Questionable Answers in Question Answering Research

Matt Crane
University of Waterloo

matt.crane@uwaterloo.ca

“Experiments vary greatly in goal and scale, but always rely on **repeatable procedure** and logical analysis of the results.”

Wikipedia: Experiment
Accessed: Today

“Based on theoretical reasoning it has been suggested that the reliability of findings published in the scientific literature decreases with the popularity of the research field.”

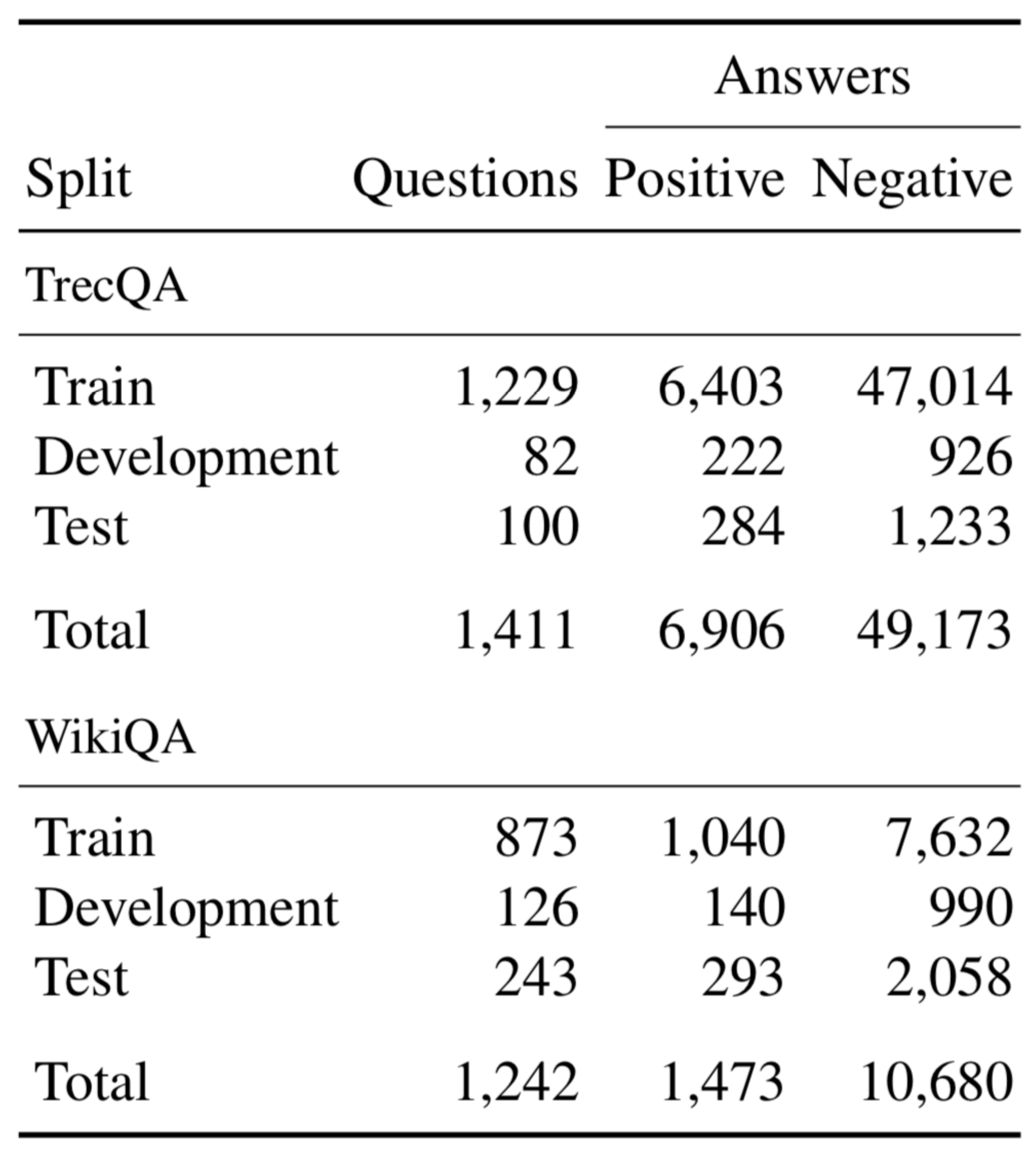
Thomas Pfeiffer and Robert Hoffmann. 2009.
Large-scale assessment of the effect of popularity on the reliability of research. *PLOS One*, 4(6):e5996.

Setup: Task & Model

- Question answering over free text: given a question and a set of candidate sentences, rank those sentences based on likelihood that the sentence contains an answer to the question
  - Example question: *what was the monetary value of the nobel peace prize in 1989 ?*
  - Example candidate sentence: *each nobel prize is worth $ 469,000 .*

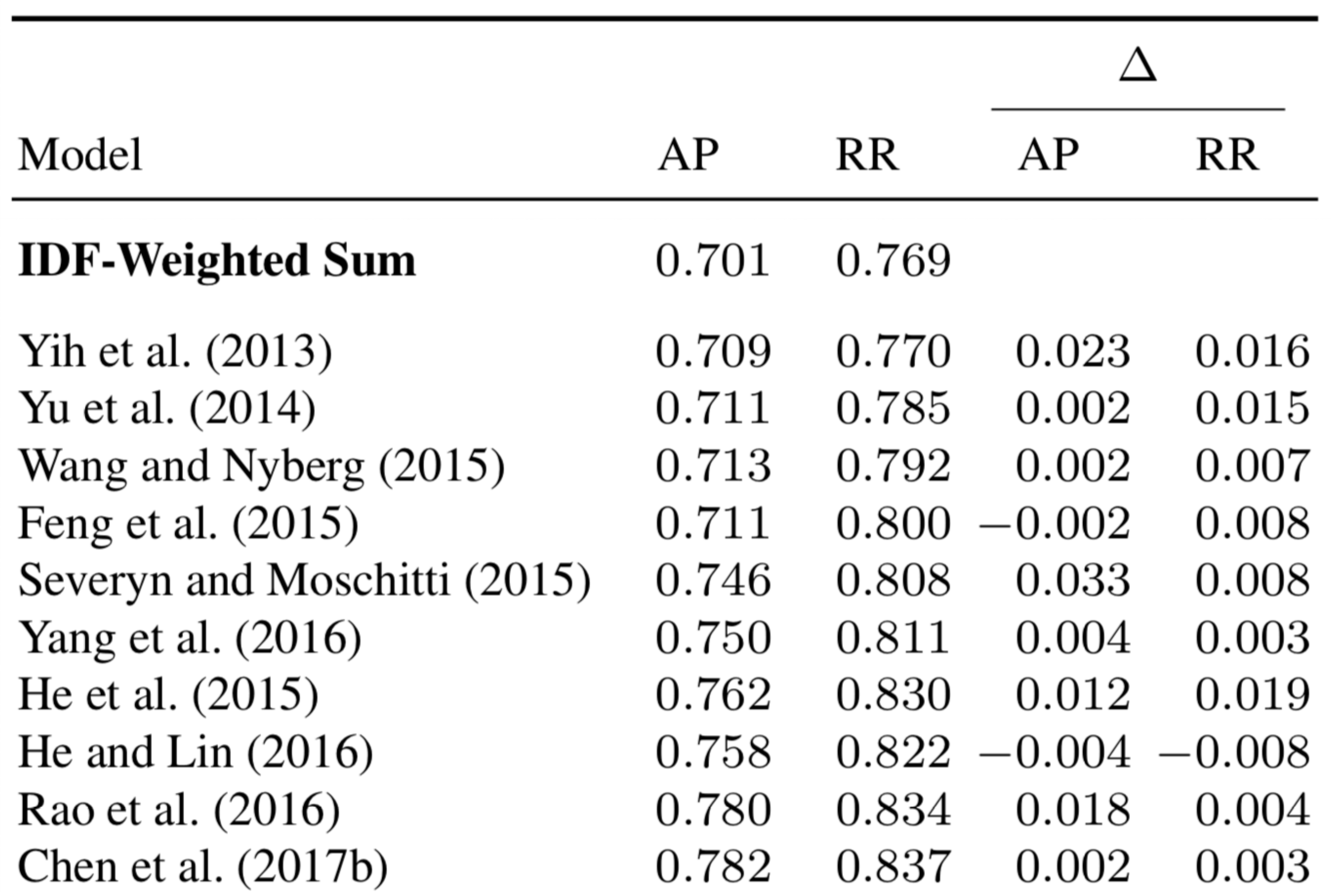
- Example model is an implementation of the Severyn & Moschitti (2015) model

Setup: Datasets



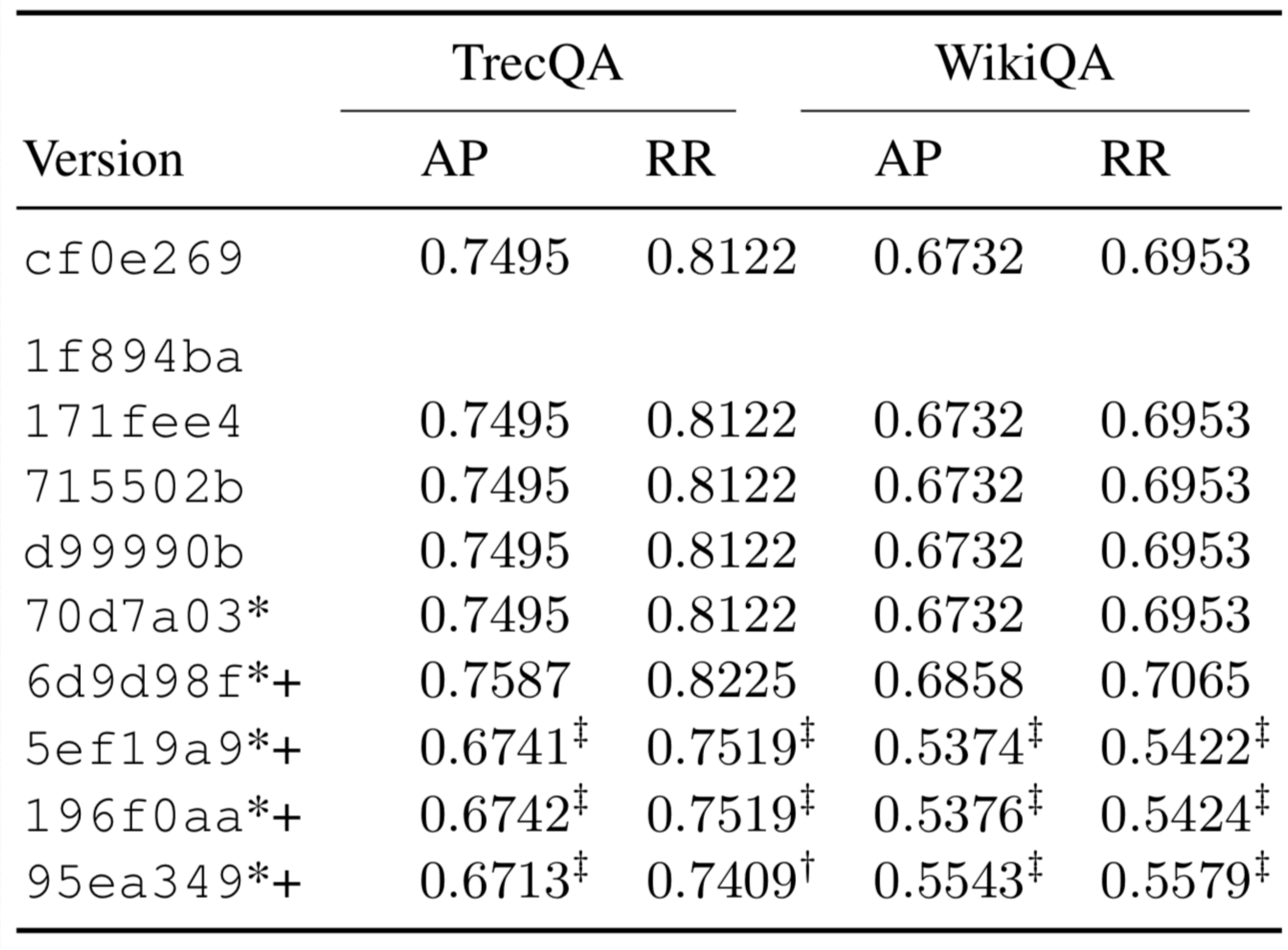
		Answers	
Split	Questions	Positive	Negative
TrecQA			
Train	1,229	6,403	47,014
Development	82	222	926
Test	100	284	1,233
Total	1,411	6,906	49,173
WikiQA			
Train	873	1,040	7,632
Development	126	140	990
Test	243	293	2,058
Total	1,242	1,473	10,680

Setup: Current Progress on TrecQA



			Δ	
Model	AP	RR	AP	RR
IDF-Weighted Sum	0.701	0.769		
Yih et al. (2013)	0.709	0.770	0.023	0.016
Yu et al. (2014)	0.711	0.785	0.002	0.015
Wang and Nyberg (2015)	0.713	0.792	0.002	0.007
Feng et al. (2015)	0.711	0.800	−0.002	0.008
Severyn and Moschitti (2015)	0.746	0.808	0.033	0.008
Yang et al. (2016)	0.750	0.811	0.004	0.003
He et al. (2015)	0.762	0.830	0.012	0.019
He and Lin (2016)	0.758	0.822	−0.004	−0.008
Rao et al. (2016)	0.780	0.834	0.018	0.004
Chen et al. (2017b)	0.782	0.837	0.002	0.003

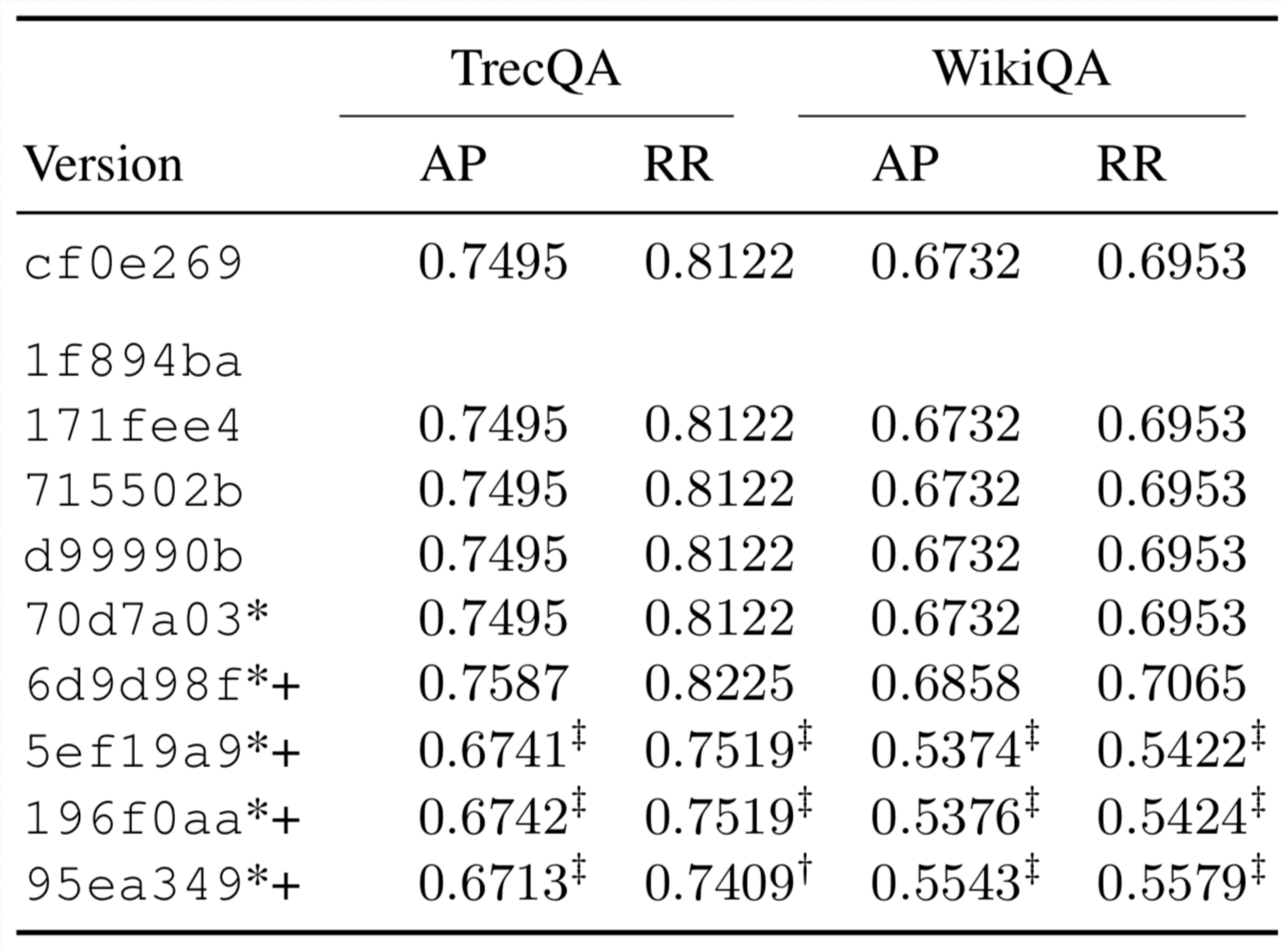
Versioning: Model Definition



	TrecQA		WikiQA	
Version	AP	RR	AP	RR
cf0e269	0.7495	0.8122	0.6732	0.6953
1f894ba				
171fee4	0.7495	0.8122	0.6732	0.6953
715502b	0.7495	0.8122	0.6732	0.6953
d99990b	0.7495	0.8122	0.6732	0.6953
70d7a03*	0.7495	0.8122	0.6732	0.6953
6d9d98f*+	0.7587	0.8225	0.6858	0.7065
5ef19a9*+	0.6741‡	0.7519‡	0.5374‡	0.5422‡
196f0aa*+	0.6742‡	0.7519‡	0.5376‡	0.5424‡
95ea349*+	0.6713‡	0.7409†	0.5543‡	0.5579‡

- Nobody writes perfect code, and when we change the code, we change the results…

Versioning: Model Definition

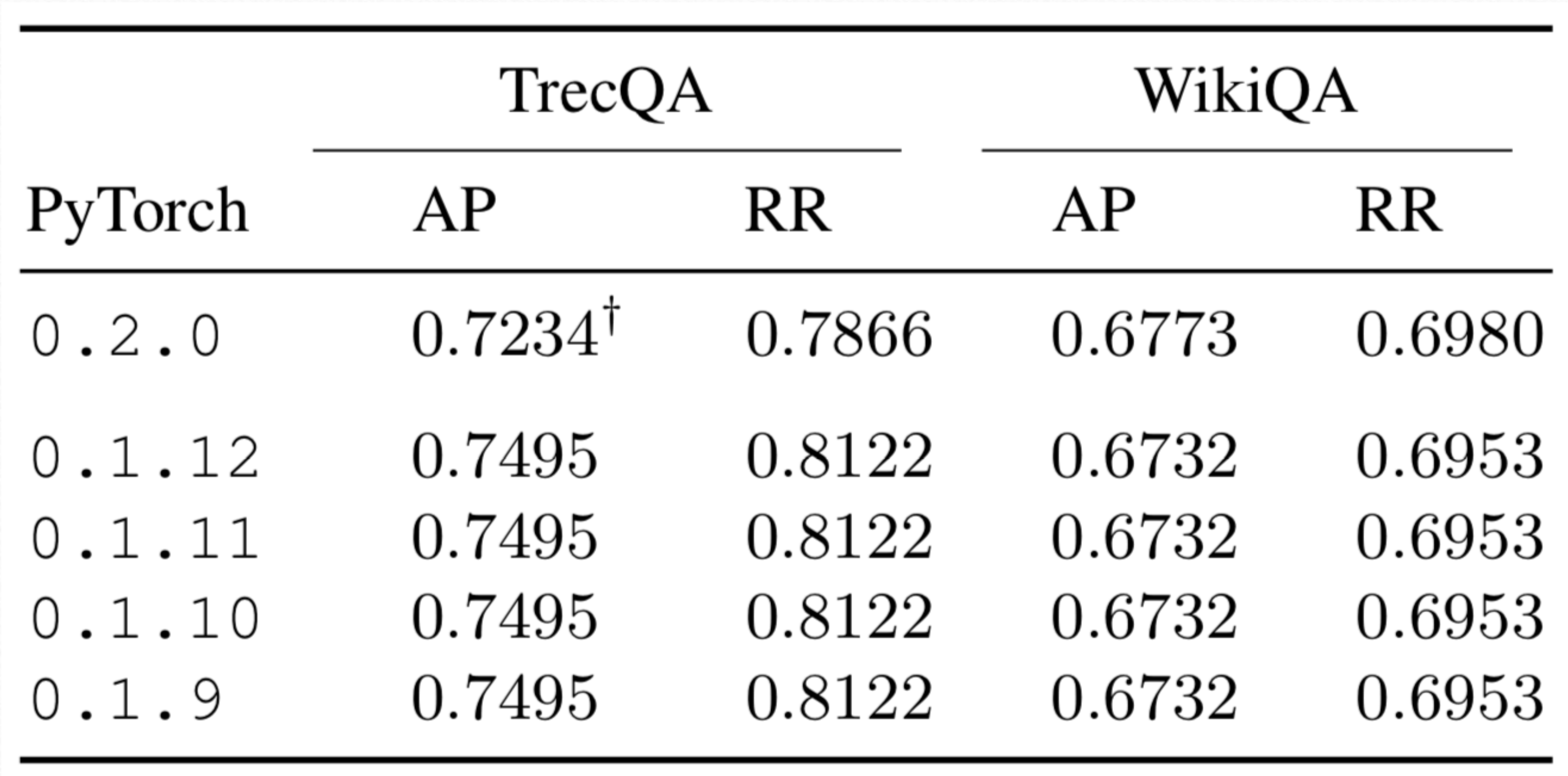


	TrecQA		WikiQA	
Version	AP	RR	AP	RR
cf0e269	0.7495	0.8122	0.6732	0.6953
1f894ba				
171fee4	0.7495	0.8122	0.6732	0.6953
715502b	0.7495	0.8122	0.6732	0.6953
d99990b	0.7495	0.8122	0.6732	0.6953
70d7a03*	0.7495	0.8122	0.6732	0.6953
6d9d98f*+	0.7587	0.8225	0.6858	0.7065
5ef19a9*+	0.6741‡	0.7519‡	0.5374‡	0.5422‡
196f0aa*+	0.6742‡	0.7519‡	0.5376‡	0.5424‡
95ea349*+	0.6713‡	0.7409†	0.5543‡	0.5579‡

- Nobody writes perfect code, and when we change the code, we change the results…

  significantly ($p < 0.01^\ddagger$, $p < 0.05^\dagger$ against `cf0e269`, paired Wilcoxon signed rank test)

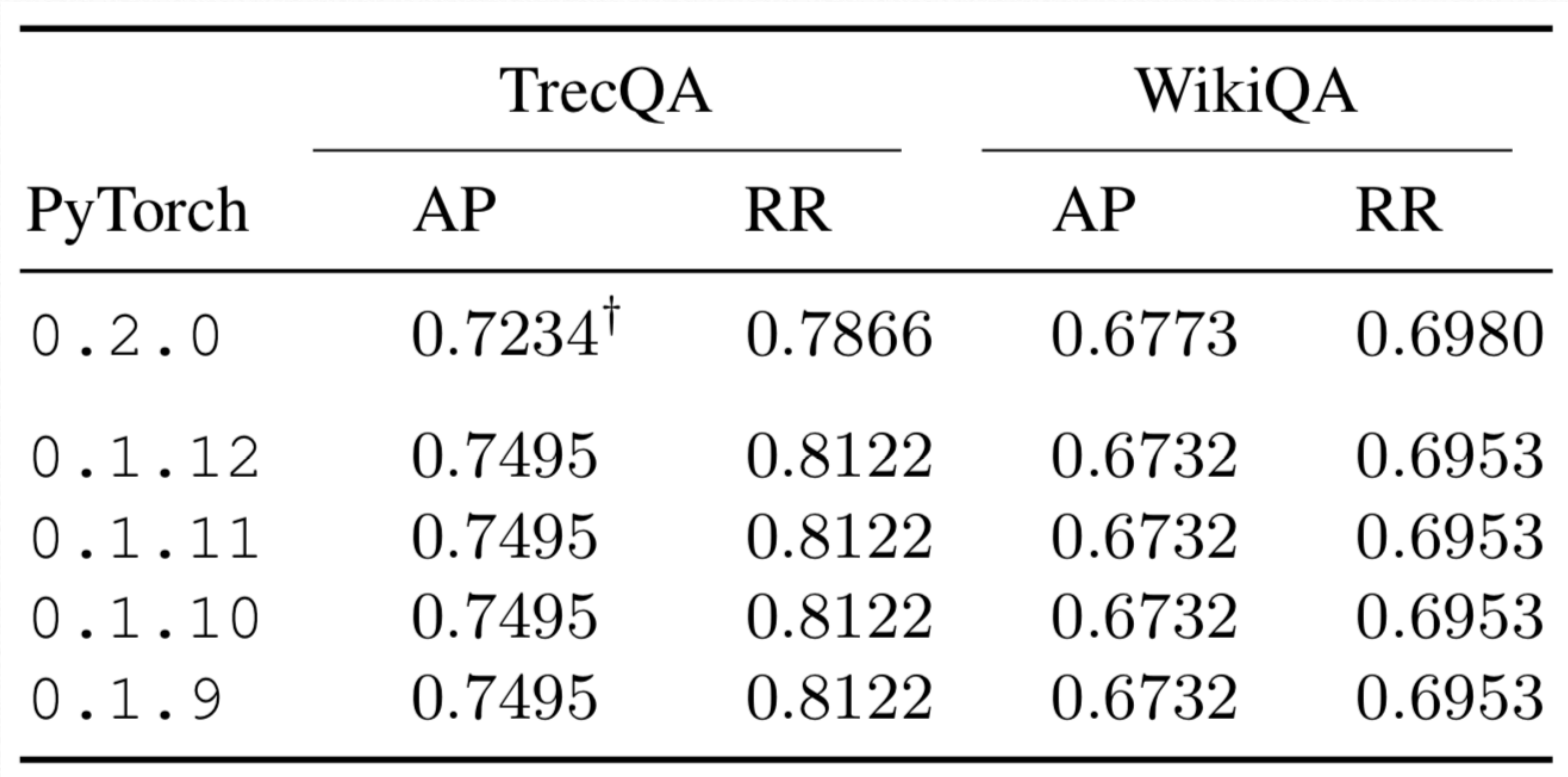
Versioning: Framework



	TrecQA		WikiQA	
PyTorch	AP	RR	AP	RR
`0.2.0`	0.7234[†]	0.7866	0.6773	0.6980
`0.1.12`	0.7495	0.8122	0.6732	0.6953
`0.1.11`	0.7495	0.8122	0.6732	0.6953
`0.1.10`	0.7495	0.8122	0.6732	0.6953
`0.1.9`	0.7495	0.8122	0.6732	0.6953

- Sometimes the framework you use makes changes, sometimes to the bits of the framework that you use…

Versioning: Framework



	TrecQA		WikiQA	
PyTorch	AP	RR	AP	RR
`0.2.0`	0.7234†	0.7866	0.6773	0.6980
`0.1.12`	0.7495	0.8122	0.6732	0.6953
`0.1.11`	0.7495	0.8122	0.6732	0.6953
`0.1.10`	0.7495	0.8122	0.6732	0.6953
`0.1.9`	0.7495	0.8122	0.6732	0.6953

- Sometimes the framework you use makes changes, sometimes to the bits of the framework that you use…

significantly ($p < 0.05^\dagger$ against `0.1.12`, paired Wilcoxon signed rank test)

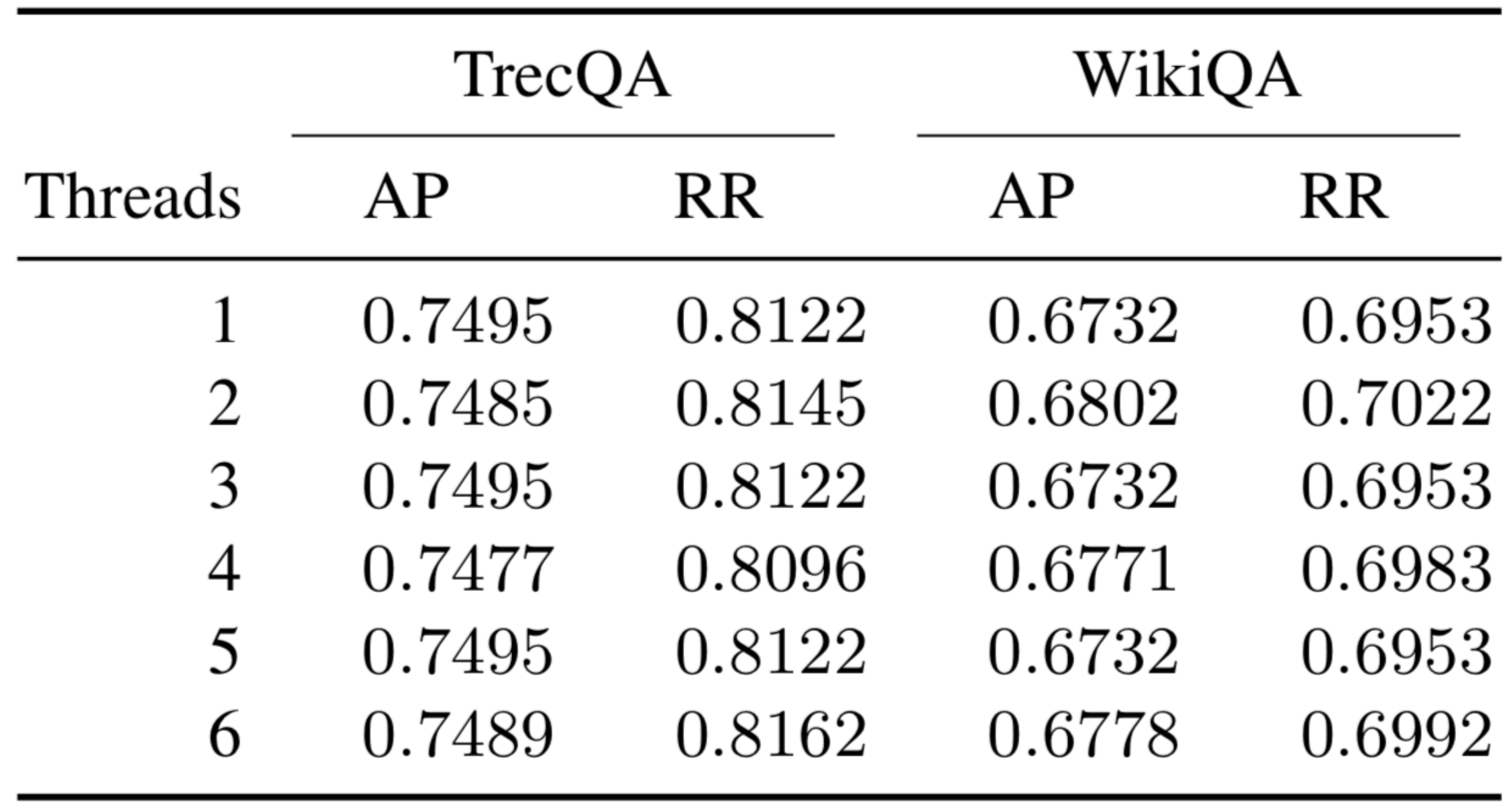
Docker?

- Docker is a containerization tool
- A container image is a lightweight, stand-alone, executable package of a piece of software that includes everything needed to run it: code, runtime, system tools, system libraries, settings
- Broadly speaking: virtual machines are to hardware what containers are to the operating system

Docker? Not Quite

- Still got different answers on different machines, the machines:
  - Intel i7-6800K (6 cores, 12 threads)
  - AMD FX-8370E (8 cores, 8 threads)
  - Intel Xeon-like on AWS EC2 (2 vCPUs)

Threading



	TrecQA		WikiQA	
Threads	AP	RR	AP	RR
1	0.7495	0.8122	0.6732	0.6953
2	0.7485	0.8145	0.6802	0.7022
3	0.7495	0.8122	0.6732	0.6953
4	0.7477	0.8096	0.6771	0.6983
5	0.7495	0.8122	0.6732	0.6953
6	0.7489	0.8162	0.6778	0.6992

- Different numbers of threads give different results, but not because of ordering, but because of workload splitting

- After fixing number of threads, now down to two answers

Hardware

- Intel gives one set of answers, AMD gives another
- Is it possible that different hardware implements the floating point specification differently?
  - Yes, but very unlikely

Hardware

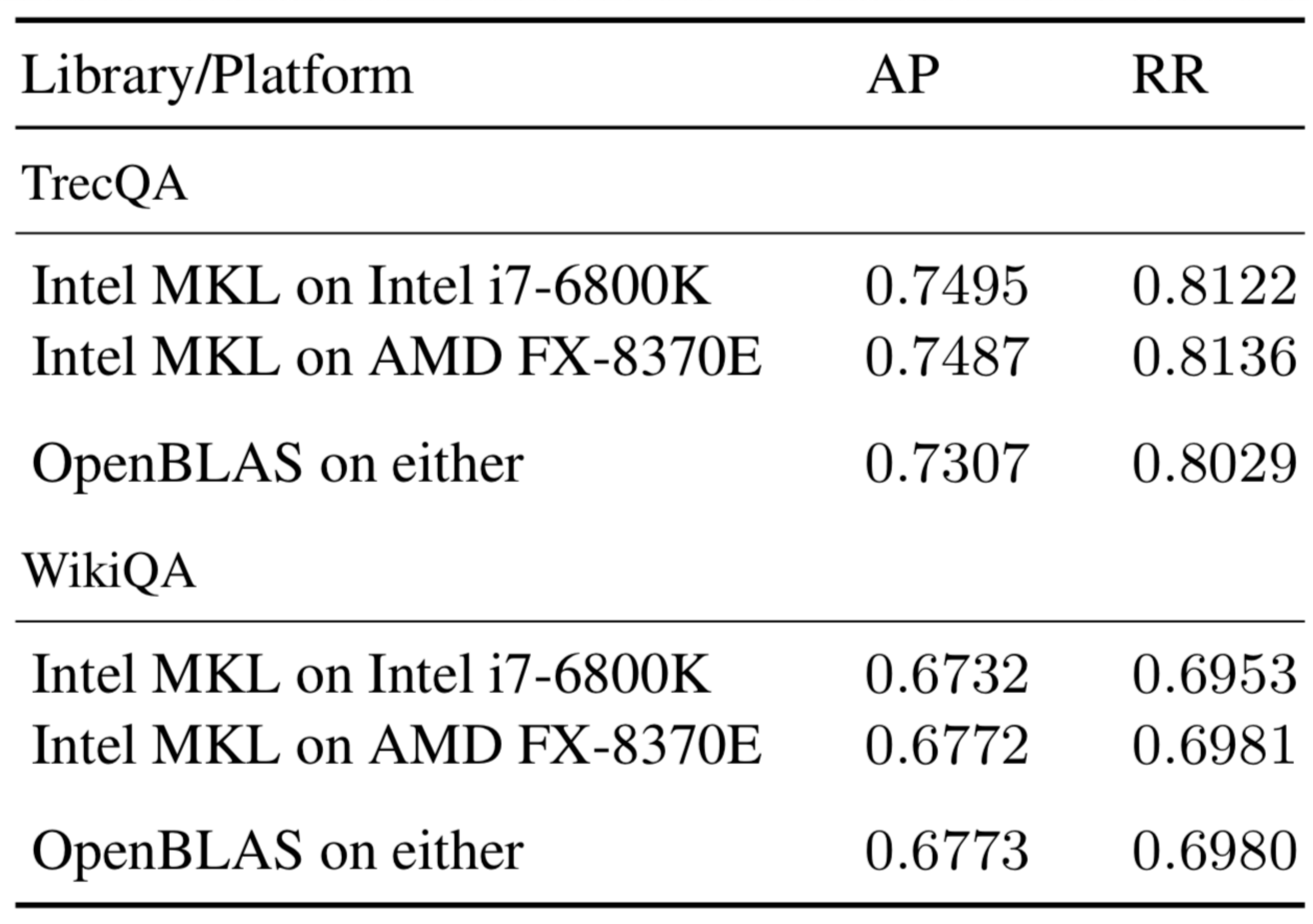
- Intel gives one set of answers, AMD gives another

- Is it possible that different hardware implements the floating point specification differently?

  - Yes, but very unlikely

- Hmmm, PyTorch ships with, and uses, *Intel’s* Math Kernel Library by default…

Hardware: A Neutral Math Library

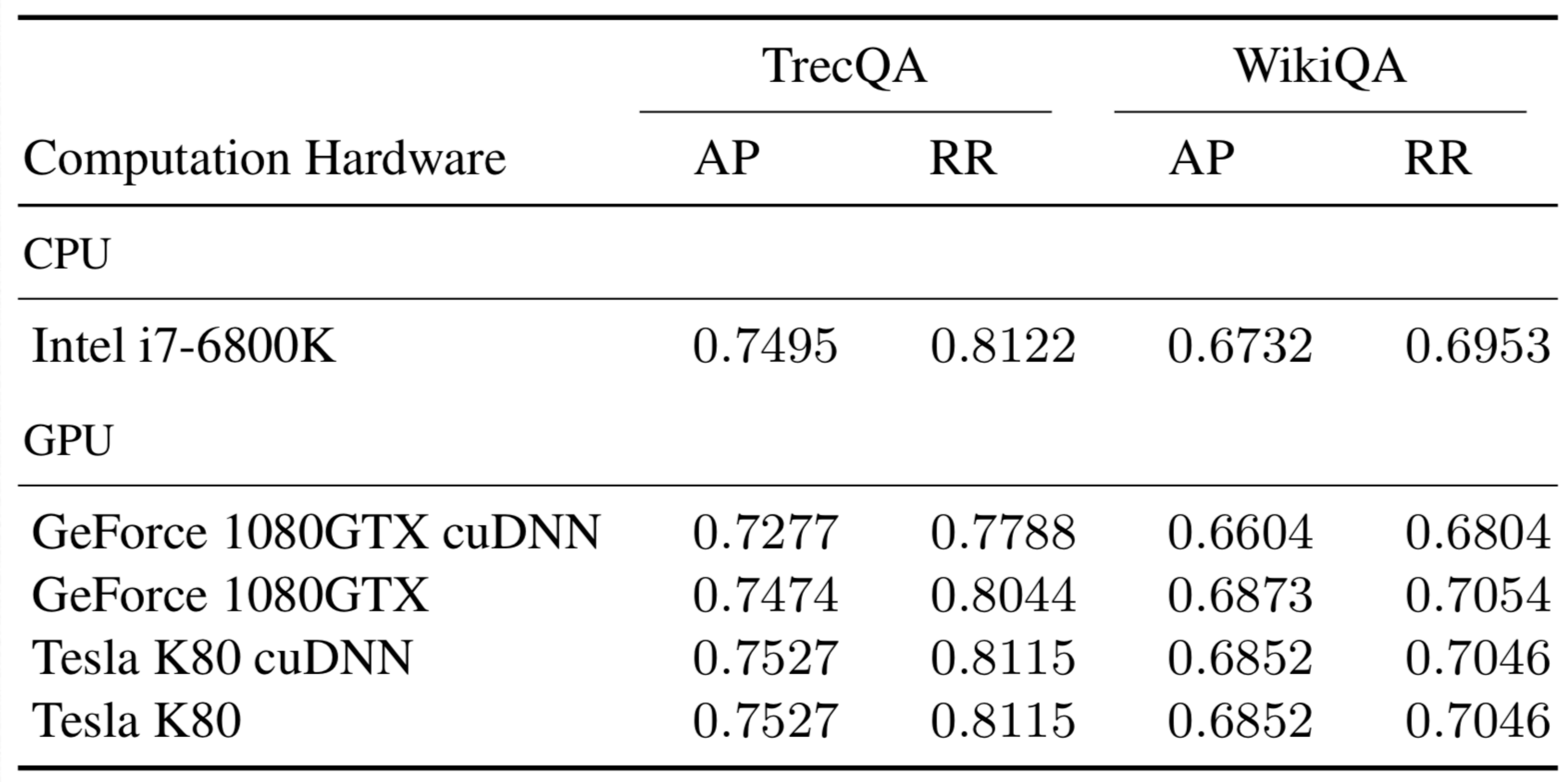


Library/Platform	AP	RR
TrecQA		
Intel MKL on Intel i7-6800K	0.7495	0.8122
Intel MKL on AMD FX-8370E	0.7487	0.8136
OpenBLAS on either	0.7307	0.8029
WikiQA		
Intel MKL on Intel i7-6800K	0.6732	0.6953
Intel MKL on AMD FX-8370E	0.6772	0.6981
OpenBLAS on either	0.6773	0.6980

Where Are We?

- Fully reproducible, repeatable, and replicable training on CPU by fixing:
  - version of model definition
  - version of framework
  - version of framework dependencies (not investigated in this case)
  - framework dependencies to be non-hardware specific
  - number of threads

What about GPU?

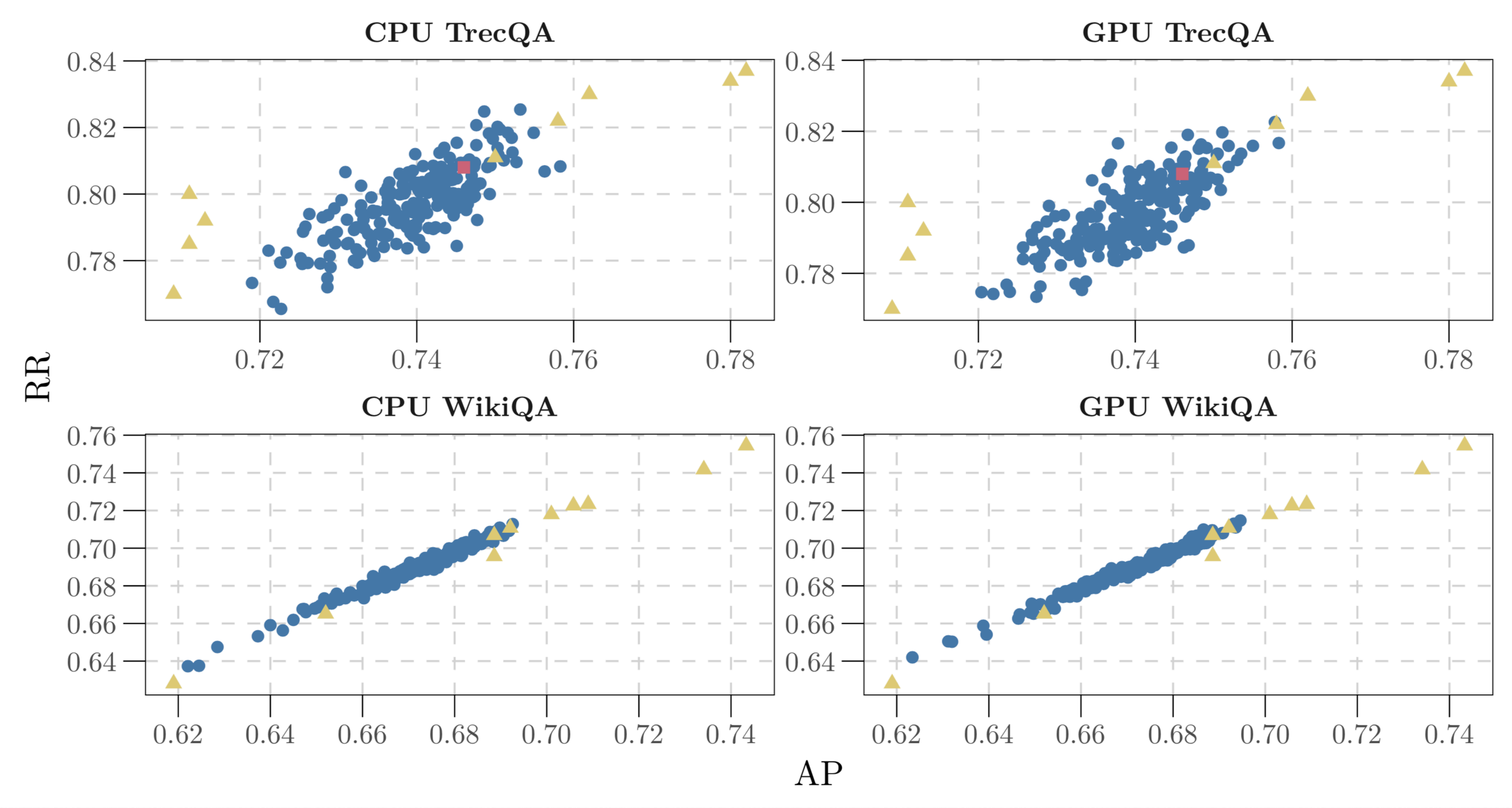


	TrecQA		WikiQA	
Computation Hardware	AP	RR	AP	RR
CPU				
Intel i7-6800K	0.7495	0.8122	0.6732	0.6953
GPU				
GeForce 1080GTX cuDNN	0.7277	0.7788	0.6604	0.6804
GeForce 1080GTX	0.7474	0.8044	0.6873	0.7054
Tesla K80 cuDNN	0.7527	0.8115	0.6852	0.7046
Tesla K80	0.7527	0.8115	0.6852	0.7046

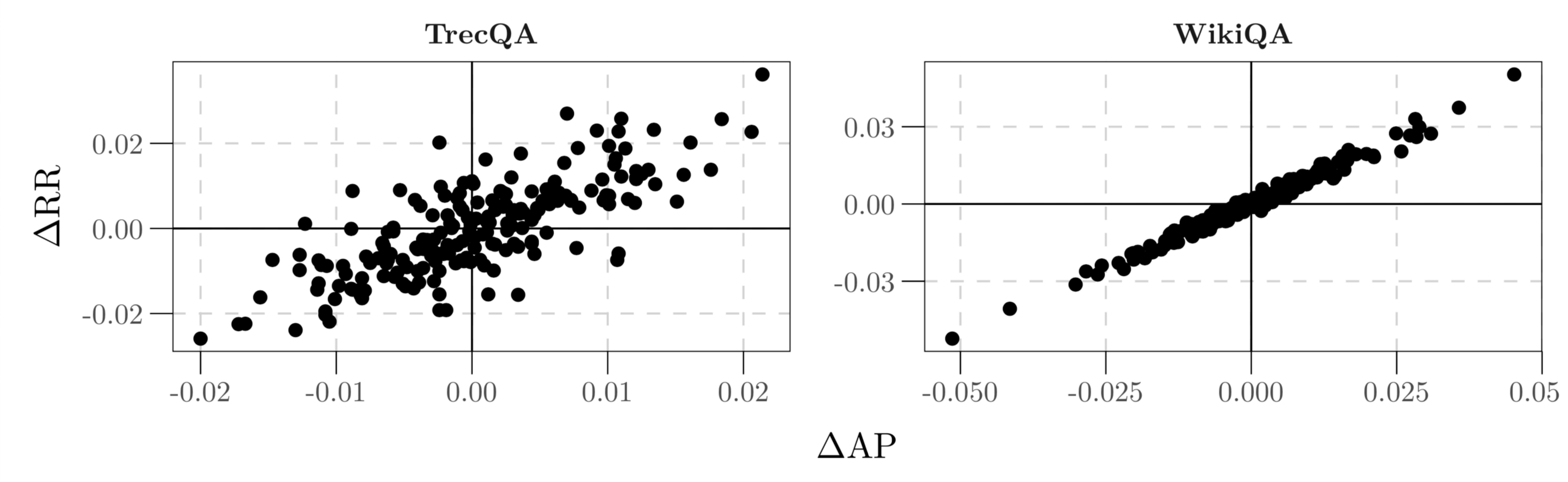
- Bajillion's of different GPUs out there, and have very little control over some aspects, as an example, we can’t fix the number of threads
- cuDNN? Enable or disable the cuDNN backend as shipped by nVidia. Has (potentially) non-reproducible kernels.

Nathan Whitehead and Alex Fit-Florea. 2011.
Precision & Performance: Floating Point and IEEE 754 Compliance for NVIDIA GPUs.

You Reap What You Sow



You Reap What You Sow: They Look Similar?



Conclusions

- All these things make a difference, and yet nobody reports them
- Nothing to really be done, if you don’t have the same hardware, then you can’t exactly reproduce the results—but at least you can compare with that caveat
- Pre-trained models are consistent—but only marginally better than believing numbers reported in a paper

Stop reporting single numbers, report on populations!