Edge cases are very easy for AI but hard for humans

Real-world tasks differ significantly from benchmarks

Al benchmarks guarantee legal compliance