Toloka's moderation solutions are out-of-the-box APIs which can be customized and empowered by humans to meet excellent quality requirements

Out-of-the-box Moderation APIs:

Text Moderation

Visual Moderation

Audio Moderation

Classes:

— Sexual
— Gibberish
— Insult/hate/offensive
— Threat/violence
— Obscenity/profanity
— Spam
— Promotions
— Personal data (PII)

+ Quick Customization:

Custom classes

Languages

+ Quality Booster:

Human verification

Toloka

Solution building: typical flow

Pre-selection of APIs

Selection of pre-trained Adaptive ML Models

- ✓ All types of content
- ✓ Key classes covered
- ✓ 9 languages supported out-of-the-box

Customization
~1-2 weeks ETA

Classes & languages are covered by pre-trained APIs
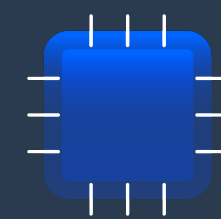
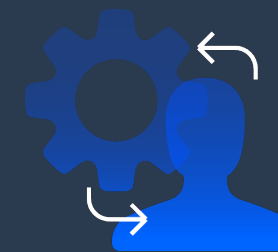For custom needs:

Custom **classes** training

Custom **languages** training

Quality Boosting
tailored to client's needs
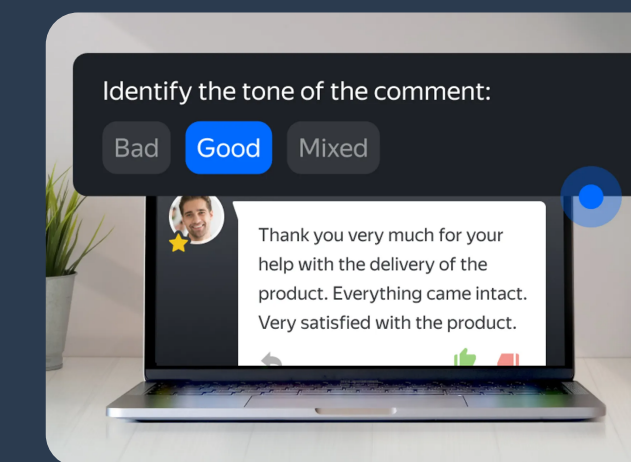
Fine tune models on client's data
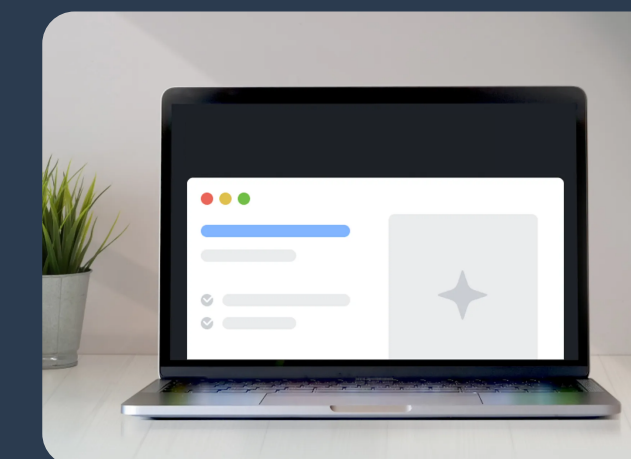
Human verification
Pre-built in Toloka

Client's own moderators verification

Ready solution for the client

Examples:

Moderation of public messages to find suspicious and illegal behavior in 3 languages. Includes language detection and classification in 7 classes

Moderation of images and text for product listings to comply with local laws

Toloka