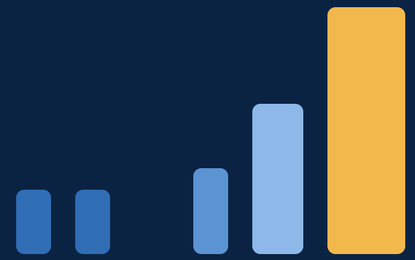
COMPASS TRAINING PROGRAM · SESSION 4



COMPASS

How it works, how it was built, and how we tested it

Platform overview and technical specifications

Dan Bugariu · Timo Bejan

18 November 2026

WHAT IT DOES

A quiet cue when the conversation drifts

COMPASS listens to a live conversation, scores every operator reply on Motivational Interviewing technique and on rapport, and taps the operator’s wrist when either starts to slip.

Scores the operator

Each finished operator reply gets two verdicts, one for MI and one for Rapport: OK, Warning or Alert. The source is never scored.

Shows its evidence

Every verdict names the behaviour it found and points to the turns it relied on, so it can be reviewed afterwards.

Keeps the conversation private

No transcript is sent to a generative AI model. Live audio is never written to disk.

ARCHITECTURE

From voice to cue in six steps

A few seconds from end of reply to verdict

01

Capture

Phone or watch microphone. Audio is processed live and never saved to disk.

02

Transcribe

Speech recognition runs on the phone. English only.

03

Attribute

The operator enrols a voiceprint once. Diarization then tells operator from source.

04

Window

When an operator reply is complete, the last three exchanges become the model input.

05

Classify

Two encoder models, one for MI and one for Rapport, each return OK, Warning or Alert.

06

Cue

A vibration pattern plays only when a level rises. The session is saved for review.

CORE FUNCTION

What the model reads, and what it returns

INPUT: THE LAST THREE EXCHANGES

SOURCE I keep wondering if I should say something to my manager, or just stay quiet.

OPERATOR You're pulled both ways. What worries you most about speaking up?

SOURCE I could lose my job. My wife doesn't work.

> OPERATOR Listen, you're overthinking this. What you need to do is take photos of the plates tonight.

VERDICT

MI Warning · Rapport Warning

BEHAVIOUR

Autonomy support (MI) · Autonomy (Rapport)

EVIDENCE

The operator turns it relied on

RECORD

Model version, probabilities, input fingerprint

Only the past is visible to the model. It never sees what the source says next.

INSIDE THE MODEL

The model places every reply on a map

1 Words in

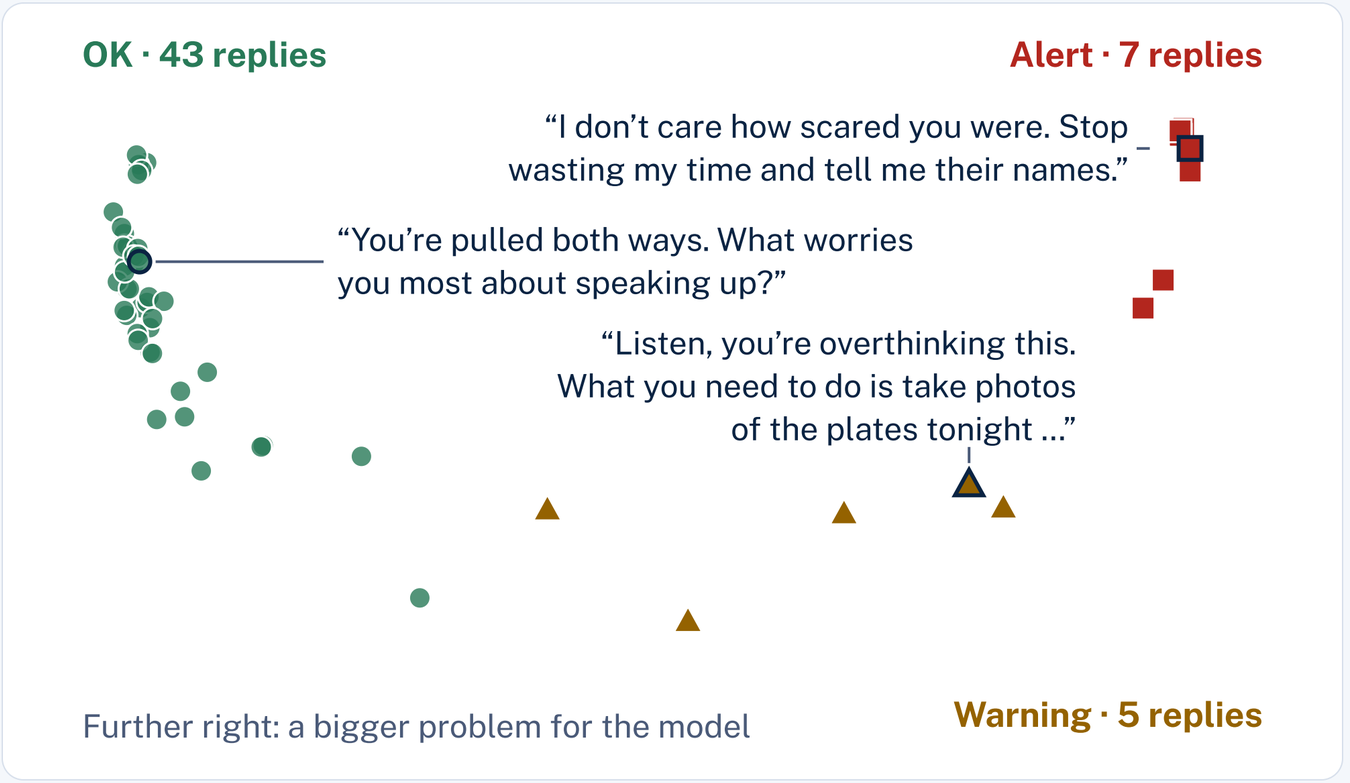
The reply, with the two exchanges before it, is cut into word pieces and read by a language model.

2 768 numbers out

The model sums up what it read as a list of 768 numbers, called an embedding. Think of it as a position on a map of meaning.

3 Position decides the cue

Replies that do the same thing land close together, whatever the wording. Training on about 82,000 labelled replies taught the model which regions are OK, Warning and Alert.



Each mark is one operator reply from today’s demo scripts, placed by the Compass v3.1 MI model. 768 numbers flattened to two for the picture.

WHAT IT SCORES

Two axes, three levels

Motivational Interviewing

Is the operator drawing out the source's own reasons, or pushing his own?

Open questions · Reflections · Affirmations · Summaries · Autonomy support · Sustain-talk management · Collaboration · Evocation · Planning · Information giving

Rapport and alliance

Is the working relationship with the source holding?

Attentiveness · Trust and respect · Welfare and security · Autonomy · Elicitation · Connected flow · Self-disclosure · Repair · Closure

OK

Nothing to signal. The watch stays silent.

Warning

A slip. Technique or relationship is starting to drift.

Alert

Serious or sustained: pressure, accusation, dismissal, threat.

Taxonomy: 200 operator indicators defined by the UVT team, positive and negative, for each axis.

HAPTIC FEEDBACK

Four cues, and you choose how each one feels

MI WARNING

Single



One short pulse

MI ALERT

Escalating



Three pulses, each stronger

RAPPORT WARNING

Double



Two even pulses

RAPPORT ALERT

Urgent



Two taps, then a long buzz

Fires when a level rises. It stays silent while the level holds and resets after the next OK reply.

You choose the pattern. For each cue: a built-in pattern, one you build yourself, or off.

Try it before the session. The phone plays any pattern on the watch, so you know what it feels like.

Sent as an alarm vibration. That is the kind the Galaxy Watch plays with the screen off.

Bar width is pulse length, bar height is strength. Shown: the phone app’s starting patterns. A custom pattern has 1 to 12 pulses.

INTERFACES

Where it runs: on the phone and the wrist

DURING THE MEETING

Phone app

No account. Speech recognition, speaker separation and scoring all run on the phone.

ON THE WRIST

Watch

Apple Watch and Galaxy Watch. Receives the cue and the session state, nothing else. Can also serve as the microphone.

AFTER THE MEETING

Session review

The transcript on the phone, with each warning and alert marked on the reply that caused it.

Live audio is never written to disk.

Sessions and voiceprints stay on the phone.

No transcript goes to a generative AI model.

HOW IT WAS BUILT

COMPASS learns from examples. We built them three ways.

An example is a conversation in which every operator reply is labelled OK, Warning or Alert.

Generated dialogues

Scenario-driven conversations written by large language models, with deliberate operator slips.

UVT-guided scenarios

Generated again, steered by the instructors' scenario material and the UVT taxonomy.

Public human conversations

Real people talking, from open research datasets, relabelled with the COMPASS rubric.

TRAINING DATA

Writing and labelling are kept apart

01

Scenario

Who, where, what is at stake, and how the operator will slip.

A rail yard on a rainy night. A switch clerk saw a man photographing a van and fears being blacklisted if she talks.

02

Conversation

A plain spoken dialogue, 18 to 38 turns. No labels yet.

“Thanks for coming in, Mara. Can you walk me through the night from your side?”

03

Labels

Every operator reply, both axes: level, behaviour, evidence turns.

Reply 9: MI Warning, open questions. Rapport Warning, connected flow.

The writer never labels its own text. A different model does.

Two annotators work independently. A third settles disagreements, or the reply is left out.

Labels use only the past, the same three exchanges the model sees during a conversation.

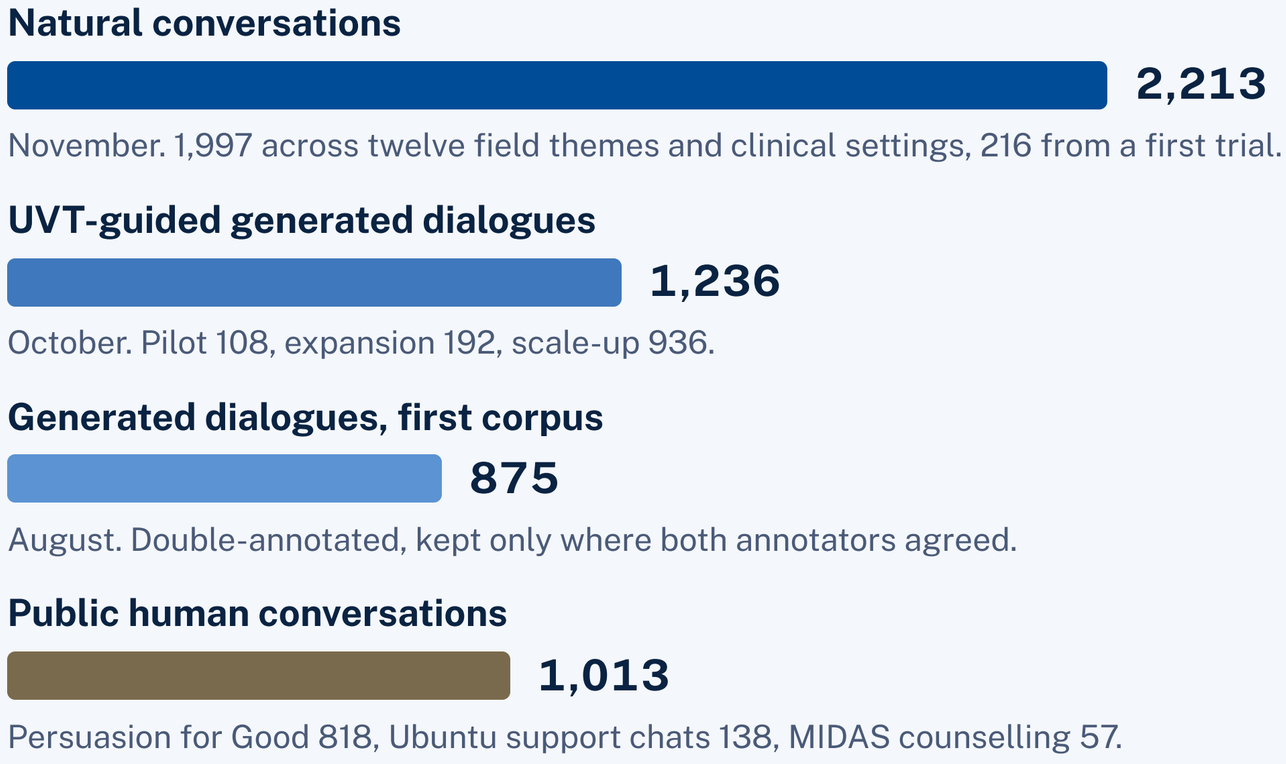
Example from the published training corpus, conversation s001-01.

TRAINING DATA

What Compass v3.1 learned from

5,337
conversations

82,286
scored operator replies



Training and validation together. Four in five replies or more are labelled OK; the rest are Warning or Alert.

HOW IT WAS BUILT

Five months, one round at a time

JUNE

Expert questionnaire

Nine experts ranked six communication models. MI and Rapport came first.

AUGUST

First corpus, first model

1,240 generated dialogues, then a rebuilt corpus of 976. Compass v2.1 runs on the phone.

SEPTEMBER

Apps and haptics

iPhone and Watch apps, the web workspace, the haptic pattern builder.

OCTOBER

Human text, larger model

UVT-guided scenarios and public human conversations. Compass v3.0 trained, with promising results.

NOVEMBER

Experts and natural data

Android phone and Galaxy Watch. Three UVT experts rate 60 dialogues. 2,000 natural conversations lead to Compass v3.1.

115 training runs across four studies in October alone. The follow-up studies found nothing better, so v3.0 stayed. In November, 18 runs with 2,000 natural conversations gave Compass v3.1.

HOW IT WAS TESTED

Tested on conversations it never trained on

Held-out conversations

Expert-rated dialogues, recorded counselling sessions and other human conversations. None of them was used in training.

Two plain questions

When the operator slipped, did COMPASS flag it? And when COMPASS flagged a reply, was there a slip?

Fixed before looking

Thresholds and test sets are frozen before a test score is read. The expert ratings also chose between training recipes.

What we compare against. Every test conversation has reference labels that say where the slips are. For 60 dialogues three UVT experts made them. For the rest, AI annotators did, cross-checked by a second and third model.

RESULTS · RESEARCH PROTOTYPE, COMPASS V3.1

Seven in ten slips caught where experts marked them

Slips caught

The operator slipped, and COMPASS flagged it.

Example: experts marked 207 MI slips, 150 flagged.

Flags that were right

COMPASS flagged a reply, and there was a slip on it or on the one next to it.

Example: 223 MI flags in the same dialogues, 179 right.

Test set	MI slips caught	MI flags right	Rapport slips caught	Rapport flags right
Expert-rated dialogues (60)*	72%	80%	70%	82%
Recorded counselling sessions (40)	70%	65%	62%	53%
Other human conversations (117)	82%	70%	46%	51%
Instructor scenarios (6)*	70%	64%	39%	44%

MI is the stronger axis: seven to eight slips in ten caught on every set, and six to eight flags in ten right. **Rapport** does as well where experts rated it. On human conversations it catches four to six slips in ten, and about half of its flags are right.

*The expert ratings chose the training recipe and the instructor scenarios guided development, so neither is a blind test.

READING THE RESULTS

Experts read the same reply differently, too

TWO EXPERTS, ONE REPLY

84%

of 635 replies got the same verdict from the two UVT experts who rated them: slip or no slip, on either axis.

WHERE ONE SAW A SLIP

49 of 153

replies that at least one of the two marked as a slip were marked by both. Most of the agreement is on clean replies.

HOW STRICT TO BE

10% to 26%

of the replies marked as a slip, from the least strict to the most strict of the three experts.

A cue is a reason to look again. The operator decides.

COMPASS and an expert give the same verdict on 78 to 87% of replies. Two experts: 82 to 87%.*

*The expert ratings chose the training recipe, so this is not a blind test.

WHAT THE TESTS TAUGHT US

Realistic conversations beat more conversations

SEPARATE WRITER AND LABELLER

21% → 64%

of the instructor-marked MI problems caught, same small model, after the corpus was rebuilt in August.

ADD REAL HUMAN TEXT

61% → 71%

of MI problems caught on held-out human conversations. False warnings on clean replies fell from 28% to 20%.

MAKE GENERATED TEXT SOUND NATURAL

72 → 77

of 92 expert-marked slips caught with the same number of flags, after 2,000 more natural conversations. More of the old kind had given no gain.

More natural training conversations gave the latest gain, a small one.

You know how a real source meeting sounds. The realism vote sets the brief for the next batch we generate.

CURRENT LIMITS

What COMPASS cannot do yet

Rapport is the weaker axis

On human conversations about five Rapport flags in ten are right and four to six slips in ten are caught. MI is at about seven in ten on both.

Most reference labels are AI-made

Three UVT experts have rated 60 dialogues so far. They agree with each other only in part, so more ratings are needed.

Short memory

Each verdict sees the last three exchanges and nothing earlier.

English only

Speech recognition and the model work in English.

Depends on who said what

A third voice, overlap or very short replies can put words in the wrong mouth.

It is a training aid

A cue says “check yourself”. It does not judge the source, the information or the outcome of the meeting.

ROADMAP

What comes next

01	**More realistic training data**	Your realism ranking from this training sets the brief for the next batch of training conversations.
02	**Test it in the lab**	Role-played meetings with instructors, where every cue can be checked against their judgement.
03	**Test it in the field**	Worn during exercises: does the cue arrive, is it felt, and does it help the operator correct course?

YOUR TURN · ABOUT EIGHT MINUTES

Which of these sounds like a real meeting?

01 Scan the code or type the address. No account, no name.

02 Read ten short excerpts, one at a time. Each was written by an AI model to train COMPASS.

03 After each one, say if it sounds more real or less real than the one before.

04 The ranking comes up on this screen when the votes are in.

What we do with it. The ten excerpts differ in who wrote them, the setting and how people talk. The ranking tells us which style to ask for in the next batch of training conversations.



NEXT · PART B

Live demo: five scripted conversations

Script	Length	What to watch for
The port gate	1:15	Good practice throughout. The watch stays silent.
The depot guard	1:45	The operator starts directing. Warning, then MI Alert, then a repair clears it.
The checkpoint driver	1:20	The operator ignores fear, then accuses. Warning, then MI Alert, then Rapport Alert.
The interpreter	3:45	A well-run meeting with two slips. Each one is cued and repaired on the next reply.
The shopkeeper	3:40	Time pressure turns into threats. Alert on both axes, then a repair and a clean ending.

Read aloud by two people. Watch the transcript, the verdict on each operator reply, and the moment the cue fires.

Scripts were checked reply by reply on Compass v3.1, including under simulated speech-recognition errors.