onyx

Welcome to Onyx › **Configuring Onyx**

Welcome to Onyx

Configuring Onyx

How to customize your deployment environment.

Common Environment Variables

All of the global configuration options that are not built into the UI are set via environment variables. This page contains an exhaustive list of all the options.

There are defaults set in the code so changing/settings these values is **not required** to use Onyx. A few notable settings that are more frequently changed however are the following:

AUTH_TYPE (default value is `disabled`)

MULTILINGUAL_QUERY_EXPANSION (you can provide a comma separated list of languages for query rephrasing such as `English,French`)

LOG_LEVEL (default is `info`)

WEB_DOMAIN (your full url in production, including the protocol- e.g. `https://www.onyx.app`)

Docker Compose

There are several ways to configure environment variables for the containers. For Docker Compose, there are many ways to pass environment variables to the container, any of the standard approaches will work. However **the preferred approach** for Onyx is to use the `.env` file. To do this, create a file called `.env` at `onyx/deployment/docker_compose/.env`. From there, populate it with the values you want to override:

```
# Configures basic email/password based login
AUTH_TYPE="basic"

# Rephrasing the query into different languages to improve search recall
MULTILINGUAL_QUERY_EXPANSION="English,Spanish,German"

# Set a cheaper/faster LLM for the flows that are easier (such as translating the
FAST_GEN_AI_MODEL_VERSION="gpt-3.5-turbo"

# Setting more verbose logging
LOG_LEVEL="debug"
LOG_ALL_MODEL_INTERACTIONS="true"
```

Kubernetes

For Kubernetes, the deployment yaml files includes an Environment ConfigMap. Simply update the values in the file **here**.

All Environment Variables

This is an extensive list of the currently supported environment variables within Onyx. There are several classes of environment variables. All of the global configuration options that are not built into the UI are set via environment variables.

You can set these in your `.env` file.

Auth Settings

These variables control authentication and user management in Onyx.

AUTH_TYPE

Controls the authentication method used by Onyx.

`disabled` : No authentication is required.

`google_oauth` : Users can log in using their Google accounts.

`basic` : Standard username/password authentication.

`oidc` : OpenID Connect, available in the enterprise edition.

`saml` : Security Assertion Markup Language, available in the enterprise edition.

SESSION_EXPIRE_TIME_SECONDS

Defines the duration of a user's session in seconds. Default is 24 hours.

ENCRYPTION_KEY_SECRET

A strong, unique string used for encryption purposes. Keep this value secret.

VALID_EMAIL_DOMAINS

Comma-separated list of allowed email domains for authentication. Leave empty to allow all domains.

GOOGLE_OAUTH_CLIENT_ID

Client ID for Google OAuth authentication, obtained from **Google Cloud Console**.

GOOGLE_OAUTH_CLIENT_SECRET

Client Secret for Google OAuth authentication, obtained from **Google Cloud Console**. Keep this value secret.

REQUIRE_EMAIL_VERIFICATION

When set to `true` , users must verify their email before accessing Onyx.

SMTP_SERVER

Hostname of the SMTP server for sending verification emails. Default is `smtp.gmail.com` .

SMTP_PORT

onyx

Port used for SMTP communication. Common values are `587` (TLS) or `465` (SSL).

›

SMTP_USER

Username for SMTP authentication, often an email address used to send verification emails.

SMTP_PASS

Password for SMTP authentication. Keep this value secret.

EMAIL_FROM

Email address used as the sender for verification emails.

NEXT_PUBLIC_FORGOT_PASSWORD_ENABLED

Set to `true` to enable the forgot password feature. Only enable this if you have configured the above SMTP settings (For email functionality).

Gen AI Settings

These variables configure the generative AI capabilities of Onyx. These are covered in more depth at the **generative AI configs**.

GEN_AI_MODEL_PROVIDER

Specifies the provider of the generative AI model (e.g., `openai`, `anthropic`, `huggingface`).

GEN_AI_MODEL_VERSION

Defines the version of the generative AI model to use (e.g., `gpt-4` for OpenAI).

FAST_GEN_AI_MODEL_VERSION

Specifies a faster (usually smaller) model version for certain tasks.

GEN_AI_API_KEY

onyx

API key for accessing the generative AI service. Keep this value secret.

›

GEN_AI_LLM_PROVIDER_TYPE

Specifies the type of LLM provider being used (e.g., `openai`, `anthropic`, `azure`).

GEN_AI_MAX_TOKENS

Maximum number of tokens to generate in AI responses.

QA_TIMEOUT

Timeout for question-answering operations in seconds.

MAX_CHUNKS_FED_TO_CHAT

Maximum number of document chunks fed into a single chat session.

DISABLE_LLM_FILTER_EXTRACTION

Set to `true` to disable LLM-based filter extraction from queries.

DISABLE_LLM_CHUNK_FILTER

Set to `true` to disable LLM-based filtering of document chunks.

DISABLE_LLM_CHOOSE_SEARCH

Set to `true` to disable LLM-based selection of search method.

DISABLE_LLM_QUERY_REPHRASE

Set to `true` to disable LLM-based query rephrasing.

DISABLE_GENERATIVE_AI

Set to `true` to disable all generative AI functionality.

DISABLE_LITELLM_STREAMING

onyx

Set to `true` to disable streaming responses when using LiteLLM.

›

LITELLM_EXTRA_HEADERS

JSON-formatted string of key-value pairs for additional headers in LiteLLM API requests.

TOKEN_BUDGET_GLOBALLY_ENABLED

Set to `true` to enable the global token budget system.

AWS Bedrock Settings

These variables are used for AWS Bedrock integration.

AWS_ACCESS_KEY_ID

AWS access key ID for Bedrock access, obtained from AWS IAM.

AWS_SECRET_ACCESS_KEY

AWS secret access key for Bedrock access, obtained from AWS IAM. Keep this value secret.

AWS_REGION_NAME

AWS region where Bedrock is deployed (e.g., `us-west-2`).

Query Options

These variables control various aspects of query processing and search behavior.

DOC_TIME_DECAY

Controls the recency bias in search results. Higher values increase preference for newer documents.

HYBRID_ALPHA

onyx

Balances keyword vs. vector search in hybrid search. Range 0-1 (0 for pure keyword, 1 for pure vector search).

EDIT_KEYWORD_QUERY

Set to `true` to enable query editing for keyword searches.

MULTILINGUAL_QUERY_EXPANSION

Set to `true` to enable multilingual query expansion.

QA_PROMPT_OVERRIDE

Custom prompt text to override the default prompt used for question-answering.

Other Services

Configuration for external services used by Onyx.

POSTGRES_HOST

Hostname or IP address of the Postgres server. Default is `relational_db`.

VESPA_HOST

Hostname or IP address of the Vespa server. Default is `index`.

WEB_DOMAIN

Fully qualified domain name used for the Onyx web interface. (e.g. **https://www.onyx.com**)

NLP Model Configurations

Advanced settings for NLP models. Modify with caution.

INDEX_BATCH_SIZE

onyx

Size of batch used when indexing documents. Overrides the default batch size for indexing operations.

›

EMBEDDING_BATCH_SIZE

Size of batch used when embedding documents during indexing and search operations. Overrides the default batch size for embedding processes.

DOCUMENT_ENCODER_MODEL

Name or path of the encoder model used for document encoding.

DOC_EMBEDDING_DIM

Dimension of document embeddings, typically matching the chosen encoder model's output dimension.

NORMALIZE_EMBEDDINGS

Set to `true` to enable normalization of embeddings.

ASYM_QUERY_PREFIX

Text prepended to queries in asymmetric semantic search.

ENABLE_RERANKING_REAL_TIME_FLOW

Set to `true` to enable reranking in real-time search flow.

ENABLE_RERANKING_ASYNC_FLOW

Set to `true` to enable reranking in asynchronous search flow.

MODEL_SERVER_HOST

Hostname or IP address of the model server. Default is `inference_model_server` .

MODEL_SERVER_PORT

onyx

Port on which the model server is listening.

›

Miscellaneous

< Various other configuration options.

DISABLE_TELEMETRY

Set to `true` to opt out of telemetry. Telemetry helps improve Onyx; no sensitive data is collected.

LOG_LEVEL

Sets the logging verbosity. Possible values: `debug`, `info`, `warning`, `error`, `critical`.

LOG_ALL_MODEL_INTERACTIONS

Set to `true` to enable logging of all prompts sent to the LLM.

LOG_VESPA_TIMING_INFORMATION

Set to `true` to enable additional logging of Vespa query performance.

LOG_ENDPOINT_LATENCY

Set to `true` to enable logging of endpoint latency information.

DEFAULT_PERMISSION_DOC_SYNC_FREQUENCY

The default interval between per document permission syncs for any connector without its own specific setting (in seconds).

SLACK_PERMISSION_DOC_SYNC_FREQUENCY

The default interval between per document permission syncs for Slack (in seconds).

GOOGLE_DRIVE_PERMISSION_GROUP_SYNC_FREQUENCY

onyx

The interval between external group permission syncs for Google Drive (in seconds).

Custom Model Server

Multilingual Setup