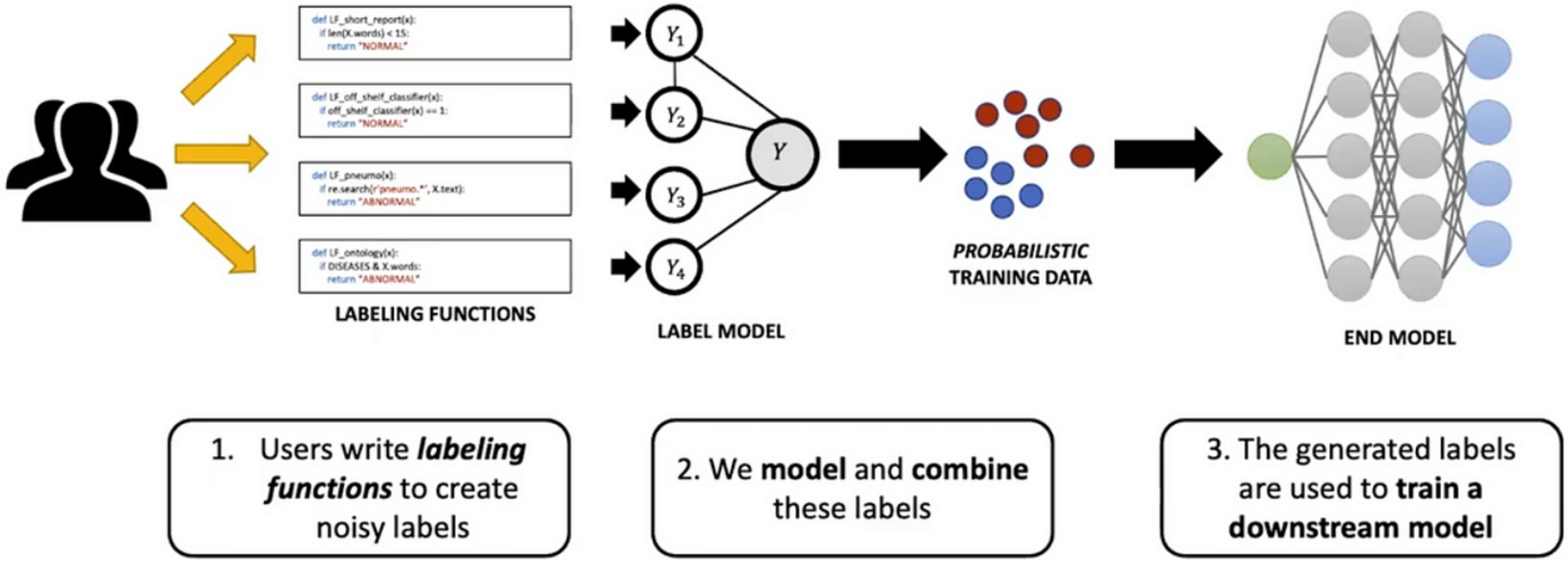
Stronger Than You Think:

Benchmarking Weak Supervision on Realistic Tasks

Tianyi Zhang*, Linrong Cai*, Jeffrey Li, Nicholas Roberts, Neel Guha, Frederic Sala

Weak Supervison (WS)



Source: Ratner, Alex. "Alex Ratner's Homepage." *ajratner.github.io*, https://ajratner.github.io/. Accessed 11 Nov. 2024.

Problems of Existing WS Benchmarks

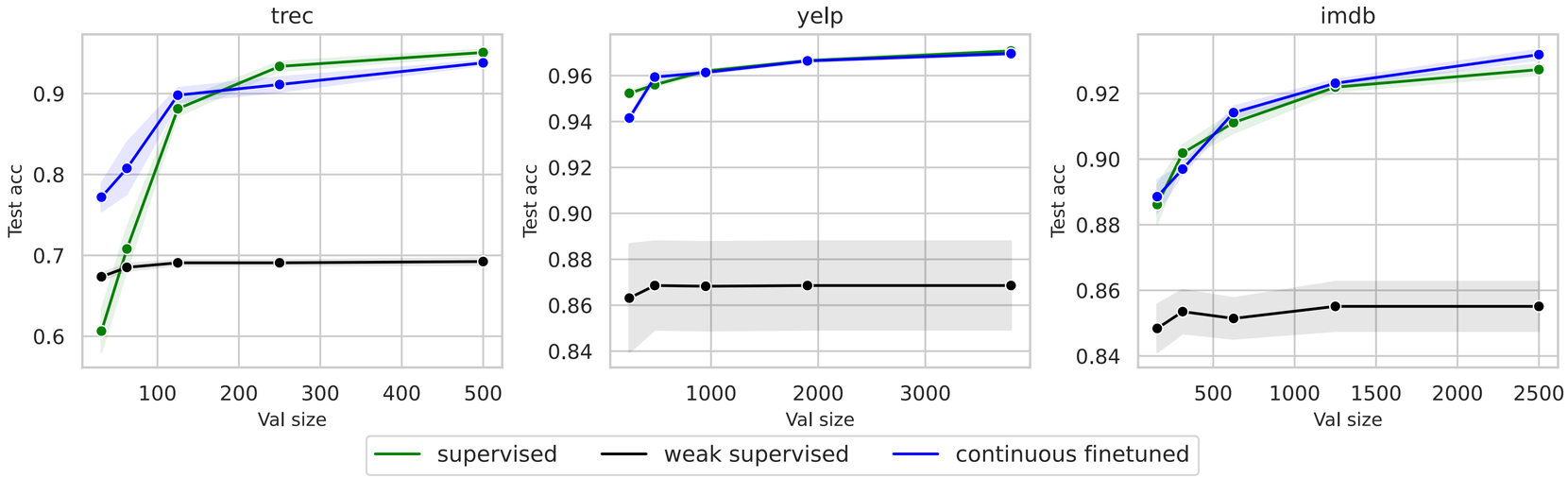
- Benchmark datasets usually have too few classes, are balanced, or aren't specialized enough to be representative of real-world tasks
- WS also depends on the quality of LFs, and LFs from current benchmarks can be improved
- Zhu et al. [1]: fine-tuning **50** manual labels can achieve comparable results against some WS

Dataset	Number of Classes
IMDB	2
ChemProt	10
TREC	6
Yelp	2
SemEval	9
AGNews	4

[1]: Dawei Zhu, Xiaoyu Shen, Marius Mosbach, Andreas Stephan, and Dietrich Klakow. Weaker than you think: A critical look at weakly supervised learning, 2023.

Existing Datasets

- On existing benchmarks, weak supervision is surpassed by hand-labeling often within < 200 labels (well below the size of the validation sets used)



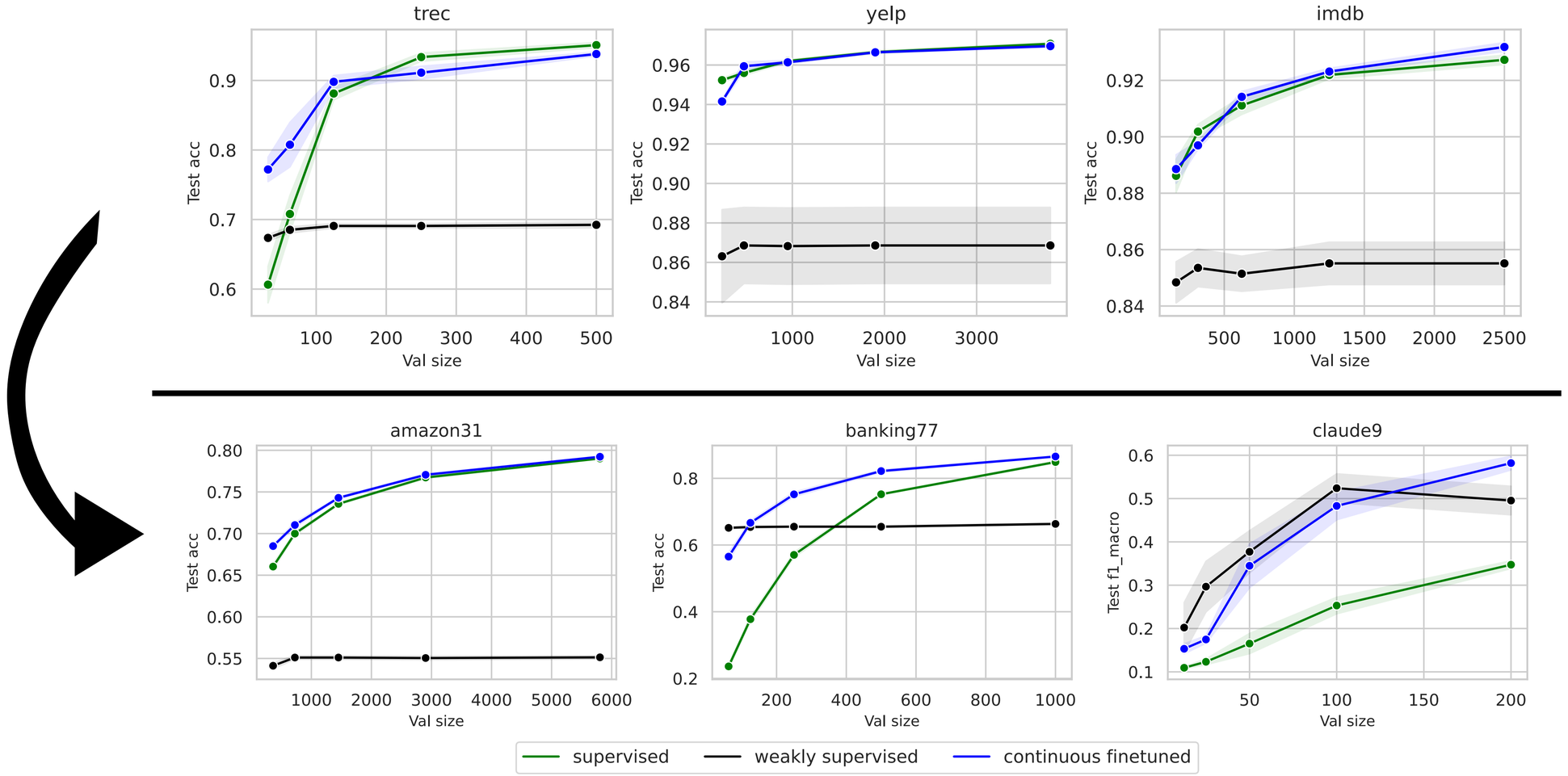
BOXWrench:

- **High-cardinality label spaces, imbalanced classes, and/or require specific domain knowledge**
- Showing that by adhering to careful LF design practices, we can write effective LFs for these tasks that can even improve upon existing benchmark LFs.

Dataset	Class	Train	Valid	Test
Banking77	77	9,003	1,000	3,080
ChemProt	10	12,600	1,607	1,607
Claude9	9	5,469	200	2057
MASSIVE{18, 60}	18, 60	11,564	3,305	1,651
Amazon31	31	131,781	5,805	17,402

Dataset	Number of Classes
IMDB	2
ChemProt	10
TREC	6
Yelp	2
SemEval	9
AGNews	4
Banking77	77
Claude9	9
MASSIVE18	18
MASSIVE60	60
Amazon31	31

BOXWRENCH Datasets



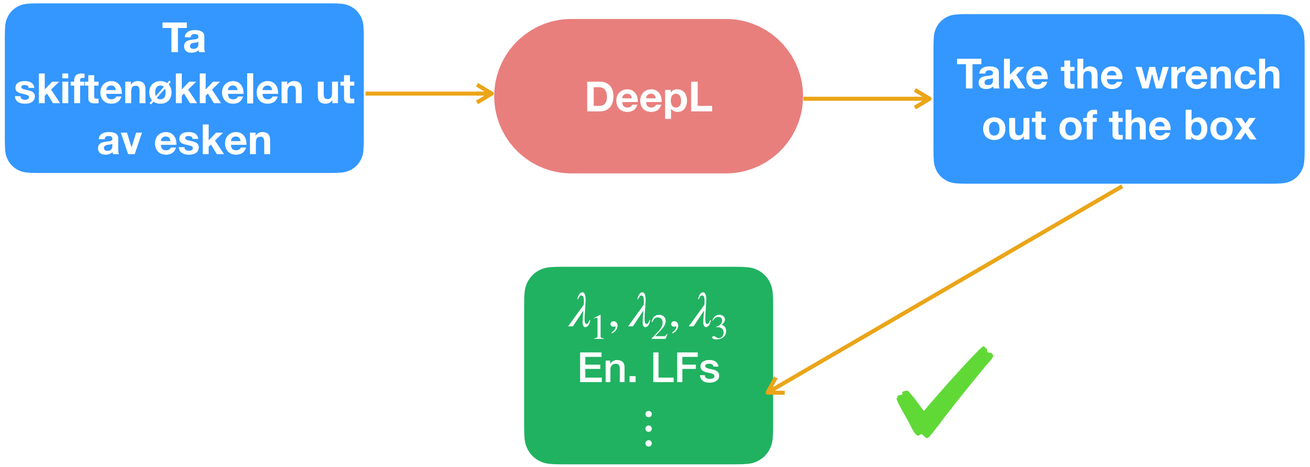
CFT vs. Supervised

- **CFT** outperforms Supervised with a clear margin
- **Majority vote** and **DavidSkene** consistently performed well

	Claude9	ChemProt[1]	MASSIVE18
6.25% Validation Size			
+Majority vote	0.1532±0.0256	**0.6246±0.0181**	**0.8110±0.0042**
+DawidSkene	**0.1533±0.0251**	0.6100±0.0199	0.7968±0.0103
+Snorkel	0.1322±0.0217	0.6168±0.0116	0.8106±0.0074
+FlyingSquid	0.1090±0.0046	0.6058±0.0209	0.7507±0.0109
+Supervised Only	0.1093±0.0048	0.5037±0.0296	0.7531±0.0136
12.5% Validation Size			
+Majority vote	0.1747±0.0198	**0.6850±0.0104**	0.8448±0.0060
+DawidSkene	**0.1840±0.0220**	0.6540±0.0244	0.8357±0.0038
+Snorkel	0.1565±0.0159	0.6485±0.0147	**0.8451±0.0048**
+FlyingSquid	0.1239±0.0143	0.6395±0.0152	0.8181±0.0092
+Supervised Only	0.1232±0.0150	0.5854±0.0219	0.8203±0.0128
25% Validation Size			
+Majority vote	0.3448±0.1136	**0.7263±0.0210**	0.8625±0.0040
+DawidSkene	**0.3623±0.1074**	0.7119±0.0136	**0.8626±0.0070**
+Snorkel	0.3028±0.0898	0.7114±0.0165	0.8585±0.0021
+FlyingSquid	0.1700±0.0248	0.6957±0.0101	0.8563±0.0086
+Supervised Only	0.1653±0.0521	0.6836±0.0234	0.8545±0.0095
50% Validation Size			
+Majority vote	**0.4830±0.0714**	**0.7783±0.0052**	0.8824±0.0052
+DawidSkene	0.4769±0.0744	0.7749±0.0027	0.8843±0.0047
+Snorkel	0.4624±0.0649	0.7669±0.0090	**0.8853±0.0054**
+FlyingSquid	0.2319±0.0565	0.7618±0.0075	0.8795±0.0049
+Supervised Only	0.2531±0.0429	0.7731±0.0099	0.8825±0.0042
100% Validation Size			
+Majority vote	**0.5819±0.0357**	**0.8199±0.0041**	**0.8993±0.0021**
+DawidSkene	0.5724±0.0377	0.8197±0.0050	0.8932±0.0019
+Snorkel	0.5573±0.0195	0.8143±0.0041	0.8988±0.0023
+FlyingSquid	0.3521±0.0374	0.8133±0.0060	0.8943±0.0050
+Supervised Only	0.3473±0.0201	0.8162±0.0065	0.8975±0.0029

Adaptability of LFs

- A key advantage of WS is that LFs can be **adapted** across task specifications.
- We study one example via the MASSIVE classification datasets
  - MASSIVE contains **52 parallel** versions in different languages
  - LFs written for English can be **reused** for other languages

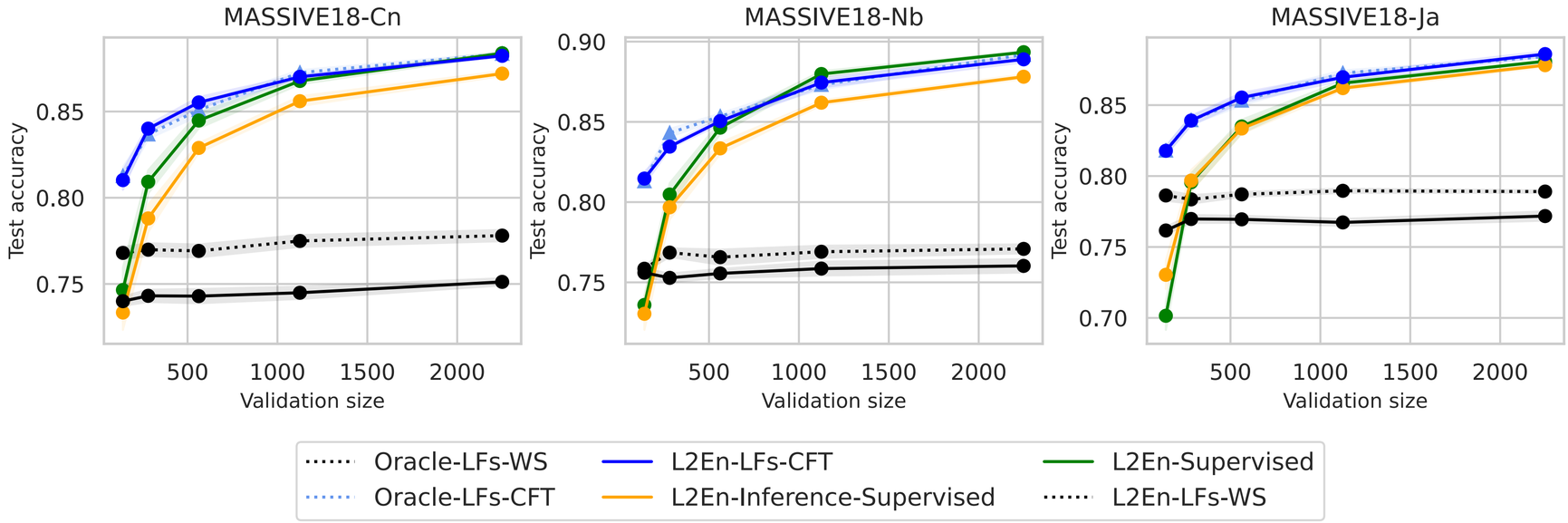



```
def lf1(x):
    return 1 if x.startswith('take') else -1
```

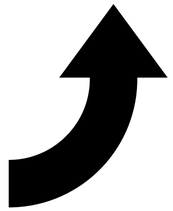
- *L2En -LFs: this method provides a more realistic scenario for reusing the English LFs.*

Generalization of existing LFs on Multi-lingual Massive

- **Green Solid vs Yellow Solid:** importance of having a language specific model.
- **Blue Solid vs Light Blue Dotted:** adaptability of WS



Thank you for listening!



Checkout our paper throught the QR code📄