LumenVox Prometheus Metrics - updated June 2025 (version 6.1.0)

New metric

Container	Measurement	Description	Type
Deployment	deployment_active_count	Number of active deployments	gauge
Deployment	deployment_total_responses_returned	Number of responses returned by the container (including success, timeouts and errors responses)	gauge vector
Deployment	deployment_active_requests	Total number of active requests	gauge
Deployment	deployment_average_request_process_time	Distribution average request processing time	histogram
Deployment	deployment_max_requests	The maximum number of simultaneous deployment requests	gauge
Deployment	deployment_total_requests	Total number of deployment requests	counter
vad	vad_audio_streams_current	Active current VAD streams	gauge
vad	vad_audio_streams_max	The maximum concurrent number of VAD streams	gauge
vad	vad_audio_streams_total	The total number of completed VAD streams	counter
vad	vad_audio_timeout_total	Total number of VAD stream timeouts	counter
vad	vad_active_requests	Total number of active requests	gauge
vad	vad_total_cpa_requests	Total number of CPA requests received	counter
vad	vad_active_cpa_requests	Total number of active CPA requests	gauge
vad	vad_average_cpa_request_process_time_dist	Distribution average CPA request processing time	histogram
vad	vad_total_cpa_responses_returned	Number of CPA responses returned by the container (including success, timeouts and errors responses)	gauge vector
vad	vad_total_amd_requests	Total number of AMD requests received	counter
vad	vad_active_amd_requests	Total number of AMD active requests	gauge
vad	vad_average_amd_request_process_time_dist	Distribution average AMD request processing time	histogram
vad	vad_total_amd_responses_returned	Number of responses returned by the container (including success, timeouts and errors responses)	gauge vector
vad	vad_total_asr_requests	Total number of ASR requests received	counter
vad	vad_active_asr_requests	Total number of active ASR requests	gauge
vad	vad_average_asr_request_process_time_dist	Distribution average ASR request processing time	histogram
vad	vad_total_responses_returned	Number of ASR responses returned by the container (including success, timeouts and errors responses)	gauge vector
vad	vad_total_transcription_requests	Total number of Transcription requests received	counter
vad	vad_active_transcription_requests	Total number of active Transcription requests	gauge

LumenVox Prometheus Metrics - updated June 2025 (version 6.1.0)

New metric

Container	Measurement	Description	Type
vad	vad_average_transcription_request_process_time_dist	Distribution average Transcription request processing time	histogram
vad	vad_total_transcription_responses_returned	Number of Transcription responses returned by the container (including success, timeouts and errors responses)	gauge vector
session	session_total_requests	Total number of requests received	counter
session	session_active_requests	Total number of active requests	gauge
session	session_average_request_process_time_dist	Distribution average request processing time	histogram
session	session_total_responses_returned	Number of responses returned by the container (including success, auto close, errors and timeout responses	gauge vector
asr	asr_total_asr_requests	Total number of requests received	counter
asr	asr_active_asr_requests	Total number of active requests	gauge
asr	asr_active_europa_requests	Total number of active backend engine requests	gauge
asr	asr_average_asr_request_process_time_dist	Distribution average request processing time	histogram
asr	asr_max_asr_requests	The maximum number of simultaneous asr requests	gauge
asr	asr_total_asr_responses_returned	Number of responses returned by the container (including success, timeouts and errors responses)	gauge vector
asr	asr_total_grammar_requests	Total number of grammar load requests received	counter
asr	asr_active_grammar_load_requests	Total number of active grammar load requests	gauge
asr	asr_average_asr_stream_request_process_time_dist	Distribution average request processing time	histogram
asr	asr_max_concurrent_grammar_load_requests	The maximum number of simultaneously active grammars load requests	gauge
asr	asr_average_grammar_load_request_process_time_dist	Distribution average grammar load request processing time	histogram
asr	asr_total_grammar_load_responses_returned	Number of responses returned by the container (including success, timeouts and errors responses)	gauge vector
asr	asr_total_transcription_requests	Total number of grammar load requests received	counter
asr	asr_active_transcription_requests	Total number of active grammar load requests	gauge

LumenVox Prometheus Metrics - updated June 2025 (version 6.1.0)

New metric

Container	Measurement	Description	Type
asr	asr_total_transcription_responses_returned	Number of responses returned by the container (including success, timeouts and errors responses)	gauge vector
asr	asr_average_asr_batch_request_process_time_dist	Distribution average grammar load request processing time	histogram
asr	asr_average_transcription_batch_request_process_time_dist	Distribution average grammar load request processing time	histogram
asr	asr_average_transcription_stream_request_process_time_dist	Distribution average grammar load request processing time	histogram
asr	asr_max_transcription_requests	The maximum number of simultaneous transcription requests	gauge
asr	asr_max_active_grammars	The maximum number of simultaneously active grammars (within the grammar cache)	gauge
asr	asr_max_active_parses	The maximum number of simultaneous active SISR parses	gauge
asr	asr_active_decodes	The active number of active decodes being processed	gauge
asr	asr_active_grammars	The active number of active grammars being processed	gauge
asr	asr_active_parses	The active number of active SISR parses being processed	gauge
asr	asr_average_sisr_parse_text_request_process_time_dist	Distribution average request processing time	histogram
asr	asr_sisr_parse_text_requests_total	Total number of requests received	counter
asr	asr_total_sisr_parse_text_responses_returned	Number of responses returned by the container (including success, timeouts and errors responses)	gauge vector
asr	asr_total_ms_audio_pushed	Total milliseconds of audio pushed into ASR	counter
asr	asr_cache_entries	This is the number of entries currently present in the ASR grammar cache.	counter
asr	asr_cache_size_bytes	This is the current size, in bytes, of the grammar cache.	counter
asr	asr_active_ms_audio_processing	Total milliseconds of audio currently being processed by the ASR	gauge

LumenVox Prometheus Metrics - updated June 2025 (version 6.1.0)

New metric

Container	Measurement	Description	Type
asr	asr_fine_tuned_results	FT - if transcription request was processed by the fine-tuned model DNN - if transcription request was processed by the DNN model Error - if there was an error processing interaction Timeout - if there was a decode timeout	gauge vector
tts	tts_total_requests	Total number of requests received	counter
tts	tts_active_requests	Total number of active requests	gauge
tts	tts_average_request_process_time_dist	Distribution average request processing time	histogram
tts	tts_total_responses_returned	Number of responses returned by the container	gauge vector
tts	tts_average_pending_queue_time	Average time of requests queued for processing	histogram
tts	tts_max_queue_size_synthesis_requests_tts1	The maximum number of simultaneous TTS1 synthesis requests at any one time since statistics were last reset (or startup)	gauge
tts	tts_active_queue_size_synthesis_requests_tts1	The current number of simultaneous TTS1 synthesis requests at being processed	gauge
tts	tts_max_pending_requests_tts1	The maximum number of pending TTS1 synthesis requests at any one time	counter
tts	tts_preprocess_load_cache_results (tts3/neural)	Used internally for testing	gauge
tts	tts_postprocess_load_cache_results (tts3/neural)	Used internally for testing	gauge
tts	tts_max_queue_size_synthesis_requests (tts3/neural)	Max TTS requests per container	gauge
tts	tts_first_result_time_max	Maximum time between client making synthesis request and receiving the first audio packed	histogram
tts	tts_first_result_time_min	Minimum time between client making synthesis request and receiving the first audio packed	histogram

LumenVox Prometheus Metrics - updated June 2025 (version 6.1.0)

New metric

Container	Measurement	Description	Type
tts	tts_first_result_time_dist	Average time between client making synthesis request and receiving the first audio packed	histogram
resource	resource_active_asr_installs	Actively installing ASR packages	gauge
resource	resource_asr_download_attempts_counter_total	The total number of ASR download attempts the resource manager has started	counter
resource	resource_asr_download_failure_counter_total	The total number of failed ASR downloads the resource manager detected	counter
resource	resource_asr_download_success_counter_total	The total number of successful ASR downloads the resource manager processed	counter
resource	resource_asr_language_packages_configured	The number of ASR packages configured for the system	gauge
resource	resource_tts_active_installs	Actively installing TTS packages	gauge
resource	resource_tts_download_attempts_counter_total	The total number of TTS download attempts the resource manager has started	counter
resource	resource_tts_download_failure_counter_total	The total number of failed TTS downloads the resource manager detected	counter
resource	resource_tts_download_success_counter_total	The total number of successful TTS downloads the resource manager processed	counter
resource	resource_tts_voice_packages_configured	The number of TTS packages configured for the system	gauge
resource	resource_active_vb_active_installs	Actively installing VB-Active packages	gauge
resource	resource_vb_active_download_attempts_counter_total	The total number of VB-Active download attempts the resource manager has started	counter
resource	resource_vb_active_download_failure_counter_total	The total number of failed VB-Active downloads the resource manager detected	counter
resource	resource_vb_active_download_success_counter_total	The total number of successful VB-Active downloads the resource manager processed	counter
resource	resource_vb_active_language_packages_configured	The number of VB-Active packages configured for the system	gauge
resource	resource_dnn_active_installs	Actively installing TTS packages	gauge
resource	resource_dnn_download_attempts_counter_total	The total number of TTS download attempts the resource manager has started	counter

LumenVox Prometheus Metrics - updated June 2025 (version 6.1.0)

New metric

Container	Measurement	Description	Type
resource	resource_dnn_download_failure_counter_total	The total number of failed TTS downloads the resource manager detected	counter
resource	resource_dnn_download_success_counter_total	The total number of successful TTS downloads the resource manager processed	counter
resource	resource_dnn_voice_packages_configured	The number of TTS packages configured for the system	gauge
resource	resource_itn_active_installs	Total number of itn resource installs	counter
resource	resource_itn_download_attempts_counter_total	Total number of itn download attempts	counter
license	license_invalid_check_ops_total	The total number of unsuccessful license check events	counter
license	license_sync_fail_ops_total	The total number of unsuccessful license sync events	counter
license	license_sync_ops_total	The total number of attempted license sync events	counter
license	license_sync_success_ops_total	The total number of successful license sync events	counter
license	license_valid_check_ops_total	The total number of successful license check events	counter
license	license_valid_licences	Number of valid license deployments	gauge
license	license_invalid_licences	Number of invalid license deployments	gauge
configuration	configuration_total_requests	Total number of requests received	counter
configuration	configuration_active_requests	Total number of active requests	gauge
configuration	configuration_max_requests	The maximum number of configuration requests	gauge
configuration	configuration_average_request_process_time_dist	Distribution average request processing time	histogram
configuration	configuration_total_responses_returned	Number of responses returned by the container (including success, timeouts and errors responses)	gauge vector
binary-storage	binary_storage_total_requests	Total number of requests received	counter
binary-storage	binary_storage_active_requests	Total number of active requests	gauge
binary-storage	binary_storage_max_requests	Maximum number of binary storage requests	gauge
binary-storage	binary_storage_average_request_process_time_dist	Distribution average request processing time	histogram
admin-portal	admin_portal_total_requests	Total number of admin portal requests	counter
admin-portal	admin_portal_average_request_process_time_dist	Distribution average request processing time	histogram

LumenVox Prometheus Metrics - updated June 2025 (version 6.1.0)

New metric

Container	Measurement	Description	Type
archive	archive_total_requests	Total number of requests received	counter
archive	archive_active_requests	Total number of active requests	gauge
archive	archive_average_request_process_time_dist	Distribution average request processing time	histogram
archive	archive_total_responses_returned	Number of responses returned by the container (including success, timeouts and errors responses)	gauge vector
archive	archive_active_execution	Total number of active archive requests currently being executed e.g. saving binary storage data	gauge
archive	archive_requests_max	Maximum number of archive requests received	gauge
archive	archive_total_execute	Total number of archive requests executed e.g. saving binary storage data	counter
deployment-portal	deployment_portal_total_requests	Total number of deployment portal requests	counter
deployment-portal	deployment_portal_active_requests	Total number of active deployment portal requests	gauge
deployment-portal	deployment_portal_requests_max	Maximum number of deployment portal requests	gauge
deployment-portal	deployment_portal_average_request_process_time_dist	Distribution average request processing time	histogram
deployment-portal	deployment_portal_total_responses_returned	Number of responses returned by the container (including success, timeouts and errors responses)	gauge vector
reporting	reporting_total_requests	Total number of requests received	counter
reporting	reporting_active_requests	Total number of active requests	gauge
reporting	reporting_average_request_process_time_dist	Distribution average request processing time	histogram
reporting	reporting_requests_max	Maximum number of reporting requests	gauge
lumenvox-api	lumenvox_api_total_requests	Total number of initial requests received	counter
lumenvox-api	lumenvox_api_total_requests_within sessions	Total number of api requests within sessions	counter
lumenvox-api	lumenvox_api_active_requests	Total number of active requests	gauge
lumenvox-api	lumenvox_api_total_responses_returned	Number of responses returned by the container (e.g. grpc error codes)	gauge vector
lumenvox-api	lumenvox_api_rmq_messages_received	Number of LumenVox API rabbitmq messages received	counter

LumenVox Prometheus Metrics - updated June 2025 (version 6.1.0)

New metric

Container	Measurement	Description	Type
lumenvox-api	lumenvox_api_rmq_messages_sent	Number of LumenVox API rabbitmq messages sent	counter
itn	itn_request_times	itn request times	histogram
itn	itn_requests_current	Active itn requests	gauge
itn	itn_requests_max	The maximum number of simultaneous TTS synthesis requests at any one time since statistics were last reset (or startup)	gauge
itn	itn_requests	Total itn requests	counter
nlu	nlu_average_request_process_time_dist	Distribution average request processing time	histogram
nlu	nlu_active_requests	Total number of active requests	gauge
nlu	nlu_total_requests	Total number of initial requests received	counter
nlu	nlu_total_responses_returned	Number of responses returned by the container	gauge vector
mrcp	mrcp_total_requests	Total number of calls (sessions) received	counter
mrcp	mrcp_active_requests	Total number of active requests	gauge
mrcp	mrcp_average_request_process_time_dist	Distribution average request processing time	histogram
mrcp	mrcp_total_responses_returned	Number of responses returned by the container (including success, timeouts and errors responses)	gauge vector
mrcp	mrcp_max_calls	The maximum simultaneous number of calls processed at one time	gauge
mrcp	mrcp_sip_calls	Total number of SIP calls processed	counter
mrcp	mrcp_sip_tcp_connections	Total number of SIP TCP calls processed	counter
mrcp	mrcp_rtsp_calls	Total number of RTSP calls processed	counter
mrcp	mrcp_garbage_collection_calls	Total ended calls that are in the process of having being garbage collected	gauge
diarization	diarization_average_request_process_time_dist	Distribution average request processing time	histogram
diarization	diarization_active_requests	Total number of active requests	gauge
diarization	diarization_total_requests	Total number of initial requests received	counter
lid (language id)	lid_average_request_process_time_dist	Distribution average request processing time	histogram
lid (language id)	lid_active_requests	Total number of active requests	gauge
lid (language id)	lid_total_requests	Total number of initial requests received	counter