AWER - A Framework for Automated Worker Evaluation Based on Free-Text Responses with No Ground Truth

Inbal Yahav

With Tomer Geva and Anat Goldstein

Coller School of Management
Tel Aviv University

Coller Lab for Artificial Intelligence and Business Analytics

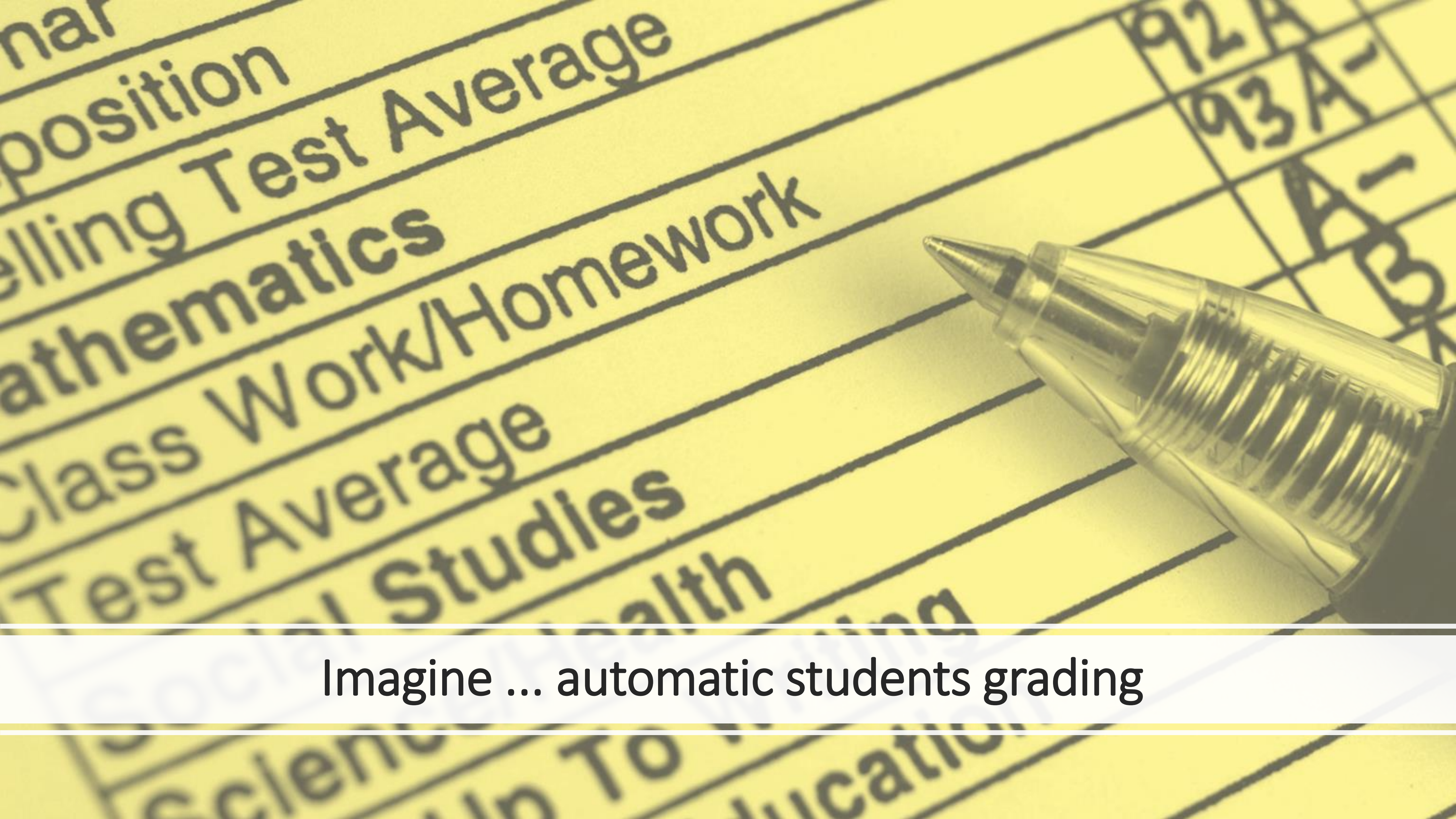
Settings and goals

M individuals answer N open questions.

Ground truth is not available.

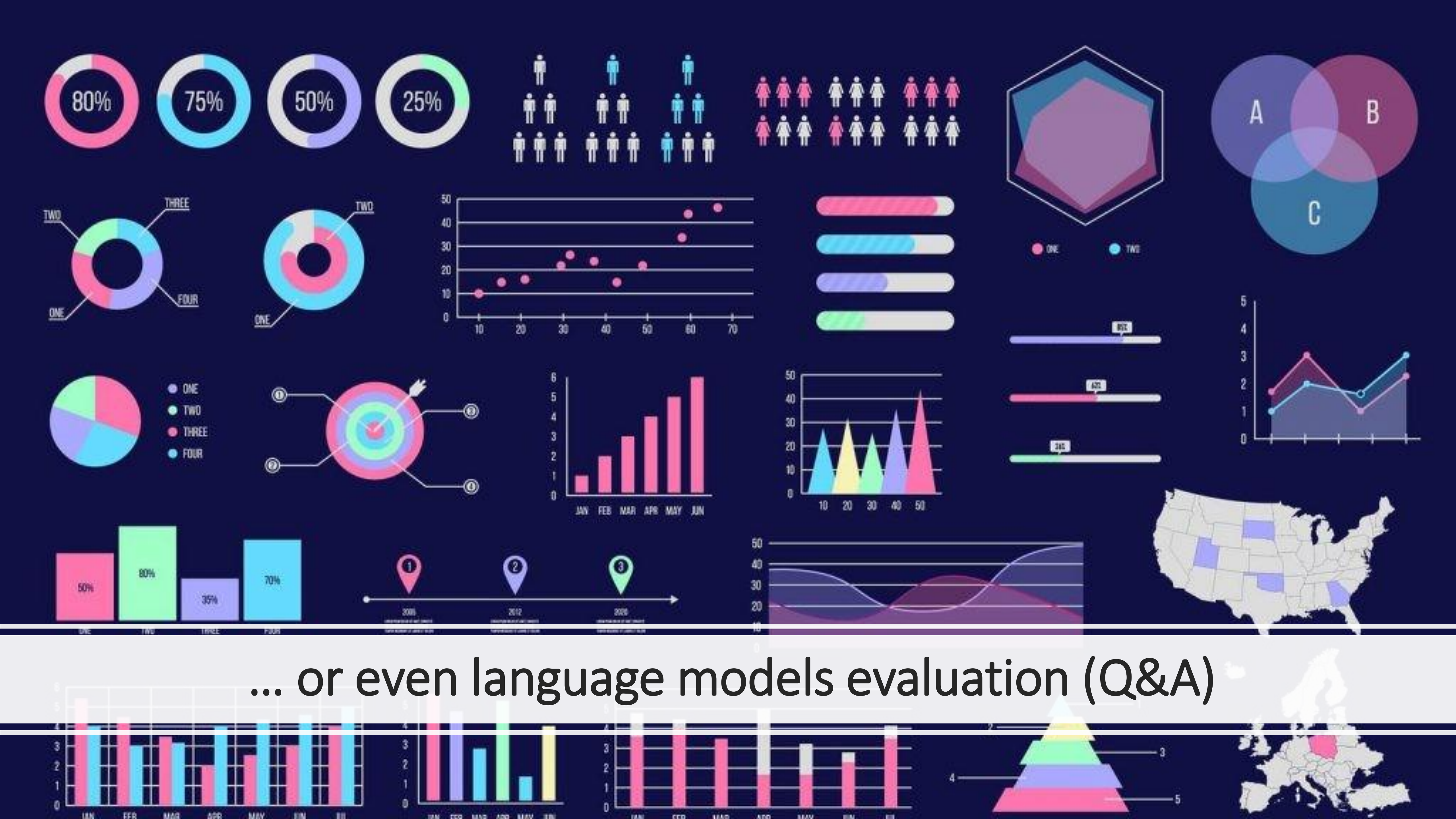
How do we grade (or rank) them?

What is the correct answer?

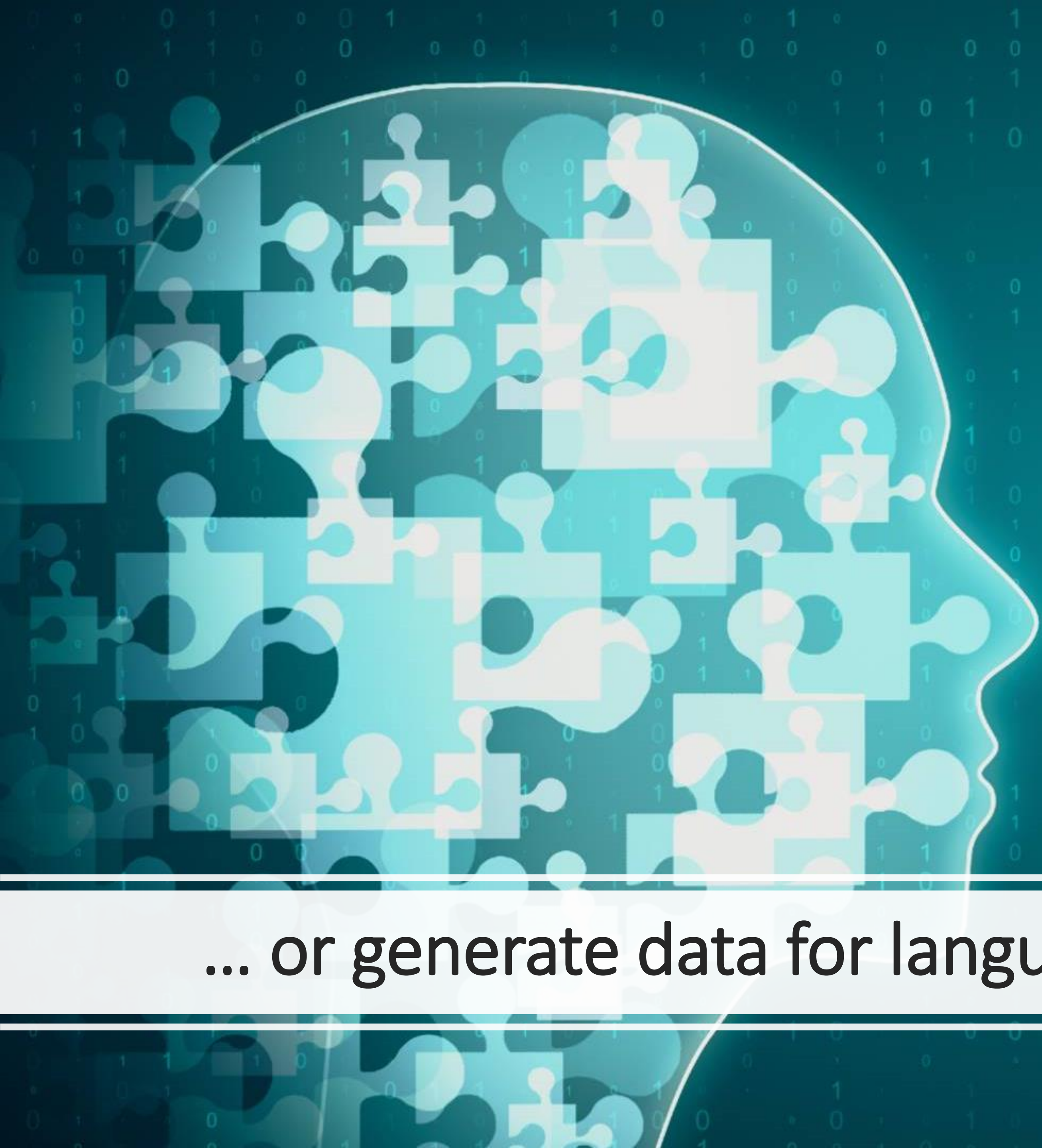


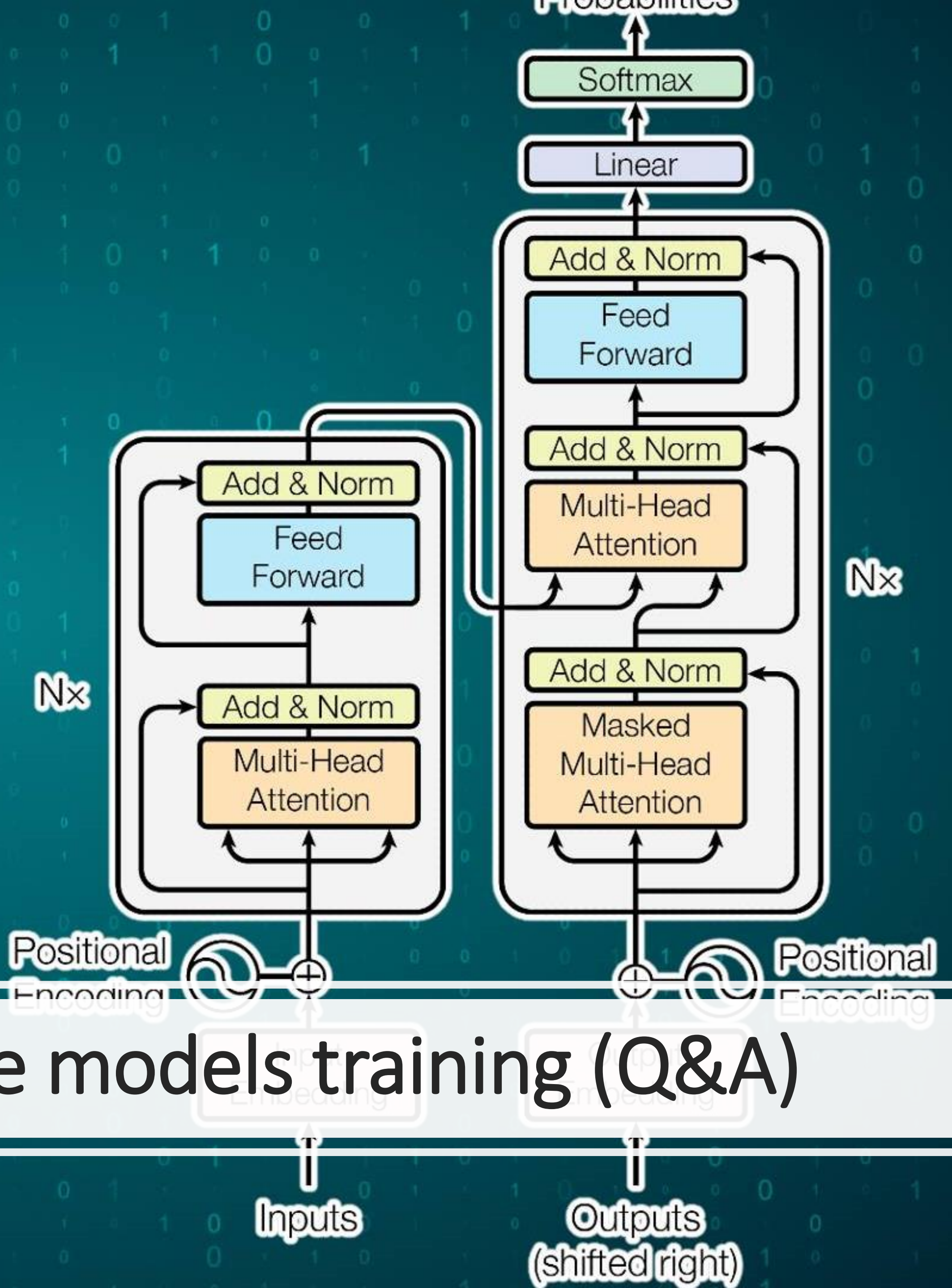
Imagine … automatic students grading

… or workers/ experts evaluation



… or even language models evaluation (Q&A)





… or generate data for language models training (Q&A)

Back to settings and goals

M workers answer N open questions.

Ground truth is not available.

	Worker #1	Worker #2	Worker #3	....	....	....	Worker #M
Question #1	Due to electricity issues	It will be too old as over 40	It won't have enough power left	Beyond reasonable repair	it will run out of electricity	A generator will stop working	As it will no longer have enough
Question #2	The viable atmosphere.	Environment atmosphere	Because the Weather conditions	Size and activity	Electric fields in the atmosphere	Weather and magnetic fields	Due to the Weather atmosphere, ……
Question #3	kitchen grade aluminium foil	Use of kitchen grade foil wrapped around t…..	Because of the Aluminium foil	Foil wrapped around wires.	Tin foil	Foil wrapped around wires.	Kitchen foil
....	**....**	**....**	**....**	**....**	**....**	**....**	**....**
....	**....**	**....**	**....**	**....**	**....**	**....**	**....**
....	**....**	**....**	**....**	**....**	**....**	**....**	**....**
Question #n	Because of elevation levels/difficulty	To make sure people train for them……	Varies depending on the course……	Depending on the course	it depends on the demands of the particular course	Depending on which course they take,	Dependant on the terrain/course

Back to settings and goals

(1) automatically assign a score to each worker according to the average **correctness** of her responses.

(2) automatically extract the correct answer for each question.

Related literature

- Automatic workers evaluation w/g ground truth: focus on *binary, numeric, or multi-category output* (e.g., Geva & Saar Tsechansky, 2021; Wang et al., 2017; Yin et al., 2021)
- Automatic **question** evaluation w/g ground truth: single work, used as baseline (Roy et al., 2016)
- Automatic Short Answer Grading (ASAG): focus on grading, when ground truth exists (e.g., Burrows et al., 2015 ; Bonthu et al., 2021)

The AWER Framework

Part 1:

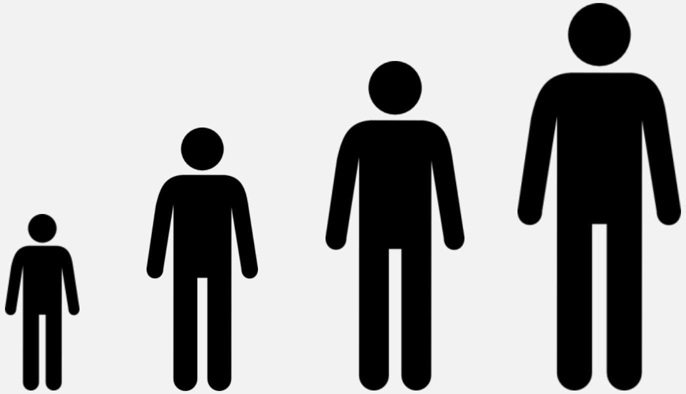
“The wisdom of the crowd”

a multidimensional voting scheme

Part 2:

“The wisdom of the wise”

an iterative re-weighting algorithm (adapted Expectation Maximization-based solution)



Multidimensional Voting

Represent **each** response as a textual vector

Example:

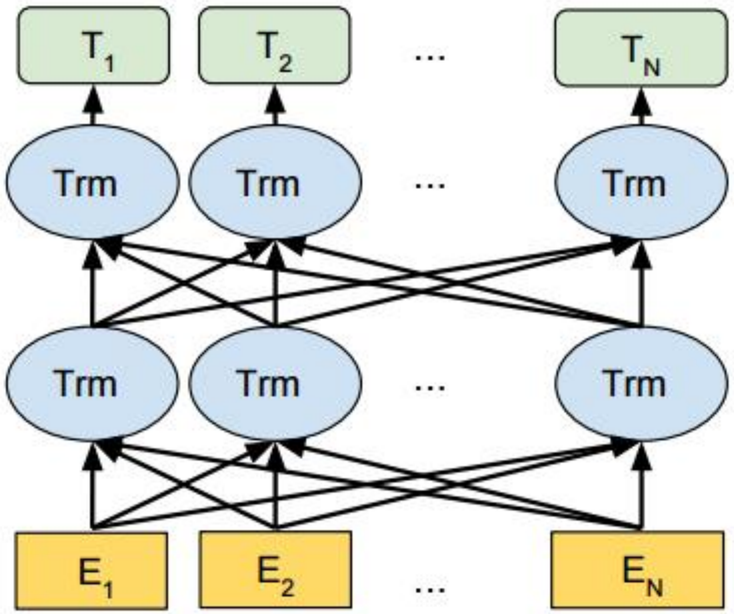
Question: “Near which planets did Voyager 1 make a flyby?“

Response: "Made a flyby next to Saturn and Jupiter."

k=1 **Saturn flyby**	k=2 **Mars flyby**	k=3 **Jupiter flyby**	……	k=K **Neptune flyby**
1	0	1		0

Note

In practice, we represent responses as embedding vectors



Multidimensional Voting

Represent **multiple** responses for a given question in a matrix

Example:

Question: “Near which planets did Voyager 1 make a flyby?“

	k=1 **Saturn flyby**	k=2 **Mars flyby**	k=3 **Jupiter flyby**	……	k=K **Neptune flyby**
Response 1	1	0	1	…….	0
Response 2	1	1	0	…….	0
…….	…….	…….	…….	…….	…….
Response M	1	0	1	…….	1

Multidimensional Voting

Compute majority vote

Example:

Question: “Near which planets did Voyager 1 make a flyby?“

	k=1 **Saturn flyby**	k=2 **Mars flyby**	k=3 **Jupiter flyby**	……	k=K **Neptune flyby**
Response 1	1	0	1	…….	0
Response 2	1	1	0	…….	0
…….	…….	…….	…….	…….	…….
Response M	1	0	1	…….	1
Majority vote	1	0	1		0

Multidimensional Voting

Compute majority vote

Example:

Question: “Near which planets did Voyager 1 make a flyby?“

	k=1 **Saturn flyby**	k=2 **Mars flyby**	k=3 **Jupiter flyby**	……	k=K **Neptune flyby**
Response 1	1	0	1	…….	0
Response 2	1	1	0	…….	0
…….	…….	…….	…….	…….	…….
Response M	1	0	1		1

Synthetic Exemplary Answer (SEA)

Multidimensional Voting

Why use a majority vote?

- Under the assumptions:
  - Workers are **independent**
  - Workers are **weak classifiers**, *for each vector element k*

→ The number of correct votes for vector element *k*, V_k, follows a **binomial distribution**

→ $\Pr\left(V_k > \frac{M}{2}\right) \rightarrow 1$ as M gets large

Multidimensional Voting

Compute the similarity between the worker's answer to the SEA,

And set:

Correctness (single question) ~ similarity

Grade ~ average correctness across all questions

SEA

k=1	k=2	k=3	……	k=K
1	0	1		0

Cosine similarity

Response

k=1	k=2	k=3	……	k=K
0	0	1		1

Iterative Re-Weighting

“Wisdom of the wise” – reweighing workers based on assessing their capabilities

<u>Iteratively</u>:

- for each question: update the voting weight of worker w_i according to the estimated workers’ grade (from a previous iteration) [Initialize: $weight_i = 1$]
- Recompute SEA, correctness, and grades

Illustration

Framework summary

0. Represent each response $R_{\mathrm{i,j}}$ ($R_{\mathrm{i,j}}$ is the response by worker W_{i} to question Q_{j}) as a vector, $\overrightarrow{text}_{i,j}$

1. For each question Q_{j} Obtain an initial estimate of $\overrightarrow{SEA}_j$ by applying an equally weighted voting mechanism on $\overrightarrow{text}_{i,j}\ \forall i$

Iterate steps 2-3 below until convergence:

2. For each $\overrightarrow{text}_{i,j}$ (representing $R_{\mathrm{i,j}}$) compute $S_{\mathrm{i,j}}$ - the similarity of $\overrightarrow{text}_{i,j}$ to the corresponding $\overrightarrow{SEA}_j$; Set the corresponding $grade_i = f(\frac{1}{n} * \sum_{j=1}^{n} S_{\mathrm{i,j}})$, where f is a normalization function across all workers' average scores.

3. For each question Q_{j}, apply a (re-)weighted voting mechanism on the numerical vectors representing the responses to generate a new *Synthetic Exemplary Answer* ($\overrightarrow{SEA}_j$) vector. Each worker's W_{i} voting $weight_i$ is proportional to the worker's estimated $grade_i$.

4. Output $grade_i\ \forall W_{\mathrm{i}}$

Modular Implementation

- Textual Representation (step 0) can be implemented using various methods such as Transformer-based embeddings, BOW, TF-IDF, etc.
- Similarity/distance (step 2): can be implemented using various measures such as Cosine similarity, Euclidian distance, or entailment

Empirical Evaluation

Three datasets:

- Computer Science course Q&A (Mohler et al., 2011): semi-synthetic simulation to define "workers"
- Purposely compiled datasets: online workers' responses to questions on Wikipedia articles (40 workers, 15 questions in each dataset). Workers recruited via Prolific.com.
- Pure numerical simulation: used to examine "special conditions"
- Baseline: Roy et al., 2016.

Main results

Semi-synthetic simulation (CS data)

Settings	Baseline	AWER	%Improvement: AWER vs. Baseline
2 quality groups; 10 workers per group	0.935	0.979	4.7%***
4 quality groups; 5 workers per group	0.941	0.978	4.0%***
10 quality groups; 2 workers per group	0.925	0.962	4.0%***

Pearson correlation values are between the model-based evaluation and the average score of two expert evaluators
*** P value < .01

Purposely compiled datasets (Wikipedia data)

Dataset	Baseline	AWER	%Improvement: AWER vs. Baseline
Movies and History	0.779	0.915	17.5***
Science / Technology and Sports	0.850	0.950	11.8%***

Bootstrap P values *** P value < .01
Pearson correlation values are between the model-based evaluation and the average score of two expert evaluators

Numerical simulation

- Three levels of workers: high, medium, and low (weak learners)
- Two types of questions: standard (majority correct), and challenging (majority incorrect among medium and low-level workers)

- **We vary:**
  - The % correct answer per worker
  - The ratio of challenging questions

Results highlights

- AWER framework provides **accurate evaluation** even when in (up to) ~**40%** of the questions the majority of responders provide **similar incorrect responses**.
- AWER framework provides **accurate evaluation** even when the **average worker correctness is only slightly above 50%.**

Ablation Study

Impact of iterative re-weighting (CS data)

Settings	Wisdom of the crowd	Wisdom of the wise	%Improvement: AWER vs. Baseline
2 quality groups; 10 workers per group	0.965	0.979	1.4%***
4 quality groups; 5 workers per group	0.964	0.978	1.5%***
10 quality groups; 2 workers per group	0.943	0.962	2.0%***

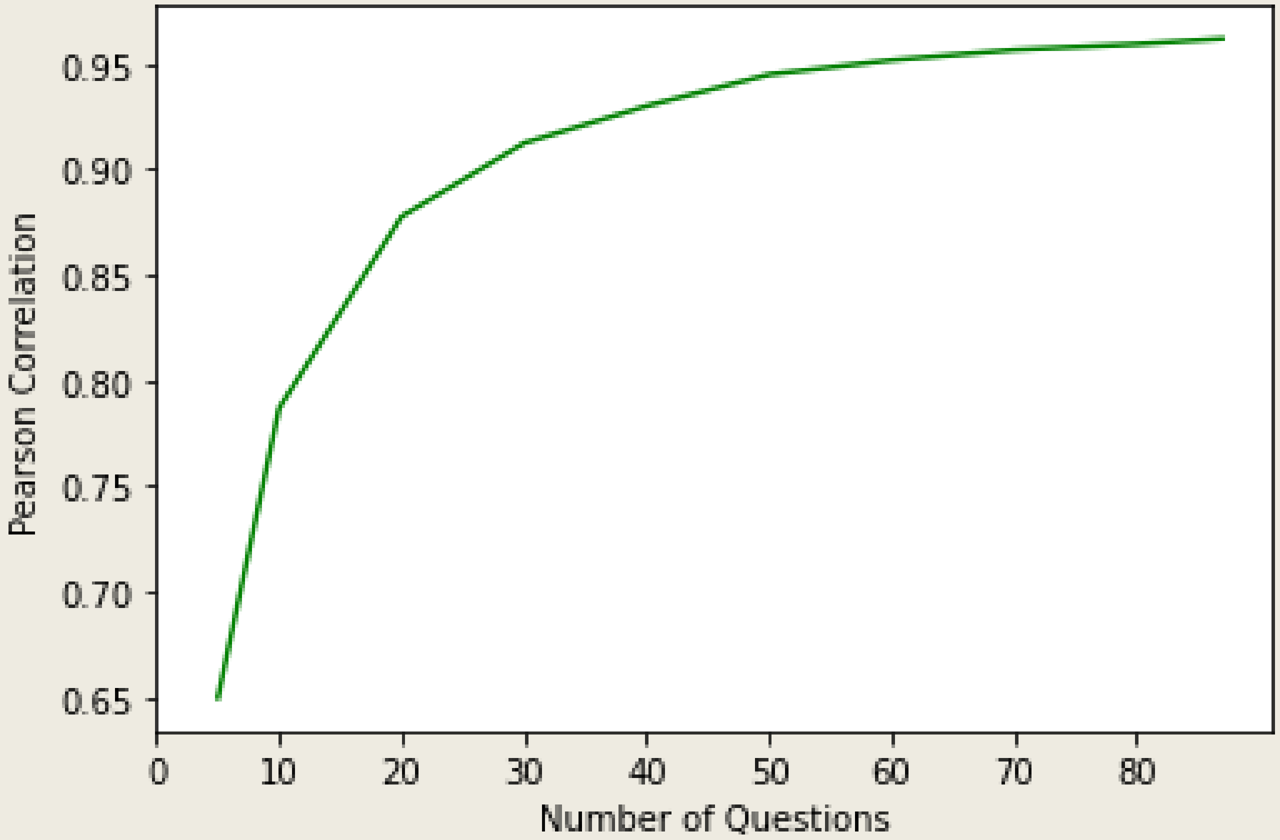
Pearson correlation values are between the model-based evaluation and the average score of two expert evaluators
*** P value < .01

Impact of iterative re-weighting (Wikipedia data)

Dataset	Wisdom of the crowd	Wisdom of the wise	%Improvement: AWER vs. Baseline
Movies and History	0.898	0.915	1.9%**
Science / Technology and Sports	0.939	0.950	1.2%***

Bootstrap P values *** P value < .01
Pearson correlation values are between the model-based evaluation and the average score of two expert evaluators

Impact of number of questions (CS data)



Additional tests

- Embeddings (RoBERTa, MPNet) vs. Bag of Words – the former performed slightly better
- Cosine vs. Euclidean distance – no significant difference

Bottom line

AWER utilizes the **wisdom of the crowd**, adjusted for textual entries, and benefits from learning **workers' capabilities**.

Still missing

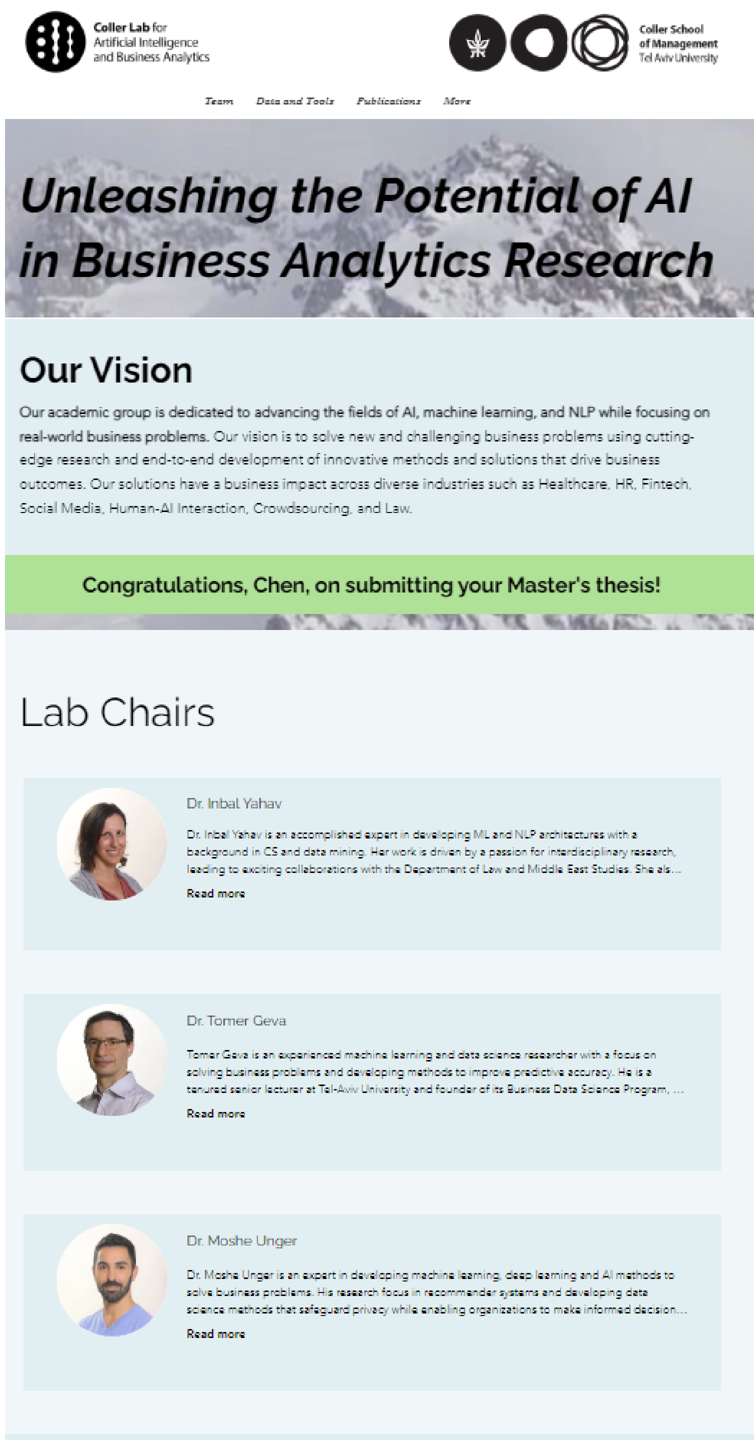
- Extracting the best response
- Evaluating language models in the question-answering task

Thank you!

Check us out!

https://www.collerlab.com/

inbalyahav@tauex.tau.ac.il



Additional slides

Future Work

- Using automated question answering methodologies algorithms to generate “ground-truth”. (Joint work with Shahar Meir and Inbal Yahav)

References:

- Bonthu, S., Rama Sree, S., & Krishna Prasad, M. H. M. (2021, August). Automated Short Answer Grading Using Deep Learning: A Survey. *In International Cross-Domain Conference for Machine Learning and Knowledge Extraction* (pp. 61-78). Springer, Cham.
- Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, *25*(1), 60–117.
- Geva, T., & Saar-Tsechansky. M. (2021). Who is a better decision maker? Data-driven expert ranking under unobserved quality. *Production and Operations Management*, *30*(1), 127–144.
- Mohler, M., Bunescu, R., & Mihalcea, R. (2011). Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 752–762).
- Roy, S., Dandapat, S., Nagesh, A., & Narahari, Y. (2016). Wisdom of students: A consistent automatic short answer grading technique. *Proceedings of the 13th International Conference on Natural Language Processing* (pp. 178–187).
- Wang, J., Ipeirotis, P. G., & Provost, F. (2017). Cost-effective quality assurance in crowd labeling. *Information Systems Research*, *28*(1), 137–158.
- Yin, J., Luo, J., & Brown S. A. (2021). Learning from crowdsourced multi-labeling: A variational Bayesian approach. *Information Systems Research*, *32*(3), 752–773.

Numerical Simulation - details

- Three types of workers with different correctness levels Q: 85%, 75%, 65%. (33 workers in each group)
- Random binary vector for correct response (dim=1,024)
- Two types of questions:
  - Standard – simulated responses are based on correct responses. Probability for inverting a response element is 1-Q
  - Challenging questions: if workers accuracy<85% then probability of a correct response element is 20%. Thus generating similar incorrect responses.
- Simulation varies the ratio of challenging questions
- Total number of question = 20.
- Number of simulation repetitions = 50.