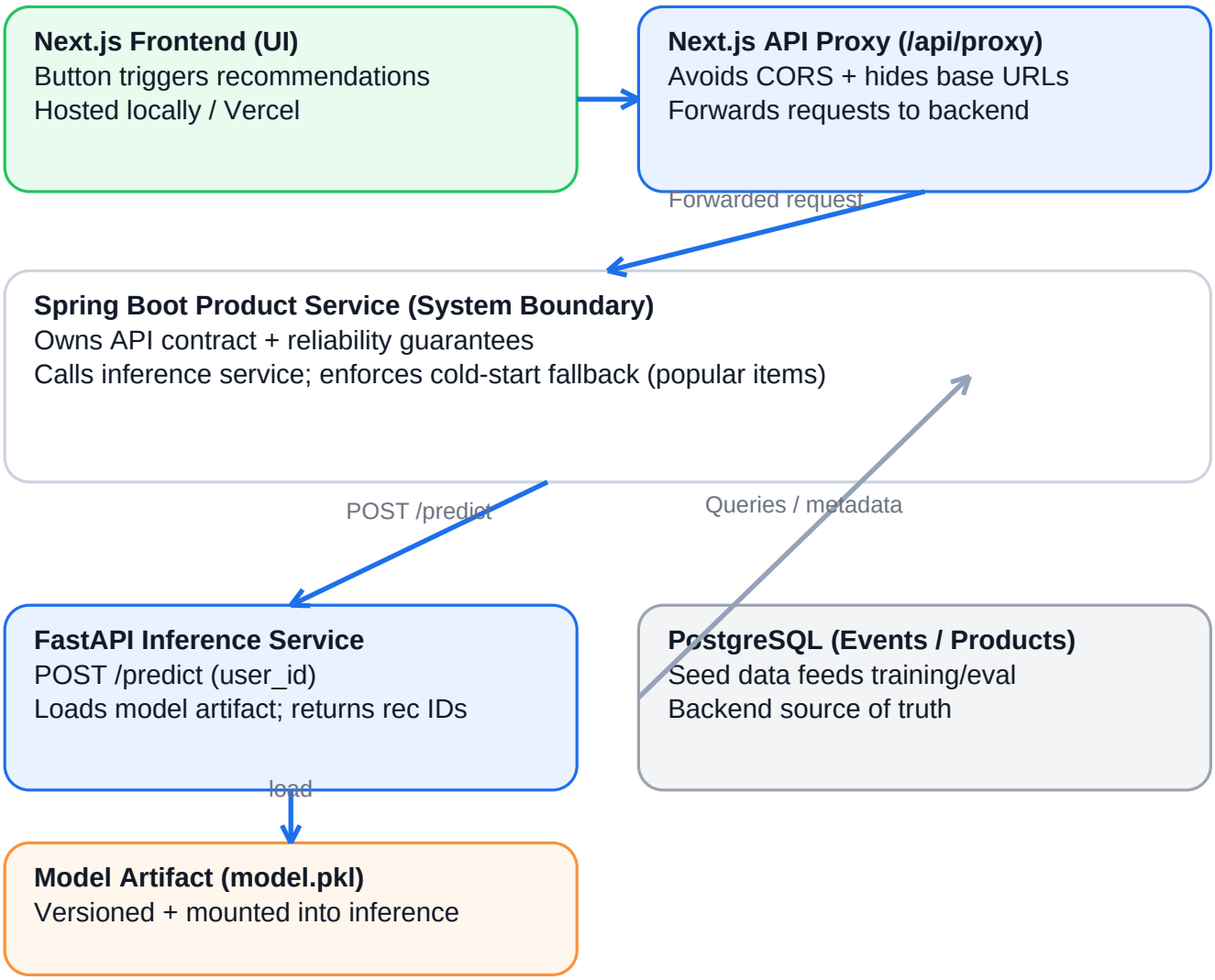
Production-Style MLOps Recommendation System

Architecture diagram + interview talking points anchor (NYC finance / trading ready)



Key production behaviors to call out in interviews

- Backend owns the reliability boundary: integrates ML when available, never lets ML violate system guarantees.
- Cold-start handling: if ML returns empty, backend returns deterministic popularity fallback (source/reason fields).
- Graceful degradation: if inference times out or errors, backend still returns valid recommendations.
- Observability: response metadata enables monitoring of fallback rate (signal for drift / sparse data / outages).