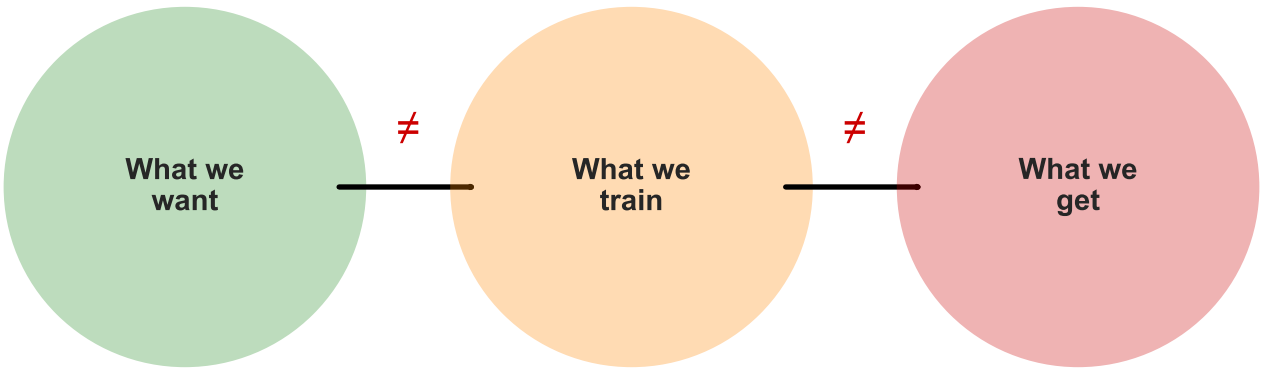
The Alignment Problem

Intent ≠ Specification ≠ Behavior



- Helpful but not harmful
- Truthful but not overly verbose
- Creative but not hallucinating