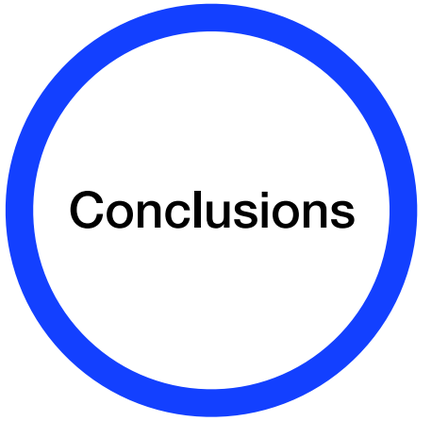

- **Proper evaluation is difficult**
- **Not clear what to measure, how**
- **Things we care about are hard to measure**
- **Many choices that can influence results**
- **Using just one error metric can give us a limited view of how these systems work. We should always try to evaluate with different methods our models**
- **Not a replacement for A/B testing, just to be more confident before it**
- **Discard models that perform poorly already during offline testing**

EBU

European Broadcasting Union