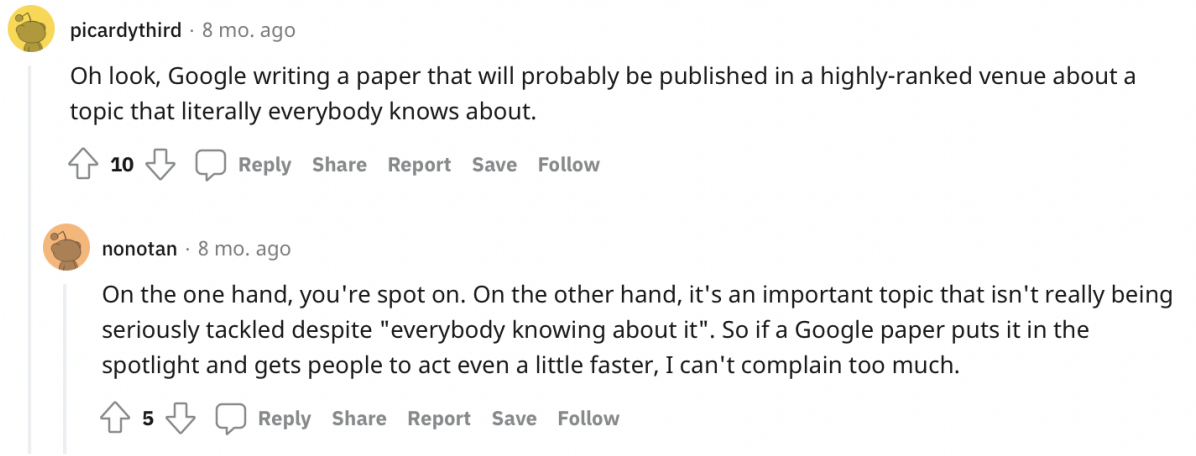
1. Работа написана летом 2021 года. Нигде не нашел ее упоминания в коференциях.
2. Авторы статьи сотрудники Google(Brain, Research). Три основных автора: Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko.

Mostafa Dehghani: не заметно, чтобы он специализировался на бенчмарках, есть одна статья со словом benchmark в названии. Является со-автором статьи про применение трансформеров на изображениях. Еще у него есть стартья “Efficient transformers: A survey”, что тоже может быть связанным с данной статьей, так как немалая часть статьи анализирует именно трансформеры и NLP бенчмарки.

Yi Tay: не похоже, что занимался исследованиями в данной области.

Alexey A. Gritsenko: аналогично.

В целом формат статьи похож на “survey” и ее ценность заключается скорее в том, что такой большой игрок как Google вообще на нее обратил внимание. Скрин из Reddit, хорошо отражающий данную мысль:



3. Нет таких статей, которые можно выделить как сильно повлиявшие на данную. Можно лишь отметить, что данная статья анализирует бенчмарки SuperGLUE, Visual Task Adaptation Benchmark, Long Range Arena, RL Unplugged

4. Из интересных статей, которые цитируют данную работу, можно выделить:
- https://arxiv.org/pdf/2112.01342.pdf - авторы меняют схему усреднения, после чего кардинально меняются значение метрики и положение моделей отнсительно друг друга
- https://proceedings.neurips.cc/paper/2021/file/f514cec81cb148559cf475e7426eed5e-Paper.pdf - авторы критикуют существующую схему оценки качества в RL и предлагают свою, основанную на статистической теории
- https://openreview.net/pdf?id=FBBWy2Sjwg - авторы поднимают ту же проблему и анализируют поведение моделей на граничных примерах, т.е. тех, которые можно отнести к обоим классам одновременно

5. Эти работы нельзя назвать прямо конкурентами, но они вышли примерно в то же время и рассматривают ту же проблему:

- https://arxiv.org/pdf/2104.14337.pdf - авторы создают новый сильный динамично меняющийся бенчмарк, отмечая проблему плохих бенчмарков
- https://aclanthology.org/2021.naacl-main.385.pdf - авторы предлагают 4 критерия хороших бенчмарков, критикуют adversarial подход из предыдущей статьи

6. Учитывая, что формат статьи скорее походит на “survey”, то в плане исследования сложно что-нибудь адекватное дополнительное предложить.

7. Из индустриальных примеров можно выделить статью “How not to Lie with a Benchmark: Rearranging NLP Leaderboards”. Другим вариантом может быть атака на существующие модели, понимая, что они были оптимизированы под определенные проблемные бенчмарки.

8. Из интересного, тоже статья про лотерею:
https://madaan.github.io/res/papers/sigbovik_real_lottery.pdf