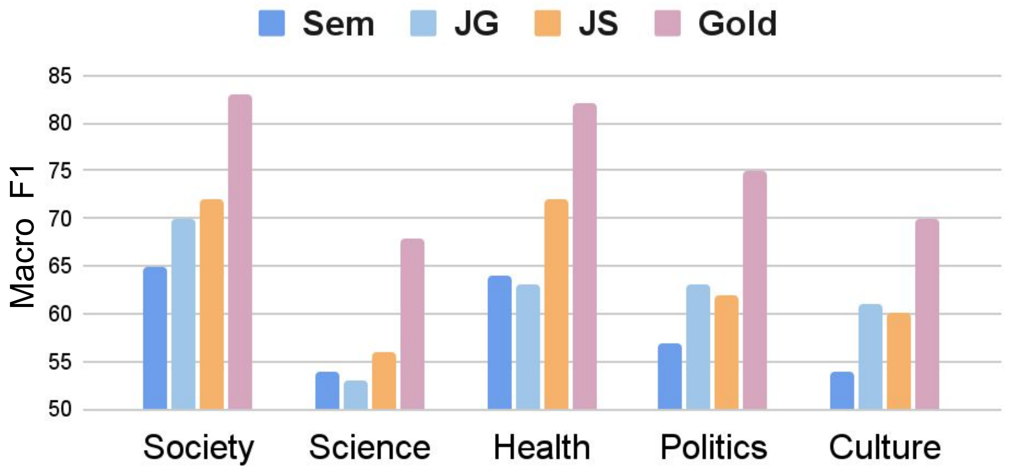


Figure: Per-domain results.

1. The scores of minor classes are much lower than the majority class.
2. Claims from science, politics and culture domains have fewer training instances as most claims in the dataset focus on the society and public health topics.

THANK YOU!

Code + Data are Available at:
http://github.com/THU-BPM/CHEF
hxm19@mails.tsinghua.edu.cn