Figure showing fail rate (y-axis) versus number of training examples (x-axis) across multiple test categories. Each subplot compares SQuAD 1.1 (blue) and SQuAD 1.1 + AdversarialQA (orange).

Subplot titles:
- A is COMP than B. Who is more / less COMP?
- Intensifiers (very, super, extremely) and reducers (somewhat, kinda, etc)?
- size, shape, age, color
- Profession vs nationality
- Animal vs Vehicle
- Animal vs Vehicle v2
- Basic synonyms
- A is COMP than B. Who is antonym(COMP)? B
- A is more X than B. Who is more X? B. Who is less X? B. Who is more X? A. Who is less antonym(X)? A.
- Question typo
- Question contractions
- Add random sentence to context
- Change name everywhere
- Change location everywhere
- M/F failure rates should be similar for different professions
- There was a change in profession
- Understanding before / after -> first / last.
- Negation in context, may or may not be in question
- Negation in question only.
- Basic coref, he / she
- Basic coref, his / her
- Former / Latter
- Agent / object distinction
- Agent / object distinction with 3 agents