Conclusion

Multilingual models uncovers a consistency bottleneck whereby the consistency does not grow monotonically on each layer.

Key Factors

- model architectures
- training strategies
- deep semantic alignments

Promising directions

- test-time calibration
- training with cross-lingual alignment objectives
- code-switching training

Necessary but not sufficient conditions

- cross-lingual representations
- shared scripts
- parallel samples