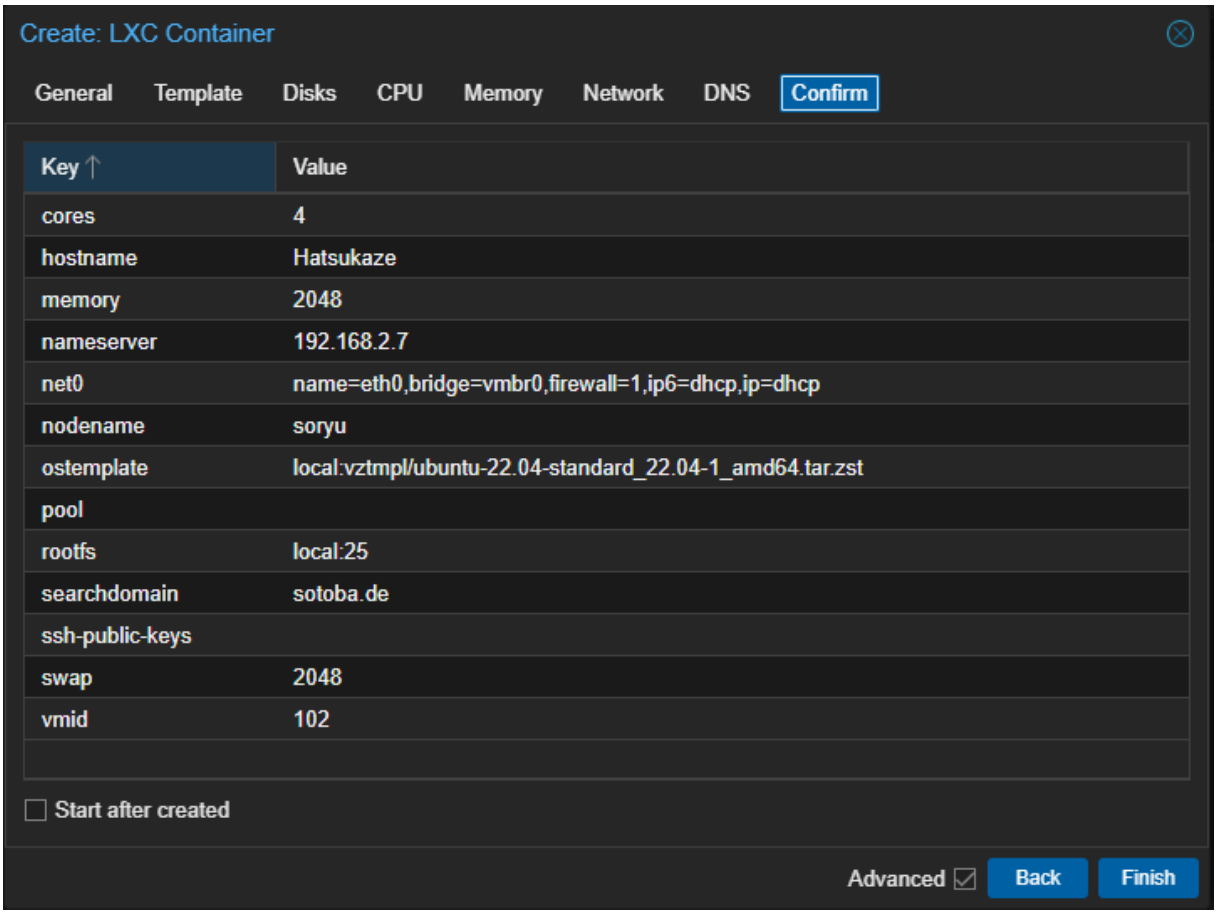
Open-LLM-V-Tuber

In diesem kleinen Projekt geht es um den Versuch den virtuellen, KI gesteuerten Chatbot. Open-LLM-V-Tuber zum Laufen zu bekommen. Der Chatbot selbst stammt vom chinesischen Entwickler Yi-Ting Chiu

Der Container

Zuerst möchte ich versuchen, den Chatbot in einem Container zu hosten. Dazu erstelle ich mir folgenden Container



Nach dem Erstellen des Containers, bringe ich ihn erstmal auf den neusten Stand

apt-get update && apt-get upgrade && reboot

```
Fetched 24.0 MB in 12s (2036 kB/s)
Reading package lists... Done
Reading package lists... Done
Building dependency tree... Done
Calculating upgrade... Done
The following packages have been kept back:
  libudev1 python3-update-manager ubuntu-advantage-tools udev update-manager-core
The following packages will be upgraded:
  apparmor apt apt-utils base-files bash bind9-dnsutils bind9-host bind9-libs bsdextrautils bsdutils busybox-static ca-certificates coreutils cpio dbus distro-info
  distro-info-data dmidecode dmsetup dpkg e2fsprogs eject file gcc-12-base gpgv gzip iptables irqbalance isc-dhcp-client isc-dhcp-common kbd less libapparmor1
  libapt-pkg6.0 libblkid1 libbpf0 libc-bin libc6 libcap2 libcap2-bin libcom-err2 libcryptsetup12 libdbus-1-3 libdevmapper1.02.1 libdrm-common libdrm2 libelf1
  libexpat1 libext2fs2 libfribidi0 libgcc-s1 libglib2.0-0 libglib2.0-data libgnutls30 libgssapi-krb5-2 libip4tc2 libip6tc2 libjson-c5 libk5crypto3 libkrb5-3
  libkrb5support0 libmagic-mgc libmagic1 libmount1 libncurses6 libncursesw6 libnetplan0 libnftables1 libnghttp2-14 libnss-systemd libntfs-3g89 libpam-cap
  libpam-modules libpam-modules-bin libpam-runtime libpam-systemd libpam0g libpcap0.8 libpcre2-8-0 libpcre3 libprocps8 libpython3-stdlib libpython3.10-minimal
  libpython3.10-stdlib libsasl2-2 libsasl2-modules-db libseccomp2 libsmartcols1 libsqlite3-0 libss2 libssl3 libstdc++6 libsystemd0 libtasn1-6 libtinfo6
  libtirpc-common libtirpc3 libunwind8 libusb-1.0-0 libuuid1 libuv1 libx11-6 libx11-data libxml2 libxtables12 locales login logrotate logsave mount nano ncurses-base
  ncurses-bin netplan.io networkd-dispatcher nftables ntfs-3g openssh-client openssh-server openssh-sftp-server openssl passwd pci.ids perl-base postfix procps
  python-apt-common python3 python3-apt python3-distro-info python3-distupgrade python3-gdbm python3-gi python3-minimal python3-pkg-resources python3.10
  python3.10-minimal rsync rsyslog ssh sudo systemd systemd-sysv systemd-timesyncd tar tcpdump tzdata ubuntu-minimal ubuntu-release-upgrader-core ubuntu-standard ufw
  util-linux uuid-runtime vim-common vim-tiny wget xxd zlib1g
158 upgraded, 0 newly installed, 0 to remove and 5 not upgraded.
98 standard security updates
Need to get 57.8 MB of archives.
After this operation, 4320 kB disk space will be freed.
Do you want to continue? [Y/n]
```

Nach dem Neustart installiere ich dann die nötigen Voraussetzungen für den Chatbot. Zuerst Git

apt install git

```
root@Hatsukaze:~# apt install git
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
  git-man libbrotli1 libcurl3-gnutls liberror-perl libgdbm-compat4 libperl5.34 librtmp1 libssh-4 patch perl perl-modules-5.34
Suggested packages:
  git-daemon-run | git-daemon-sysvinit git-doc git-email git-gui gitk gitweb git-cvs git-mediawiki git-svn diffutils-doc perl-doc libterm-readline-gnu-perl
  | libterm-readline-perl-perl make libtap-harness-archive-perl
The following NEW packages will be installed:
  git git-man libbrotli1 libcurl3-gnutls liberror-perl libgdbm-compat4 libperl5.34 librtmp1 libssh-4 patch perl perl-modules-5.34
0 upgraded, 12 newly installed, 0 to remove and 5 not upgraded.
Need to get 13.1 MB of archives.
After this operation, 71.5 MB of additional disk space will be used.
Do you want to continue? [Y/n]
```

Danach ffmpeg

apt install ffmpeg

```
root@Hatsukaze:~# apt install ffmpeg
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
  alsa-topology-conf alsa-ucm-conf fontconfig fontconfig-config fonts-dejavu-core i965-va-driver intel-media-va-driver libaacs0 libaom3 libasound2 libasound2-data
  libass9 libasyncns0 libavc1394-0 libavcodec58 libavdevice58 libavfilter7 libavformat58 libavutil56 libbdplus0 libblas3 libbluray2 libbs2b0 libcaca0
  libcairo-gobject2 libcairo2 libcdio-cdda2 libcdio-paranoia2 libcdio19 libchromaprint1 libcodec2-1.0 libdatrie1 libdav1d5 libdc1394-25 libdecor-0-0
  libdecor-0-plugin-1-cairo libdeflate0 libdrm-amdgpu1 libdrm-intel1 libdrm-nouveau2 libdrm-radeon1 libflac8 libflite1 libfontconfig1 libfreetype6 libgbm1
  libgdk-pixbuf-2.0-0 libgdk-pixbuf2.0-bin libgdk-pixbuf2.0-common libgfortran5 libgl1 libgl1-amber-dri libgl1-mesa-dri libglapi-mesa libglvnd0 libglx-mesa0 libglx0
  libgme0 libgomp1 libgraphite2-3 libgsm1 libharfbuzz0b libiec61883-0 libigdgmm12 libjack-jackd2-0 libjbig0 libjpeg-turbo8 libjpeg8 liblapack3 liblilv-0-0 libllvm15
  libmfx1 libmp3lame0 libmpg123-0 libmysofa1 libnorm1 libogg0 libopenal-data libopenal1 libopenjp2-7 libopenmpt0 libopus0 libpango-1.0-0 libpangocairo-1.0-0
  libpangoft2-1.0-0 libpciaccess0 libpgm-5.3-0 libpixman-1-0 libpocketsphinx3 libpostproc55 libpulse0 libquadmath0 librabbitmq4 libraw1394-11 librsvg2-2
  librsvg2-common librubberband2 libsamplerate0 libsd12-2.0-0 libsensors-config libsensors5 libserd-0-0 libshine3 libsnappy1v5 libsndfile1 libsndio7.0 libsodium23
  libsord-0-0 libsoxr0 libspeex1 libsphinxbase3 libsratom-0-0 libsrt1.4-gnutls libssh-gcrypt-4 libswresample3 libswscale5 libthai-data libthai0 libtheora0 libtiff5
  libtwolame0 libudfread0 libva-drm2 libva-x11-2 libva2 libvdpau1 libvidstab1.1 libvorbis0a libvorbisenc2 libvorbisfile3 libvpx7 libwayland-client0 libwayland-cursor0
  libwayland-egl1 libwayland-server0 libwebp7 libwebpmux3 libx11-xcb1 libx264-163 libx265-199 libxcb-dri2-0 libxcb-dri3-0 libxcb-glx0 libxcb-present0 libxcb-randr0
  libxcb-render0 libxcb-shape0 libxcb-shm0 libxcb-sync1 libxcb-xfixes0 libxcursor1 libxfixes3 libxi6 libxinerama1 libxkbcommon0 libxrandr2 libxrender1 libxshmfence1
  libxss1 libxv1 libxvidcore4 libxxf86vm1 libzimg2 libzmq5 libzvbi-common libzvbi0 mesa-va-drivers mesa-vdpau-drivers ocl-icd-libopencl1 pocketsphinx-en-us
  va-driver-all vdpau-driver-all x11-common
```

Als nächstes brauche ich das Python Umgebungsmanagement

wget -qO- https://astral.sh/uv/install.sh | sh

```
root@Hatsukaze:~# wget -qO- https://astral.sh/uv/install.sh | sh
downloading uv 0.7.8 x86_64-unknown-linux-gnu
--2025-05-27 16:36:33--  https://github.com/astral-sh/uv/releases/download/0.7.8/uv-x86_64-unknown-linux-gnu.tar.gz
Resolving github.com (github.com)... 140.82.121.3
Connecting to github.com (github.com)|140.82.121.3|:443... connected.
HTTP request sent, awaiting response... 302 Found
Location: https://objects.githubusercontent.com/github-production-release-asset-2e65be/699532645/55b0de0c-6b6e-4516-b6f2-50931b850e08?X-Amz-Algorithm=AWS4-HMAC-SHA256&X
-Amz-Credential=releaseassetproduction%2F20250527%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Date=20250527T163528Z&X-Amz-Expires=300&X-Amz-Signature=3e92c293128dd93816559591
e76c87c753feb38f7cb816789631c6484ce34c9b&X-Amz-SignedHeaders=host&response-content-disposition=attachment%3B%20filename%3Duv-x86_64-unknown-linux-gnu.tar.gz&response-co
ntent-type=application%2Foctet-stream [following]
--2025-05-27 16:36:33--  https://objects.githubusercontent.com/github-production-release-asset-2e65be/699532645/55b0de0c-6b6e-4516-b6f2-50931b850e08?X-Amz-Algorithm=AWS
4-HMAC-SHA256&X-Amz-Credential=releaseassetproduction%2F20250527%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Date=20250527T163528Z&X-Amz-Expires=300&X-Amz-Signature=3e92c2931
28dd93816559591e76c87c753feb38f7cb816789631c6484ce34c9b&X-Amz-SignedHeaders=host&response-content-disposition=attachment%3B%20filename%3Duv-x86_64-unknown-linux-gnu.tar
.gz&response-content-type=application%2Foctet-stream
Resolving objects.githubusercontent.com (objects.githubusercontent.com)... 185.199.111.133, 185.199.109.133, 185.199.108.133, ...
Connecting to objects.githubusercontent.com (objects.githubusercontent.com)|185.199.111.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 17486328 (17M) [application/octet-stream]
Saving to: '/tmp/tmp.KPnyFkqnrP/input.tar.gz'

/tmp/tmp.KPnyFkqnrP/input.tar.gz          47%[=================================>                                  ]   7.91M  1.97MB/s    eta 5s
```

Active Directory

Da dieser Container teil des Active Directory sein soll, bereite ich die Einladung vor. Dazu installiere ich zuerst Samba.

apt install sssd-ad sssd-tools adcli realmd krb5-user libnss-sss libpam-sss samba-common-bin oddjob oddjob-mkhomedir

```
root@Hatsukaze:~# apt install sssd-ad sssd-tools adcli realmd krb5-user libnss-sss libpam-sss samba-common-bin oddjob oddjob-mkhomedir
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
  cracklib-runtime dirmngr gnupg gnupg-l10n gnupg-utils gpg gpg-agent gpg-wks-client gpg-wks-server gpgconf gpgsm krb5-config ldap-utils libassuan0 libavahi-client3
  libavahi-common-data libavahi-common3 libbasicobjects0 libc-ares2 libcollection4 libcrack2 libcups2 libdhash1 libgpgme11 libgssrpc4 libini-config5 libipa-hbac0
  libkadm5clnt-mit12 libkadm5srv-mit12 libkdb5-10 libksba8 libldap-2.5-0 libldap-common libldb2 libnfsidmap1 libnl-3-200 libnl-route-3-200 libnpth0 libpam-pwquality
  libpath-utils1 libpolkit-agent-1-0 libpolkit-gobject-1-0 libpwquality-common libpwquality1 libpython3.10 libref-array1 libsasl2-modules libsasl2-modules-gssapi-mit
  libsmbclient libsss-certmap0 libsss-idmap0 libsss-nss-idmap0 libtalloc2 libtdb1 libtevent0 libwbclient0 pinentry-curses pkexec policykit-1 polkitd python3-gpg
  python3-ldb python3-samba python3-sss python3-systemd python3-talloc python3-tdb samba-common samba-dsdb-modules samba-libs sssd sssd-ad-common sssd-common
  sssd-dbus sssd-ipa sssd-krb5 sssd-krb5-common sssd-ldap sssd-proxy wamerican
Suggested packages:
  dbus-user-session pinentry-gnome3 tor parcimonie xloadimage scdaemon krb5-k5tls cups-common krb5-doc libsasl2-modules-ldap libsasl2-modules-otp libsasl2-modules-sql
  pinentry-doc heimdal-clients python3-markdown python3-dnspython libsss-sudo
The following NEW packages will be installed:
  adcli cracklib-runtime dirmngr gnupg gnupg-l10n gnupg-utils gpg gpg-agent gpg-wks-client gpg-wks-server gpgconf gpgsm krb5-config krb5-user ldap-utils libassuan0
  libavahi-client3 libavahi-common-data libavahi-common3 libbasicobjects0 libc-ares2 libcollection4 libcrack2 libcups2 libdhash1 libgpgme11 libgssrpc4 libini-config5
  libipa-hbac0 libkadm5clnt-mit12 libkadm5srv-mit12 libkdb5-10 libksba8 libldap-2.5-0 libldap-common libldb2 libnfsidmap1 libnl-3-200 libnl-route-3-200 libnpth0
  libnss-sss libpam-pwquality libpam-sss libpath-utils1 libpolkit-agent-1-0 libpolkit-gobject-1-0 libpwquality-common libpwquality1 libpython3.10 libref-array1
  libsasl2-modules libsasl2-modules-gssapi-mit libsmbclient libsss-certmap0 libsss-idmap0 libsss-nss-idmap0 libtalloc2 libtdb1 libtevent0 libwbclient0 oddjob
  oddjob-mkhomedir pinentry-curses pkexec policykit-1 polkitd python3-gpg python3-ldb python3-samba python3-sss python3-systemd python3-talloc python3-tdb realmd
  samba-common samba-common-bin samba-dsdb-modules samba-libs sssd sssd-ad sssd-ad-common sssd-common sssd-dbus sssd-ipa sssd-krb5 sssd-krb5-common sssd-ldap
  sssd-proxy sssd-tools wamerican
0 upgraded, 90 newly installed, 0 to remove and 5 not upgraded.
Need to get 20.3 MB of archives.
After this operation, 85.8 MB of additional disk space will be used.
Do you want to continue? [Y/n]
```

Erstelle die sssd.conf Datei mit folgenden Daten.

nano /etc/sssd/sssd.conf

[sssd]
services = nss, pam, ssh
config_file_version = 2
domains = sotoba.de

[domain/sotoba.de]
id_provider = ad
access_provider = ad
ad_domain = sotoba.de
krb5_realm = SOTOBA.DE
realmd_tags = manages-system joined-with-adcli
cache_credentials = True
default_shell = /bin/bash
fallback_homedir = /home/%u@%d
ldap_id_mapping = True
use_fully_qualified_names = False

Ändere die Berechtigungen der Datei

chmod 600 /etc/sssd/sssd.conf
chown root:root /etc/sssd/sssd.conf
ls -l /etc/sssd/sssd.conf

```
root@Hatsukaze:~# chmod 600 /etc/sssd/sssd.conf
chown root:root /etc/sssd/sssd.conf
ls -l /etc/sssd/sssd.conf
-rw------- 1 root root 365 May 27 16:18 /etc/sssd/sssd.conf
root@Hatsukaze:~#
```

Und lade den Container in die Domäne ein und Schau, ob der Dienst richtig gestartet wurde

adcli join sotoba.de
systemctl status sssd

```
root@Hatsukaze:~# adcli join sotoba.de
Password for Administrator@SOTOBA.DE:
root@Hatsukaze:~# systemctl status sssd
* sssd.service - System Security Services Daemon
     Loaded: loaded (/lib/systemd/system/sssd.service; enabled; vendor preset: enabled)
     Active: active (running) since Tue 2025-05-27 16:21:52 UTC; 385ms ago
   Main PID: 2309 (sssd)
      Tasks: 5 (limit: 9329)
     Memory: 50.5M
        CPU: 1.517s
     CGroup: /system.slice/sssd.service
             |-2309 /usr/sbin/sssd -i --logger=files
             |-2319 /usr/libexec/sssd/sssd_be --domain sotoba.de --uid 0 --gid 0 --logger=files
             |-2321 /usr/libexec/sssd/sssd_nss --uid 0 --gid 0 --logger=files
             |-2322 /usr/libexec/sssd/sssd_ssh --uid 0 --gid 0 --logger=files
             `-2323 /usr/libexec/sssd/sssd_pam --uid 0 --gid 0 --logger=files

May 27 16:21:49 Hatsukaze sssd_ssh[2315]: Starting up
May 27 16:21:49 Hatsukaze sssd_pam[2314]: Starting up
May 27 16:21:49 Hatsukaze sssd_nss[2316]: Starting up
May 27 16:21:49 Hatsukaze sssd_ssh[2317]: Starting up
May 27 16:21:49 Hatsukaze sssd_pam[2318]: Starting up
May 27 16:21:51 Hatsukaze sssd_be[2319]: Starting up
May 27 16:21:51 Hatsukaze sssd_ssh[2322]: Starting up
May 27 16:21:52 Hatsukaze sssd_nss[2321]: Starting up
May 27 16:21:52 Hatsukaze sssd_pam[2323]: Starting up
May 27 16:21:52 Hatsukaze systemd[1]: Started System Security Services Daemon.
root@Hatsukaze:~#
```

Der Chatbot

Nun lade ich den Bot von Github runter

git clone https://github.com/Open-LLM-VTuber/Open-LLM-VTuber --recursive

```
root@Hatsukaze:~# git clone https://github.com/Open-LLM-VTuber/Open-LLM-VTuber --recursive
Cloning into 'Open-LLM-VTuber'...
remote: Enumerating objects: 5377, done.
remote: Counting objects: 100% (340/340), done.
remote: Compressing objects: 100% (106/106), done.
remote: Total 5377 (delta 271), reused 238 (delta 233), pack-reused 5037 (from 2)
Receiving objects: 100% (5377/5377), 45.46 MiB | 1.99 MiB/s, done.
Resolving deltas: 100% (3422/3422), done.
Submodule 'frontend' (https://github.com/Open-LLM-VTuber/Open-LLM-VTuber-Web) registered for path 'frontend'
Cloning into '/root/Open-LLM-VTuber/frontend'...
remote: Enumerating objects: 2954, done.
remote: Counting objects: 100% (711/711), done.
remote: Compressing objects: 100% (179/179), done.
remote: Total 2954 (delta 552), reused 581 (delta 532), pack-reused 2243 (from 1)
Receiving objects: 100% (2954/2954), 32.69 MiB | 2.01 MiB/s, done.
Resolving deltas: 100% (1852/1852), done.
remote: Enumerating objects: 5, done.
remote: Counting objects: 100% (3/3), done.
remote: Compressing objects: 100% (3/3), done.
remote: Total 5 (delta 0), reused 0 (delta 0), pack-reused 2 (from 1)
Unpacking objects: 100% (5/5), 574.25 KiB | 1.63 MiB/s, done.
From https://github.com/Open-LLM-VTuber/Open-LLM-VTuber-Web
 * branch            59808e04234378d50fc33d49275814191ab2a569 -> FETCH_HEAD
Submodule path 'frontend': checked out '59808e04234378d50fc33d49275814191ab2a569'
root@Hatsukaze:~#
```

Dann brauche ich das Programm uv, das ich mit wget heruntergeladen habe. Jedoch zeit er bei dem Befehl

uv –version

```
root@Hatsukaze:~# uv --version
uv: command not found
```

Um das zu ändern, bearbeite ich die. bahrc Datei und füge folgende Anweisung ans Ende der Datei

nano /root/.bashrc

export PATH=/root/.local/bin:$PATH

```
  GNU nano 6.2                                                    /root/.bashrc
if [ -x /usr/bin/dircolors ]; then
    test -r ~/.dircolors && eval "$(dircolors -b ~/.dircolors)" || eval "$(dircolors -b)"
    alias ls='ls --color=auto'
    #alias dir='dir --color=auto'
    #alias vdir='vdir --color=auto'

    alias grep='grep --color=auto'
    alias fgrep='fgrep --color=auto'
    alias egrep='egrep --color=auto'
fi

# some more ls aliases
alias ll='ls -alF'
alias la='ls -A'
alias l='ls -CF'

# Alias definitions.
# You may want to put all your additions into a separate file like
# ~/.bash_aliases, instead of adding them here directly.
# See /usr/share/doc/bash-doc/examples in the bash-doc package.

if [ -f ~/.bash_aliases ]; then
    . ~/.bash_aliases
fi

# enable programmable completion features (you don't need to enable
# this, if it's already enabled in /etc/bash.bashrc and /etc/profile
# sources /etc/bash.bashrc).
#if [ -f /etc/bash_completion ] && ! shopt -oq posix; then
#    . /etc/bash_completion
#fi

. "$HOME/.local/bin/env"
export PATH=/root/.local/bin:$PATH
```

Ob es funktioniert hat. Zeigt sich nach einem reboot

```
Ubuntu 22.04.5 LTS Hatsukaze tty1

Hatsukaze login: root
Password:
Welcome to Ubuntu 22.04.5 LTS (GNU/Linux 6.8.12-10-pve x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/pro
New release '24.04.2 LTS' available.
Run 'do-release-upgrade' to upgrade to it.

Last login: Tue May 27 16:28:09 UTC 2025 on lxc/tty1
root@Hatsukaze:~# uv --version
uv 0.7.8
root@Hatsukaze:~#
```

Nun Wechsel ich in das Stammverzeichnis des Bots und führe den Befehl uv sync aus, was den Chatbot installiert

```
root@Hatsukaze:~# cd Open-LLM-VTuber
root@Hatsukaze:~/Open-LLM-VTuber# uv sync
Using CPython 3.10.12 interpreter at: /usr/bin/python3.10
Creating virtual environment at: .venv
Resolved 275 packages in 9.27s
      Built srt==3.5.3
⠙ Preparing packages... (79/110)
protobuf                        ------------------------------ 959.49 KiB/1.00 MiB
pygments                        ------------------------------ 1007.45 KiB/1.17 MiB
soundfile                       ------------------------------ 992.00 KiB/1.25 MiB
aiohttp                         ------------------------------ 1020.59 KiB/1.52 MiB
networkx                        ------------------------------ 1.03 MiB/1.64 MiB
pydantic-core                   ------------------------------ 973.19 KiB/1.90 MiB
torchaudio                      ------------------------------ 939.00 KiB/3.22 MiB
uvloop                          ------------------------------ 948.81 KiB/3.65 MiB
virtualenv                      ------------------------------ 1004.79 KiB/4.10 MiB
silero-vad                      ------------------------------ 1022.80 KiB/4.79 MiB
sympy                           ------------------------------ 1.03 MiB/5.90 MiB
ruff                            ------------------------------ 994.38 KiB/10.76 MiB
onnxruntime                     ------------------------------ 990.05 KiB/12.71 MiB
nvidia-cuda-cupti-cu12          ------------------------------ 962.38 KiB/13.17 MiB
numpy                           ------------------------------ 991.18 KiB/17.40 MiB
nvidia-nvjitlink-cu12           ------------------------------ 1005.77 KiB/20.09 MiB
sherpa-onnx                     ------------------------------ 1010.74 KiB/21.16 MiB
nvidia-cuda-nvrtc-cu12          ------------------------------ 923.00 KiB/23.50 MiB
azure-cognitiveservices-speech  ------------------------------ 930.34 KiB/37.89 MiB
scipy                           ------------------------------ 986.89 KiB/38.72 MiB
nvidia-curand-cu12              ------------------------------ 990.61 KiB/53.70 MiB
nvidia-cusolver-cu12            ------------------------------ 960.00 KiB/122.01 MiB
nvidia-cusparselt-cu12          ------------------------------ 963.56 KiB/143.11 MiB
nvidia-nccl-cu12                ------------------------------ 1005.03 KiB/179.91 MiB
nvidia-cusparse-cu12            ------------------------------ 1010.38 KiB/197.84 MiB
nvidia-cufft-cu12               ------------------------------ 974.67 KiB/201.66 MiB
triton                          ------------------------------ 981.62 KiB/241.37 MiB
nvidia-cublas-cu12              ------------------------------ 955.00 KiB/346.60 MiB
nvidia-cudnn-cu12               ------------------------------ 1005.08 KiB/633.96 MiB
torch                           ------------------------------ 1011.56 KiB/731.17 MiB
```

Nun starte ich das Hauptprogramm

uv run run_server.py

```
root@Hatsukaze:~/Open-LLM-VTuber# uv run run_server.py
2025-05-27 17:40:56.680 | INFO     | __main__:<module>:86 - Running in standard mode. For detailed debug logs, use: uv run run_server.py --verbose
2025-05-27 17:40:56 | INFO     | __main__:run:57 | Open-LLM-VTuber, version v1.1.3
2025-05-27 17:40:56 | WARNING  | upgrade:sync_user_config:352 | Warning: conf.yaml not found
2025-05-27 17:40:56 | WARNING  | upgrade:sync_user_config:353 | Copying default configuration from template
2025-05-27 17:40:56 | INFO     | src.open_llm_vtuber.service_context:init_live2d:156 | Initializing Live2D: shizuku-local
2025-05-27 17:40:56 | INFO     | src.open_llm_vtuber.live2d_model:_lookup_model_info:142 | Model Information Loaded.
2025-05-27 17:40:56 | INFO     | src.open_llm_vtuber.service_context:init_asr:166 | Initializing ASR: sherpa_onnx_asr
2025-05-27 17:40:56 | INFO     | src.open_llm_vtuber.asr.sherpa_onnx_asr:__init__:81 | Sherpa-Onnx-ASR: Using cpu for inference
2025-05-27 17:40:56 | WARNING  | src.open_llm_vtuber.asr.sherpa_onnx_asr:_create_recognizer:166 | SenseVoice model not found. Downloading the model...
2025-05-27 17:40:56 | WARNING  | src.open_llm_vtuber.asr.utils:check_and_extract_local_file:160 | Local file not found or not a tar.bz2 archive: models/sherpa-onnx-sens
e-voice-zh-en-ja-ko-yue-2024-07-17.tar.bz2
2025-05-27 17:40:56 | INFO     | src.open_llm_vtuber.asr.sherpa_onnx_asr:_create_recognizer:176 | Local file not found. Downloading...
2025-05-27 17:40:56 | INFO     | src.open_llm_vtuber.asr.utils:download_and_extract:82 | Downloading https://github.com/k2-fsa/sherpa-onnx/releases/download/asr-model
s/sherpa-onnx-sense-voice-zh-en-ja-ko-yue-2024-07-17.tar.bz2 to ./models/sherpa-onnx-sense-voice-zh-en-ja-ko-yue-2024-07-17.tar.bz2...
sherpa-onnx-sense-voice-zh-en-ja-ko-yue-2024-07-17.tar.bz2:   5%|▌                                                    | 51.7M/999M [00:26<11:32, 1.44MiB/s]
```

Jetzt fehlt noch die KI. Dazu lade ich zuerst einmal LLama runter

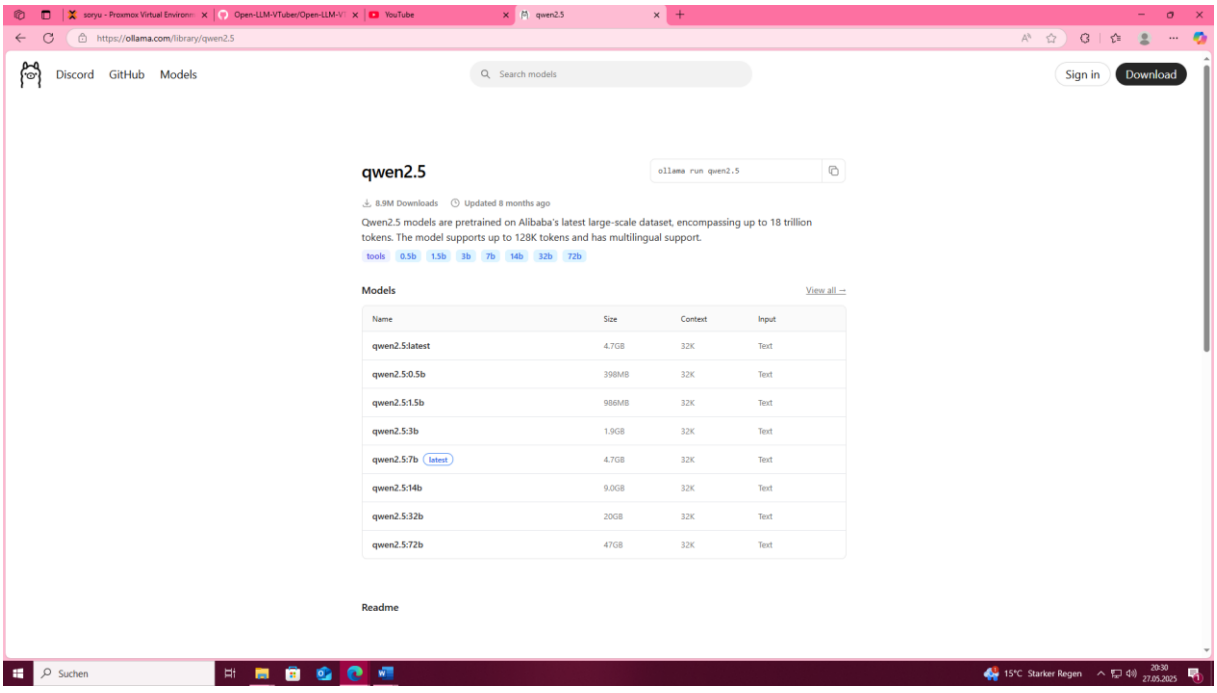
curl -fsSL https://ollama.com/install.sh | sh

```
root@Hatsukaze:~# curl -fsSL https://ollama.com/install.sh | sh
>>> Installing ollama to /usr/local
>>> Downloading Linux amd64 bundle
#                                                                                2.4%
```

Danach lade ich ein passendes Model runter
Informationen über verscheinde Modelle finde ich hier:

[Ollama Search]

[qwen2.5:1.5b]



Aber da die KI zuerst einmal nur im CPU-Modus läuft, wähle ich ein schlankes Modell zum Testen wie qwen:0.5b und lade es wie folgt runter. Danach schaue ich nach ob alles vor Ort ist und starte die KI

ollama pull qwen:0.5b
ollama list
ollama run qwen:0.5b